Lily Eckhart, Justin Eldridge, Emily Prewett

DSCI 478-001

Dr. Emily King

7th May, 2024

Publication Bias in Nutrition Research: A Z-Curve Analysis

Introduction:

When reading news articles or scrolling through social media headlines, one might get the impression that science is an uninterrupted string of successes where researchers continuously validate their predictions and confirm their hypotheses. In reality, countless studies yield null results, but these are simply not publicized because they are uninteresting to a general audience. This leads to a distorted view of the scientific method where positive findings are the rule, and non-findings the exception.

This distortion is not solely due to media bias. Scientific journals themselves often prioritize publishing studies with statistically significant results. As the psychologist Stuart Ritchie notes, in an ideal world, the methodological rigor of a study would be all that matters when it comes to the decision to publish (Ritchie, 2021). If everyone agrees that the study is a well-designed test of an important question, it should be published regardless of the result. In practice, however, the outcome often plays an almost deterministic role in publication– a phenomenon known as publication bias.

In favoring results with a positive outcome over null results, publication bias leads to a disproportionate amount of false positives being published. False positives cause problems because they seem to be genuine findings, when in reality, they are just statistical flukes. When these false positives are published frequently enough, it can bias meta-analyses and literature reviews, which leads to future resources being wasted building on results that aren't true in the first place. This heavily contributes to the "replication crisis", which is a term coined to describe the phenomenon of often-cited results in fields like psychology being impossible to reproduce.

To evaluate the extent of the replication crisis in psychology, a large group of researchers published a paper in 2015 detailing their attempts to replicate studies from prominent psychology journals. They chose to replicate 100 studies from three top psychology journals and found that only a measly 39% of them were successfully replicated (Open Science Collaboration, 2015). This alarming figure illustrates that the literature has become infested with flimsy, unreplicable results which were published for their novelty rather than their reliability. While psychology has begun to confront this issue, the extent of publication bias in other fields is unclear.

We chose to examine the prevalence of publication bias and numerical errors in nutrition research. Nutritional advice is notoriously inconsistent, with guidance from the US Department of Health and Human Services on foods like fats, coffee, and eggs flip-flopping multiple times over the years (Hickerson, 2024). We suspect that publication bias, with its propensity for false positives, might contribute to this instability.

Nutrition studies face certain challenges that could amplify the problem: small sample sizes, self-reported survey data prone to social desirability bias, and observational study designs riddled with potential confounds. These methodological issues create favorable conditions for p-hacking, which is the manipulation of analyses to produce statistically significant results. By investigating the presence of publication bias and numerical inconsistencies in this field, we hope to assess the reliability of nutrition research.

Background: Tools for Detecting Publication Bias and Numerical Errors

While there are several methods for identifying publication bias in a body of research, we will primarily be focused on identifying skewed distributions of Z-scores. For example, consider a sample of Z-scores that contains a large number of just-barely-significant results (e.g., Z=1.97, p=0.049) and a small number of just-barely-non-significant results (e.g., Z=1.95, p=0.051). This is a skewed Z-score distribution because there is a sharp increase in the number of results between the two nearly identical Z-scores. If there were no publication bias, the frequency of results would instead be very similar for both.

Z-Curve Analysis:

Z-curve analysis, developed by the psychologist Ulrich Schimmack in 2020, is a statistical technique that estimates the prevalence of publication bias based on the distribution of reported z-scores reconstructed from p-values, t-values, F test statistics, or chi-squared tests (Schimmack, 2020). A key advantage of this method is its ability to handle censored p-values (e.g., “p<0.05”), which are common in published studies. In the figure below, we can see what the presence of publication bias looks like.

