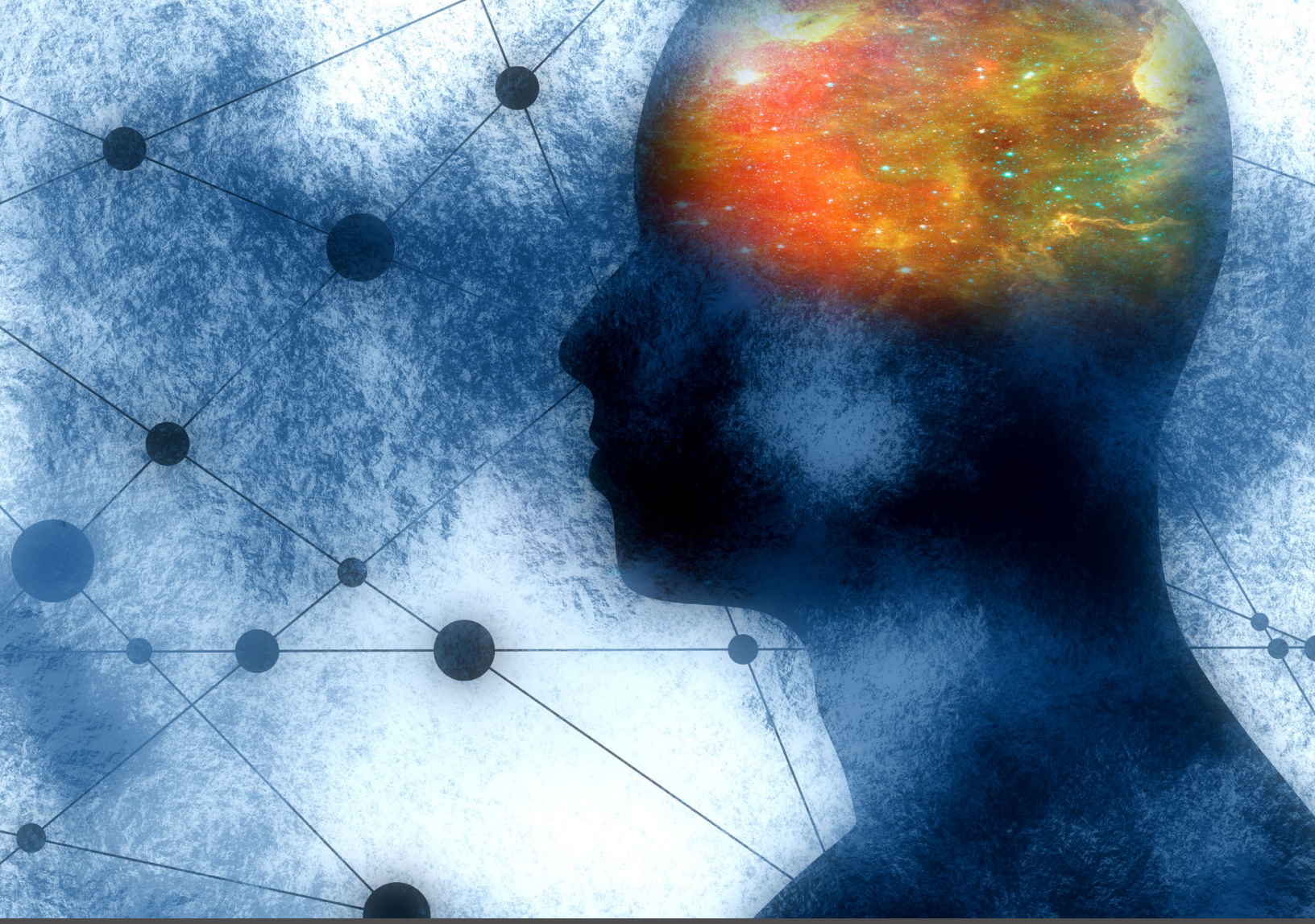

A CASE STUDY

Misreported Research in Psychology: A Text and Data Mining Project

A researcher at the University of California, Berkeley, uses TDM to analyze six prominent peer-reviewed psychology journals and quantify concerns raised in the replication crisis.

proquest.com

To talk to the sales department, contact us at
1-800-779-0137 or **sales@proquest.com**.

ProQuest®

Yuyang Zhong, an undergraduate researcher at the University of California, Berkeley (UC Berkeley), explores past research articles in six prominent peer-reviewed psychology journals to quantify the concerns raised in the recent replication crisis. Zhong is a student in the UC Berkeley Department of Psychology and the Division of Computing, Data Science, and Society.

