the WINNOWER

MATHEMATICS

π

NWO, Gender bias and Simpson's paradox

CASPER ALBERS

CORRESPONDENCE:
casper@casperalbers.nl

DATE RECEIVED:
September 24, 2015

This blog post is an abridged and translated version of my blog post in Dutch. A version of this blog post, further abridged to fit within the 500 word limit, has been submitted to PNAS as comment to the paper Gender contributes to personal research funding success in The Netherlands.

In the early seventies, the University of California, Berkeley received sincere negative attention due to supposed gender bias in graduate admissions. The data for fall 1973 clearly seemed to point in this direction:

	Nr. of applications	admissions
Male	8442	44%
Female	4321	35%

Out of 8442 male applicants, 44% was admitted, whereas out of the 4321 female applicants, only 35% was admitted. The χ^2-test on the 2×2 frequency table (or any other sensible test for 2×2 tables) will give a very significant result, with a *p*-value smaller than one in a billion. A scrutiny of the data in Science by Bickel, Hammel and O'Connel (1975) revealed that there was no evidence for gender bias. This apparent counterintuitive result was due to the interaction with an external variable. Not all departments at the university had the same admission rate, and there was a relation between the proportion of female applications and the admission rate.

Competitive departments such as English received relatively many female applications, whereas departments such as chemistry, with a surplus of male applications, where much less selective. When studying the male/female admissions on a departmental level, the supposed gender bias disappeared. (For the fall 1973 data, there even was evidence of bias *in favour* of women.) This paradox is termed *spurious correlation* or *Simpson's paradox*, after the British statistician Edward Simpson. (For a recent open access paper on Simpson's paradox in psychological science, see Kievit, Frankenhuis, Waldorp and Borsboom, 2013.)

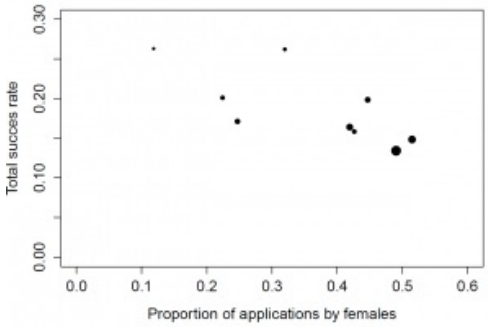
The authors, correctly, point at another pitfall: although there seemed to be evidence of bias (in favour of women) for fall 1973, there is no such evidence for other years. A significant result once in a number of years, could just be coincidence.

In the analysis by Van der Lee and Ellemers the same two flaws occur in a setting not too dissimilar from the one discussed above. Based on the results of $n = 2{,}823$ grant applications to the "VENI programme" of the Netherlands Organisation for Scientific Research, NWO, in the years 2010, 2011 and 2012, the authors conclude that the data "provide compelling evidence of gender bias in personal grant applications to obtain research funding". One of the main results this claim is based upon the following table:

	applications	Succesfull
Male	1635	17,7%
Female	1188	14,9%

When applying a standard χ^2-test to the data, the authors find a just significant *p*-value of .045. It is not only questionable to denote a *p*-value this close to 0.05 as "compelling evidence", due to Simpson's paradox, this *p*-value simply is wrong.

In the supplementary table S1 (Van der Lee and Ellemers, 2015), available online without paywall, a breakdown of the 2,823 grant applications per discipline is presented. The proportion of female applicants varies from 11.8% (physics) to 51.4% (health sciences), and the total succes rate varies from 13.4% (social sciences) to 26.3% (chemical sciences).



PROPORTION OF APPLICATIONS BY FEMALE SCIENTISTS VS TOTAL SUCCESS RATE. SIZE OF THE MARKERS IS PROPORTIONAL TO NUMBER OF APPLICATIONS WITHIN THE DISCIPLINE.

The figure above visualises these data and immediately shows a clear negative relation between the proportion of female applicants and the total succes rate (i.e. the rate for men and women combined). In four out of the nine disciplines, women have a higher succes rate than men, and in five out of nine, men have a higher succesrate than women. When taking into account that multiple comparisons are performed, for none of the disciplines the gender bias – either in favour of women or in favour of men – is significant (at the $\alpha = .05$ level). Thus, when taking into account the spurious correlation, the "compelling evidence" is lost.

Bickel *et al.* (1975) pointed at a second pitfall, concerning focussing on the year(s) where the difference was signicant and ignoring the other year(s) where it was not. Again, a similar situation occurs here. NWO publishes the results of all VENI rounds since its establishment in 2002 until 2015 (except for 2012) on its website. In some years, such as 2011, men received relatively more grants than women; and in other years, such as 2010 and 2015, the reverse was true. The *z*-test for log-odds ratio only provides a significant sign of gender bias in favour of men for the years 2010 ($z = 2.002$, $p = .023$) and 2011 ($z = 1.752$, $p = .040$) and a significant gender bias in favour of women for 2002 ($z = 2.005$, $p = .022$). When applying the Bonferroni correction for multiple comparisons none of these gender biases are significant.

Conclusion. Van der Lee and Ellemers failed to recognise the dependence of the results on the different NWO disciplines. Futhermore, they focused on results during a three-year, whereas the results of the other periods in which VENI-grants where provided did not confirm the just significant results for 2010-2012. As a consequence, the conclusion of "*compelling evidence of gender bias*" is inappropriate. In the data, there is no evidence for gender bias (which does *not* have to mean that there is no gender bias). In discussions on institutional sexual discrimination, it is important to stay factual.

Furthermore, I find it worrying that this analysis gets published. Simpson's paradox is one of statistics most well-know paradoxes (I teach it yearly to a new batch of psychology students in Groningen) and PNAS is a high-ranking journal with an impact factor of nearly ten. This paper – where conclusions are drawn on basis of flawed methodology – is not an exception. Apparently, the current peer-review system is inadequate in filtering out methodological flaws in papers. If a system doesn't work, it should

be changed.

Final note. The paper by Van der Lee and Ellemers focusses on more tests than just the one criticised by me here. However, these other tests make use of related data (e.g. the number of applicants that go through to the interview-stage) and it is not unlikely that Simpson's paradox plays a role there too. (The data provided in the paper was insufficient for me to check this.) And even if it does not: the authors are providing interpretations to effects with tiny effect sizes (partial eta-squareds of 0.006(!))… Furthermore, the paper contains a section on "language use" in NWO documents. My comments do not apply to this section.