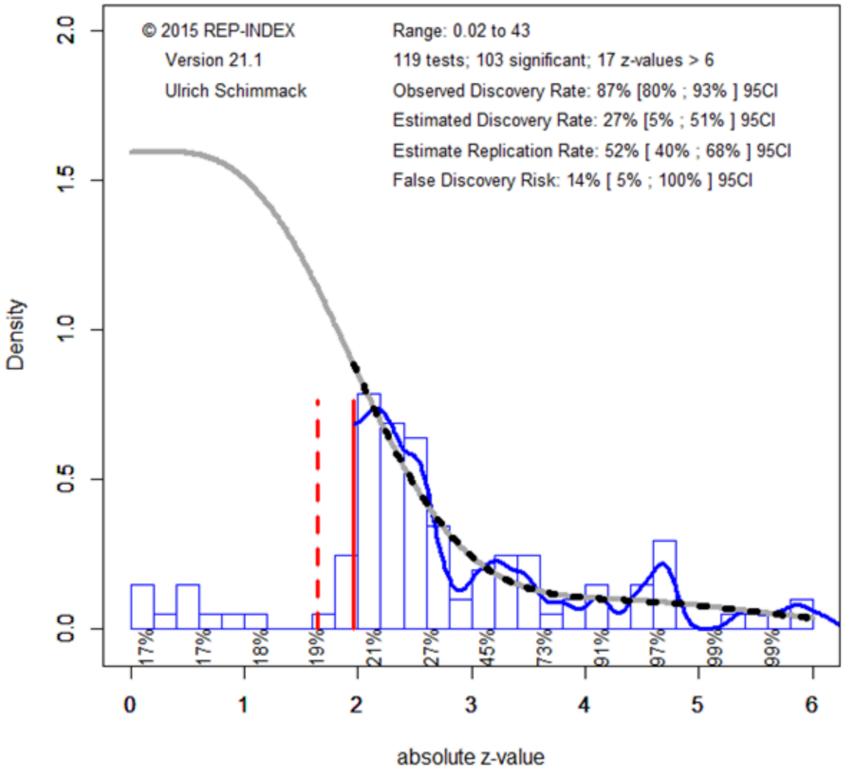

Figure 1: Z-curve example output that suggests publication bias. https://replicationindex.com/2021/04/25/z-curve-an-even-better-p-curve/

In Figure 1, we can see that there are a large number of barely significant results just to the right of the red vertical line and a tiny number of results that are barely significant just to the left of the red vertical line. This steep drop in frequency from just significant to just not significant results is not consistent with random sampling error and is indicative of publication bias. Note too that there is a large discrepancy between the observed discovery rate (proportion of statistically significant results) and the expected discovery rate. If publication bias were not present, the plotted line would be much flatter with comparable numbers of results on either side of the red line.

The method has several important outputs:

- **Observed Discovery Rate (ODR):** The proportion of statistically significant results in the observed studies. If there are significant levels of publication bias, we would expect a large proportion of published results to be statistically significant. For example, in the Z-curve shown in Figure 1, we can see that the ODR was nearly 90%. If there was little to no publication bias, we would expect the ODR to be much closer to 50%.
- **Expected Discovery Rate (EDR):** The proportion of statistically significant results that would be expected without publication bias (i.e., mean power of all studies before selection). Put another way, EDR estimates the proportion of statistically significant results if all studies were available (i.e., if researchers emptied their file cabinets). A large mismatch between the ODR and EDR is an indicator of publication bias.
- **Expected Replication Rate (ERR):** The mean power of only the statistically significant studies. This predicts the frequency of statistically significant results in replication studies. In Figure 1, we can see that while 87% of published results were statistically significant, we would only expect half of them to replicate successfully.

Z-curve is especially valuable in real-world applications as it tolerates imprecise p-values while still estimating the prevalence of bias.

Initial Approach and Abandoned Methods:

We initially planned to complement our Z-curve analysis with a review of numerical inconsistencies using the statcheck package in R. This tool, described as a spell-checker for statistics, attempts to verify that reported statistics (e.g., t-values, F-values, and p-values) within a paper are internally consistent. We did not use statcheck in our study, so we will not delve into the technical details, but the logic is similar to verifying that the sides of a triangle match the Pythagorean theorem; if you know two of the side lengths, you can compute the third.

We decided against using statcheck because it proved unreliable, even in its intended domain. Although it claimed to work with APA-formatted papers, it frequently failed to process papers from APA journals. The latest version of this package, whose documentation has not been updated, relies on the pdf_text() function from the pdftools package to extract text from PDFs, a method that frequently failed unless we extracted the text manually. Given this unreliability, inflexibility, and time constraints, we decided to abandon this approach.

Nonetheless, during our review, we came across a number of numerical issues. Some papers reported impossible values such as p=0.00 or p=1.00. Others appeared to perform multiple subgroup analyses without a correction for multiple comparisons, like Bonferroni. These issues hinted at quality control problems in the literature, but without a reliable tool, we were unable to document them systematically.