Introduction

It is very common across disciplines of research like psychology, biology, and other sciences to use statistical significance to indicate the effectiveness of its models. Statistical significance is represented by a simple number called "p-value." Widely accepted as a statistical convention, a p-value greater than 0.05 means the relationship related to such value is not statistically significant. Conversely, a p-value less than 0.05 reflects a statistically significant relationship. P-value relates to "the probability something happens by chance" – a very small p-value means the more likely something did not happen by chance.

From its inception by Sir Ronald Fisher in the early 1920s, this simple diagnostic calculation became the center of scientific research. Researchers increasingly used p-value as the gold standard of finding "true" results. Statistical significance has, in many occasions, taken over real scientific finding, and this is deeply problematic and concerning. Over the past two decades, the field of psychology became increasingly aware of bad research being published, with some challenging the existing statistical framework (Ioannidis, 2005), and others failing to replicate a large amount of studies across the various psychology domain (Open Science Collaboration, 2015). Moreover, self-reports from thousands of researchers reveal troubling confessions of bad and unethical research practices, ranging from p-hacking (i.e., manipulating methods/variables to reach the desirable p-value), to modifying hypotheses to fit significant results, to falsifying data (John, Loewenstein, & Prelec 2012). Collectively, the field has arrived at a point of inflection, with mounting pressure to address the crisis of replication.

To date, there remain debates on how serious the replication crisis is, and how exactly researchers can circumvent the pitfall of bad science. No previous work, however, had been dedicated to quantify the magnitude of the issues by looking at how statistics were reported. This project does exactly that.

To this date, there remains debates on how serious the replication crisis is, and how exactly researchers can circumvent the pitfall of bad science.

With the support of our wonderful library staff at UC Berkeley and the American Psychological Association (APA), I secured special permission – on a pilot basis – to access a corpus of full-text research articles from six prominent academic journals[1] across psychology domains, ranging from 1985 to 2020, as well as pilot access to ProQuest's text and data mining (TDM) resource, TDM Studio. With the power of TDM techniques, this project set out to capture trends of how results are reported, with an emphasis on p-values. Part 1 of this project explored the time trend in the usage of specific p-value thresholds (e.g., $p < 0.001$, $p < 0.01$, $p < 0.05$). Part 2 of this project captured actual test statistics reported in addition to p-values, which allowed me to compare the reported p-values to ones recalculated from test statistics in order to identify the prominence of misreported results.

Feature Extraction

A Python-based script was set up to extract useful metadata tags from each of the article files in the corpus. Using Regular Expression (RegEx), a common tool used in text pattern matching and data cleaning, I extracted p-values as well as test statistics (i.e., F-statistic and t-score, commonly used in statistical models in psychology research) reported in each article.

For part one of this project, I extracted a total of 192,896 p-values from 6048 articles in the Journal of Personality and Social Psychology (JPSP), in which 166,041 were filtered out and used in analysis (i.e., only those with "<" or "=" signs, below 0.1). For part two of this project, I extracted a total of 212,589 F-test and t-test statistics from 13,220 articles across six journals. Since I was only interested in values reported significant with non-significant recalculation, I filtered out 167,261 p-values for analysis.

p-value Reporting Trends

Part one of this project revealed some interesting trends in how p-values are reported. Prior to the publication of the Open Science Framework paper in 2015, we can see visibly the use of specific thresholds in the blue curve — the distribution of all p-values reported from 1985 to 2015 — with visible bumps at each multiple of 0.01. The orange curve, which shows the distribution of p-values after 2016, is much smoother, though still showing a smaller, but apparent, bump at 0.05. It speaks to the trend that the field has moved towards using equality over inequality, allowing reporting of p-values as accurately as possible.

