PROBLEMATIC SCIENCE

Flawed Statistics

The reproducibility crisis has revealed many kinds of technical problems in medical studies; and Wansink committed a large number of them in his behavioral research. Several researchers have narrowed their focus and studied the effects of p-hacking on scientific research. Megan Head's 2015 study looked at p-values in papers across a range of disciplines and found evidence that p-hacking is "widespread throughout science."[57] However, Head and her co-authors downplayed the significance of that finding and argued that most p-hacking probably just confirmed hypotheses that were fundamentally true. A 2016 paper coauthored by Ioannidis seemed to demolish those reassurances,[58] but another paper revisiting Head's study argued that she and her co-authors overestimated the evidence for p-hacking.[59] A separate paper that examined social science data found "encouragingly little evidence of false-positives or p-hacking in rigorous policy research,"[60] but the qualifier "rigorous" sidesteps the question of how much policy research does *not* meet rigorous standards. Still, these initial results suggest that while p-hacking significantly afflicts many disciplines, it is not pervasive in any of them.

P-hacking may not be as widespread as one might fear, but it appears that many scientists who routinely use p-values and statistical significance testing misunderstand those concepts, and therefore employ them improperly in their research.[61] In March 2016, the Board of Directors of the American Statistical Association issued a "Statement on Statistical Significance and *p*-Values" to address common misconceptions. The Statement's six enunciated principles included the admonition that "by itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis."[62]

Such warnings are vital, but, as the Wansink affair illustrates, scientists also make many other sorts of errors in their use of statistical tests.[63] The mathematics of advanced statistical methods are difficult, and many programs of study do not adequately train their graduates to master them.[64] The development of powerful statistical software also makes it easy for scientists who don't fully understand statistics to let their computers perform statistical tests for them. Jeff Leek, one of the authors of the popular blog *Simply Statistics*, put it bluntly in 2014: "The problem is not that people use p-values poorly, it is that the vast majority of data analysis is not performed by people properly trained to perform data analysis."[65]

> *"The problem is not that people use p-values poorly, it is that the vast majority of data analysis is not performed by people properly trained to perform data analysis."*
> *– Jeff Leek*

Faulty Data

Statistical analysis isn't the only way research goes wrong. Scientists also produce supportive statistical results from recalcitrant data by fiddling with the data itself. Researchers commonly edit their data sets, often by excluding apparently bizarre cases ("outliers") from their analyses. But in doing this they can skew their results: scientists who systematically exclude data that undermines their hypotheses bias their data to show only what they want to see.

Data based on self-report surveys is especially unreliable, particularly when the reporting involves essentially subjective mental states.[66] The crisis of reproducibility suggests that research based on self-report surveys should be scrutinized with even greater skepticism than research based on externally verifiable data.

Scientists can easily bias their data unintentionally, but some deliberately reshape their data set to produce a particular outcome. One anonymized survey of more than 2,000 psychologists found that 38% admitted to "deciding whether to exclude data after looking at the impact of doing so on the results."[67] Few researchers have published studies of this phenomenon, but anecdotal evidence suggests it is widespread. In neuroscience,

> *there may be (much) worse things out there, like the horror story someone (and I have reason to believe them) told me of a lab where the standard operating mode was to run a permutation analysis by iteratively excluding data points to find the most significant result. … The only difference from* [sic] *doing this and actually making up your data from thin air … is that it actually uses real data – but it might as well not for all the validity we can expect from that.*[68]

THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

Figure 13: Machine Learning

Researchers can also bias their data by ceasing to collect data at an arbitrary point, perhaps the point when the data that has already been collected finally supports their hypothesis. Conversely, a researcher whose data doesn't support his hypothesis can decide to keep collecting additional data until it yields a more congenial result. Such practices are all too common. The survey of 2,000 psychologists noted above also found that 36% of those surveyed “stopped data collection after achieving the desired result.”[69]

One anonymized survey of more than 2,000 psychologists found that 38% admitted to “deciding whether to exclude data after looking at the impact of doing so on the results.”