Final Approach: Z-Curve Analysis

Sampling Scheme:

We used the Web of Science database to collect research articles published in the last ten years under the topic “nutrition”. Each team member gathered the first 30 articles from a unique three-year window. We deliberately avoided filtering by article type, such as clinical trials, to avoid biasing our sample towards higher-quality, federally funded randomized controlled trials. Instead, we sought a more representative sample of the field more broadly.

Data Collection:

From each article, we recorded bibliographic information and any reported statistical results tied to the main hypotheses. This included z-scores, p-values (both exact and inequalities), t-values, chi-square values, and confidence intervals. Since the zcurve() function does not support confidence intervals directly, we used cell formulas in Google Sheets to back-calculate p-values from the reported bounds. We also recorded reasons for exclusion as necessary– such as lack of an experimental hypothesis, failure to report results, or irrelevance (e.g., reviews or opinion pieces).

Variables:

Column Name	Description	Column Name	Description
Journal	Journal that the article was published in	**Excluded**	Indicates if an article was excluded for any reason (Yes/No)

Year	Year the article was published	**Unclear_outcome**	Indicates exclusion due to unclear experimental outcome (Yes/No)
Article_number	Article number (if applicable)	**Not_reported**	Indicates exclusion due to unreported relevant statistics (Yes/No)
Volume	Volume of the journal	**Not_relevant**	Indicates exclusion due to article irrelevance (Yes/No)
Issue	Issue number of the journal	**Citation**	Full APA 7th edition citation for the article
DOI	Digital Object Identifier for the article	**Note**	Any special notes or further explanation of why the article was excluded
Z	Z-score if reported in the article	**n**	Sample size (used for calculating degrees of freedom)
p	p-value if reported in the article	**Upper**	Upper bound of a reported confidence interval
t	t-statistic if reported in the article	**Lower**	Lower bound of a reported confidence interval
chi	Chi-square statistic if reported in the article		

Z-Curve Implementation:

We formatted the extracted data using the zcurve_data() function, which converts textual representations like “p=0.05” into the format required by zcurve(). After preparing the data, we ran the Z-curve analysis and generated annotated plots to visualize the results.

Challenges and Limitations:

- **Irrelevant or Inaccessible Articles:** Many papers turned out to be non-empirical or descriptive, offering no testable hypotheses. Despite our institutional access, several relevant papers were behind exorbitant paywalls.
- **Incomplete Reporting:** Several articles failed to report results for non-significant findings. Others avoided providing precise p-values, instead stating thresholds (e.g., “p<0.05”). While Z-curve can handle the thresholds, this reduced the precision of our analysis.
- **Identifying Relevant Statistics:** It was sometimes unclear which of the reported statistics addressed the study’s main hypotheses versus exploratory analyses or manipulation checks.
- **Human Error:** Our manual review and recording of relevant statistics introduced the possibility of mistakes–such as overlooking relevant tests or including irrelevant ones.
- **Imbalanced Data:** Some studies included multiple hypotheses and subgroup analyses, generating many p-values, while others contained a single test. As a result, a small subset of articles contributed the majority of our data, giving disproportionate weight to a few sources. This can be seen in Figure 2 below.