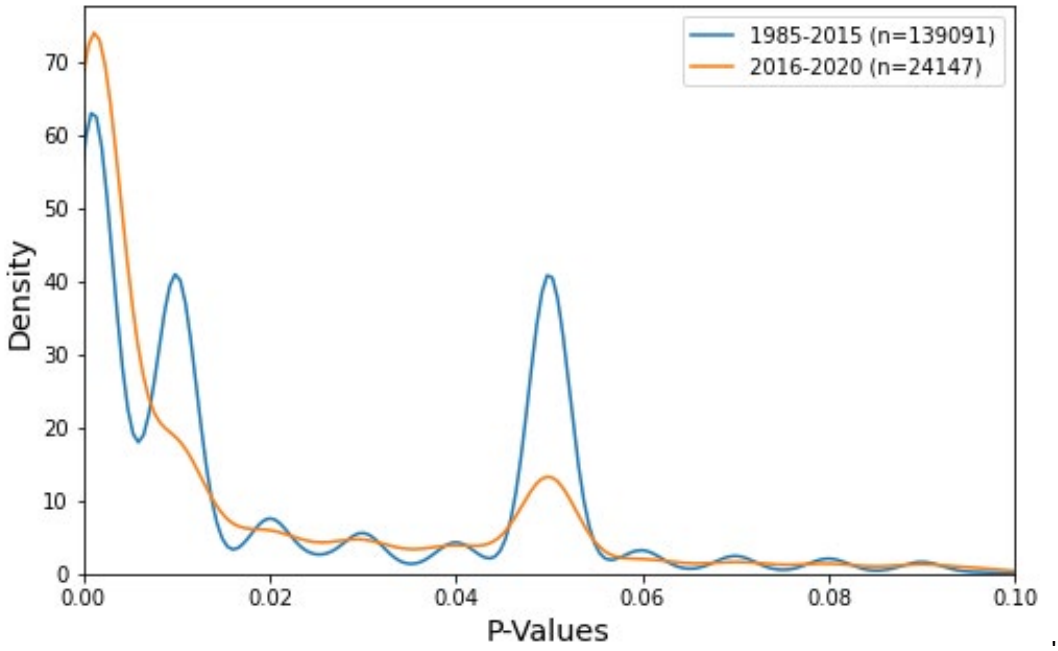

Figure 1. Kernel Density Estimation (smoothed estimate of the distribution) of p-values, segregated by prior to and after 2015. Visible smoothing over time is shown, speaking to the change in how p-value was reported.

Recalculating p-values

The second part of this project concerns the validity of p-values reported in journal articles. The recalculation of p-values showed expected but surprising results. Putting the sporadic scatter of points aside, we can see that the majority of reported p-values can be validated via recalculation from their generating test statistic. However, if we zoom into the region on the bottom left corner, we can see a decent number of points showing a significant reported value, but non-significant recalculation (i.e., on the top quadrant of the identity line).

From that perspective, I filtered out only reported p-values that are significant ($p < 0.05$), and specifically looked at the ones with non-significant recalculations. I found 2,800 values meeting these criteria, which means that about 1.67% of significant p-values captured in this process were actually invalid and false. Journal-wise calculation also showed similar results, with averages over all the years falling between 1% and 2%. Even though these numbers seem small, it is an alarming amount that signals falsification of statistical significance in peer-reviewed publications. Another concerning aspect seen in this plot is the scatter of points away from the line on the diagonal, which signifies even more misreporting when the results are non-significant in the first place.