Another sort of problem arises when scientists try to combine, or “harmonize,” multiple preexisting data sets and models in their research—while failing to account sufficiently for how such harmonization magnifies the uncertainty of their conclusions. Claudia Tebaldi and Reto Knutti concluded in 2007 that the entire field of probabilistic climate projection, which often relies on combining multiple climate models, had no verifiable relation to the actual climate, and thus no predictive value. Absent “new knowledge about the [climate] processes and a substantial increase in computational resources,” adding new climate models won’t help: “our uncertainty should not continue to decrease when the number of models increases.”[70]

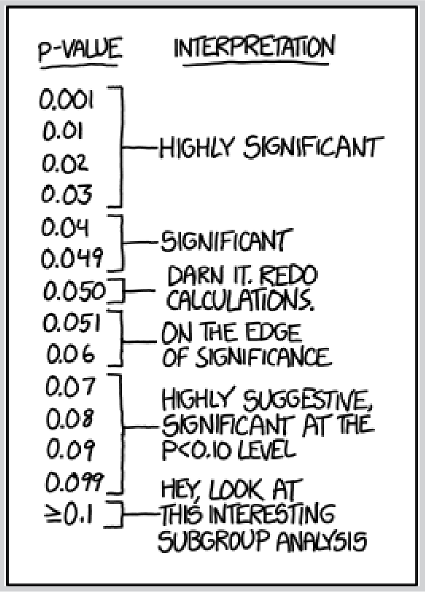
P-VALUE
INTERPRETATION
0.001
0.01
0.02
0.03
HIGHLY SIGNIFICANT
0.04
0.049
SIGNIFICANT
0.050
DARN IT. REDO CALCULATIONS.
0.051
0.06
ON THE EDGE OF SIGNIFICANCE
0.07
0.08
0.09
0.099
HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
≥0.1
HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

Figure 14: P-Values, Interpreted

Pervasive Pitfalls

Necessary and legitimate research procedures drift surprisingly easily across the line into illegitimate manipulations of the techniques of data collection and analysis. Researcher decisions that seem entirely innocent and justifiable can produce “junk science.” In a 2014 article in the *American Scientist*, Andrew Gelman and Eric Loken called attention to the many ways researchers’ decisions about how to collect, code, analyze, and present data can vitiate the value of statistical significance.[71] Gelman and Loken cited several researchers who failed to find a hypothesized effect for a population as a whole, but did find the effect in certain subgroups. The researchers then formulated explanations for why they found the postulated effect among men but not women, the young but not the old, and so on. These researchers’ procedures amounted not only to p-hacking but also to the deliberate exclusion of data and hypothesizing after the fact: they were guaranteed to find significance somewhere if they examined enough subgroups.

Researchers allowed to choose between multiple measures of an imperfectly defined variable often decide to use the one which provides a statistically significant result. Gelman and Loken called attention to a study that purported to find a relationship between women's menstrual cycles and their choice of what color shirts to wear.[72] They pointed out that the researchers framed their hypothesis far too loosely:

> *Even though Beall and Tracy did an analysis that was consistent with their general research hypothesis—and we take them at their word that they were not conducting a "fishing expedition"—many degrees of freedom remain in their specific decisions: how strictly to set the criteria regarding the age of the women included, the hues considered as "red or shades of red," the exact window of days to be considered high risk for conception, choices of potential interactions to examine, whether to combine or contrast results from different groups, and so on.*[73]

These researchers' procedures were equivalent to p-hacking, the deliberate exclusion of data, and hypothesizing after the fact: they were guaranteed to find significance somewhere if they examined enough subgroups.

Would Beall and Tracy's hypothesis have produced statistically significant results if they had made different choices in analyzing their data? Perhaps. But a belief in the very hypothesis whose validity they were attempting to confirm could have subtly influenced at least some of their choices.