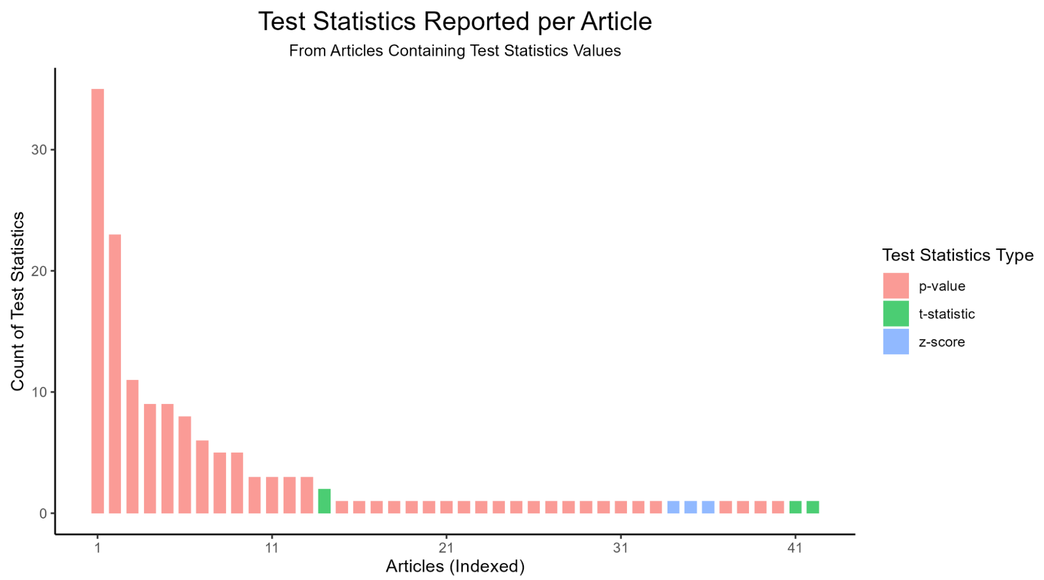

Figure 2: 42 out of the 90 total articles collected contain test statistics, and of those 42 the distribution of the number of reported test statistics is displayed by type. Code for this graph found in test_statistics_hist.qmd in Github.

The first three challenges reduced our sample size, making our results less robust than we would have liked. While Z-curve can handle censored p-values, there is necessarily a loss of precision when putting the resulting z-scores into bins. This means that as the number of censored p-values increased in our data, the precision of our results, especially at the tails, was reduced. The last challenge, too, affected our results. Since a large proportion of our p-values were taken from just a few papers, our overall results largely depended on the behaviour of a few researchers and studies. It is conceivable that if another set of papers were used in our analysis, the results could be quite different. To make our analysis less sensitive to the specific studies included, we would have to greatly expand our sample size.

Results:

After collecting our data and cleaning it appropriately, we ran the data through the zcurve() function and obtained the following plot.

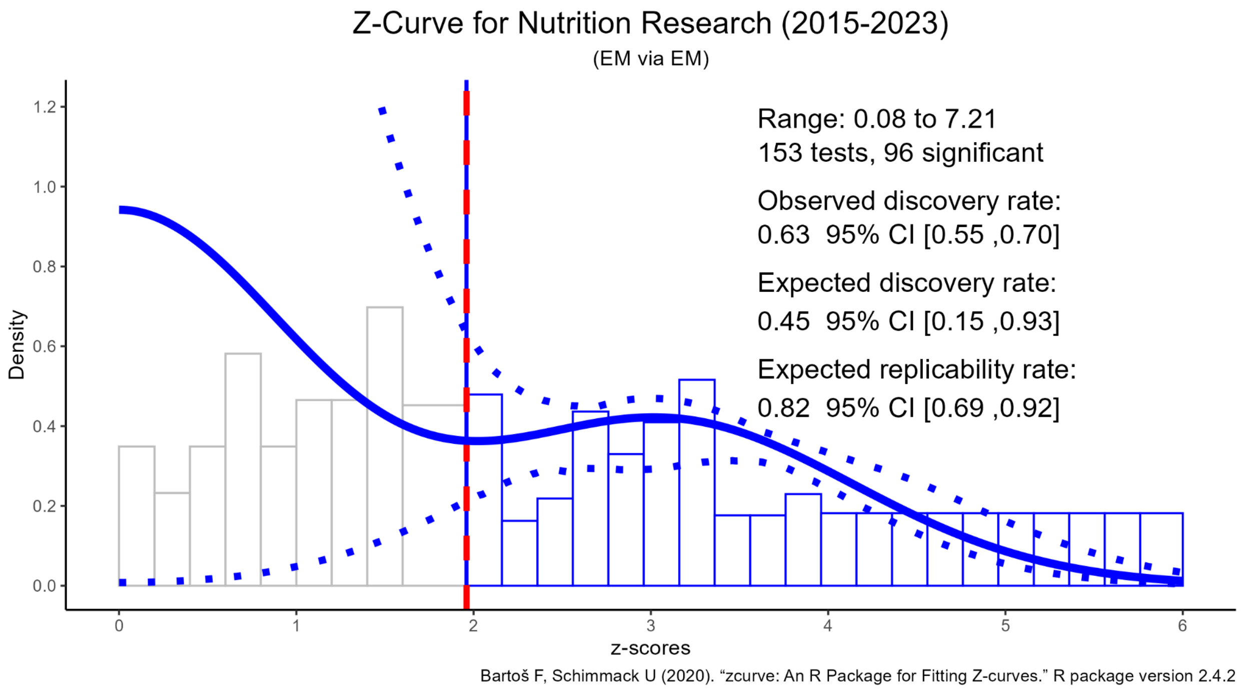

Figure 3: Our Z-curve package output when computing our dataset. Code found in final_results.qmd.