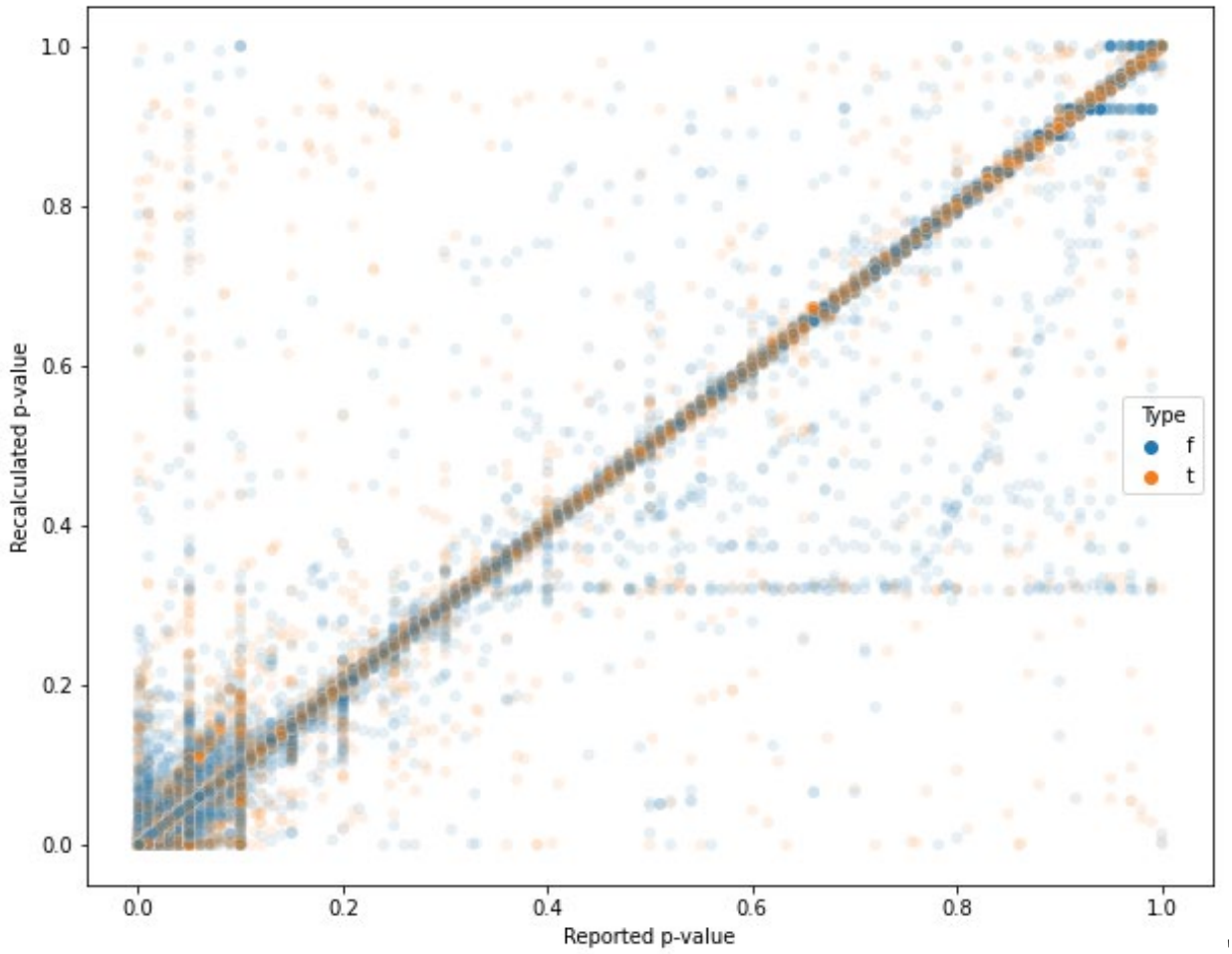

Figure 2. Scatter plot of pairs of recalculated versus reported p-values, capturing only those of F-test and t-test. In a perfect world, we would expect all points to fall on or around the identity line on the diagonal. However, this plot clearly shows abnormalities both around significant and non-significant areas.

Conclusion

With the help of TDM Studio, I was able to quantify a small but significant amount of misreporting research statistics. These range from calculation errors to rounding errors to blatant falsification of results. These results confirm the seriousness of the collective concerns of replication and reproducibility, which remains a crisis facing psychology and many other scientific disciplines.

This is not to mention that facts and science have been under attack throughout the COVID-19 pandemic, and there seems to be a loss of the "objective truth." Consequently, this project and its finding posits the field to take a small step towards finding and restoring that objectivity in what we do.

This project would not have been possible without the power of text data mining. One can only imagine the granularity of the work required should this project be conducted by a human, manually reviewing each article. The use of text mining tools like RegEx allowed for speedy and efficient extraction and analysis for a huge corpus of text data. TDM Studio provides streamlined access for projects similar in scope, allowing researchers like me to delve into historical text data in ways that would have been otherwise impossible.

This project would not been possible without the power of text data mining.

About TDM Studio

TDM Studio is a text and data mining solution for research across disciplines and enables researchers with or without knowledge of coding.

ProQuest's workflow solution for text and data mining is designed for research, teaching and learning. TDM Studio provides access to sought-after content including current and historical newspapers, primary sources, scholarly journals, and dissertations and theses. It empowers researchers, students and faculty to analyze documents by uncovering connections and patterns that lead to career-defining discoveries.

Learn more at www.proquest.com/go/tdm-studio
or contact your ProQuest representative today.

Footnote

1 *Journal of Personality and Social Psychology, Journal of Experimental Psychology: Learning, Memory, and Cognition, Developmental Psychology, Journal of Abnormal Psychology, Journal of Applied Psychology,* and *American Psychologist*

References

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. PLOS Medicine, 2(8), e124. https://doi.org/10.1371/journal.pmed.0020124

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices with Incentives for Truth Telling. Psychological Science, 23(5), 524–532. https://doi.org/10.1177/0956797611430953

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251). https://doi.org/10.1126/science.aac4716