Figure 3 shows the distribution of z-scores obtained from our review of Nutrition research papers from 2015 to 2024. We can see that the resulting curve is rather flat with a distinct plateau around the red line at z=1.96 (corresponding to p=0.05). The most prominent feature of the graph is that the number of just significant results (to the right of the red line) and the number of just not significant results (to the left of the red line) are highly comparable. This symmetrical pattern is consistent with what one would expect from random sampling error rather than selective reporting. That is, rather than a steep drop-off in frequency, we see a smooth transition. This is in contrast to the first z-curve shown in the introduction, where there is a stark drop in the frequency of results immediately to the left of the red line. This indicates that there is little to no publication bias present in the data. If there were systematic selection for statistically significant results, this would show up visually as a steep slope from the left side of the line to the right.

There is yet more to be gleaned from this graph. In the upper right, we can see that the observed discovery rate (ODR) is 0.63, meaning that 63% of the reported results were statistically significant. This is intuitively consistent with a lack of publication bias. If there were systematic selection for significance, we would expect that a much higher percentage of reported results would be significant. This is further supported by the relative agreement between the observed discovery rate and the estimated discovery rate (EDR), which was found to be 0.45. Recall that the EDR estimates the proportion of statistically significant results if all studies were available (meaning none were left in the file drawer). While these estimates are very noisy, with very wide confidence intervals, there is a lot of overlap between the EDR and ODR. This overlap suggests that the proportion of significant results in all studies is comparable to the proportion of significant results in the published literature.

Finally, the estimated replication rate (ERR) of 0.82 is quite good. Recall that the ERR estimates the proportion of studies with significant results that, if replicated, would also yield significant results. In other words, we estimate that 82% of the studies would replicate successfully, yielding statistically significant results. This is quite good and in stark contrast to the first z-curve, where the presence of publication bias indicated that just over half of the studies would replicate successfully. The high ERR we found is due to the agreement between the ODR and EDR. If the rate of significance in published research is consistent with the estimated rate in all studies, we would expect fewer false positives in the literature. If there aren't many false positives in the research, we would expect most of them to replicate successfully, as the ERR we found indicates.

Conclusions:

Our project set out to estimate the extent of publication bias in recent nutrition research using Z-curve analysis. In our study, we did not find strong evidence of systematic selection for significance. There was considerable overlap between the ODR and EDR, indicating that the rate of significant results in the published literature is comparable to the rate of significance across all studies. This is what we would expect if there were little to no publication bias. The high ERR we obtained indicates that most of the studies would replicate successfully. These results should still be taken with caution. Since so much of our data came from a small number of sources, the results may be sensitive to the particular articles included.

Although we were unable to examine the frequency of numerical errors due to limitations of the statcheck package, our findings offer some reassurance about the integrity of nutrition science.

Z-curve proved to be a strong and flexible tool, particularly in its ability to handle censored p-values. However, it cannot detect computational or reporting errors. Future work in this area would benefit from more robust tools for detecting numerical inconsistencies. It would also be good to see comparisons of Z-curves for pre-registered with non-pre-registered ones.

Overall, our results suggest that nutrition science is difficult to conduct methodologically but may not suffer from the same degree of publication bias and p-hacking as other fields.

Works Cited:

Ritchie, S. (2021). SCIENCE FICTIONS : exposing fraud, bias, negligence and hype in science. Vintage.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. In Science (Vol. 349, Issue 6251). American Association for the Advancement of Science (AAAS). https://doi.org/10.1126/science.aac4716

Hickerson, A. (2024, August 22). How US dietary recommendations have changed over the last 50 years. Northwell Health. https://www.northwell.edu/news/the-latest/us-dietary-recommendations-changes-last-50-years

Ulrich Schimmack. (2020, January 10). Z-Curve.2.0. Replicability-Index. https://replicationindex.com/2020/01/10/z-curve-2-0/