Extraction and Evaluation of Statistical Information from Social and Behavioral Science Papers

Sree Sai Teja Lanka
The Pennsylvania State University
University Park, PA, USA
szl577@psu.edu

Sarah Rajtmajer
The Pennsylvania State University
University Park, PA, USA
smr48@psu.edu

Jian Wu
Old Dominion University
Norfolk, VA, USA
jwu@cs.odu.edu

C. Lee Giles
The Pennsylvania State University
University Park, PA, USA
clg20@psu.edu

ABSTRACT

With substantial and continuing increases in the number of published papers across the scientific literature, development of reliable approaches for automated discovery and assessment of published findings is increasingly urgent. Tools which can extract critical information from scientific papers and metadata can support representation and reasoning over existing findings, and offer insights into replicability, robustness and generalizability of specific claims. In this work, we present a pipeline for the extraction of statistical information (p-values, sample size, number of hypotheses tested) from full-text scientific documents. We validate our approach on 300 papers selected from the social and behavioral science literatures, and suggest directions for next steps.

ACM Reference Format:
Sree Sai Teja Lanka, Sarah Rajtmajer, Jian Wu, and C. Lee Giles. 2021. Extraction and Evaluation of Statistical Information from Social and Behavioral Science Papers. In *Companion Proceedings of the Web Conference 2021 (WWW '21 Companion), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3442442.3451363

1 INTRODUCTION

In parallel with major shifts toward transparency in scientific process and output, advances in information extraction and natural language processing have opened up new avenues for representation and reasoning over the large and growing scientific literature. Ideally, the community's confidence in a published finding is informed by a long and well-understood history of related work as greater scientific context. At present, this context is largely qualitative – gathered through keyword-based searches and investigator-led exploration of an ad-hoc sample of similar papers. But looking forward toward the vision of a queryable scholarly record, critical information and metadata can be extracted and aggregated to inform greater context for individual claims.

One critical piece of this context, particularly in hypothesis-driven work, is statistical information reported alongside a claim and associated hypothesis test(s), e.g., t-test, F-test, or chi-squared test. The most commonly reported piece of statistical information is the p-value, or the probability of obtaining a result at least as extreme as the observed result of a statistical hypothesis test, assuming that the null hypothesis is correct [7] [11]. In addition to p-values, studies typically report the test statistic and sample size, and may report other descriptive statistics of the dataset.

The work we present here builds on publicly-available statistical extraction software (Statcheck, [8]). We improve upon this tool, expanding the breadth of statistical tests considered and adding the extraction of sample size and number of hypotheses tested. The tool we have built ingests a scientific article in PDF and converts it to text, tokenizes sentences, and searches the text for specified regular expressions in order to output p-values, sample sizes and number of hypotheses present in the paper. We validate our approach on 300 papers selected from the social and behavioral science literature and offer a comparison of our tool to the Statcheck baseline.

This work falls into the broader category of mathematical formula extraction tools. For example, SymbolScraper[1] uses heuristic methods to extract symbol labels and bounding boxes from born-digital PDF files. The output is an XML file containing all symbols and their positions on each page. Although the extracted information is comprehensive, the patterns of statistical expressions become much less explicit and it is non-trivial to accurately restore those patterns at the symbol level. A learning based method was proposed in [12], which trained a CRF model to extract in-line mathematical expressions. The method achieved an $F_1 = 89\%$ on a corpus of manually annotated ACL papers. However, the authors used an in-house PDF analysis tool for data preparation, which is not publicly available. Recently, deep learning was applied to develop a mathematical formula extraction tool called ScanSSD [15]. ScanSSD outputs bounding boxes of math equations and images cropped from the input PDF file. However, recognizing text and symbols from the images requires additional OCR tools. The current model also extracts a fraction of false positives based on our qualitative assessments.[2]

WWW '21 Companion, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8313-4/21/04.
https://doi.org/10.1145/3442442.3451363

[1] https://github.com/zanibbi/SymbolScraper
[2] https://ws-dl.blogspot.com/2020/06/2020-06-05-math-formula-extraction-from.html

2 STATISTICAL FEATURES

In hypothesis-driven work, and sometimes in exploratory work, a statistical test is performed to offer evidence in support or refute of a null hypothesis. Common statistical tests include various t-tests, chi-squared tests, binomial tests, and ANOVA. Each of these outputs a test statistic, which measures how closely observed data matches the distribution expected under the null hypothesis of that test, or the assumption that no statistical relationship exists between two sets of observed data and measured phenomena. We focus our attention in this work on three critical statistical features of an empirical study: p-values, sample size and number of hypotheses tested (see Table 1 for examples).

Feature	Example Text
p-value w/o test statistic	p = 0.01, p < 0.03, p > 0.07
p-value w/ test statistic	t (10) = 1.3, p = 0.01
Sample Size	N = 100, n = 50
Hypothesis Test	t, z, F

Table 1: Statistical features and exemplar representations.

p-value. In testing a null hypothesis H_0 against an alternative hypothesis H_1 based on data x_{obs}, the p-value is defined as the probability, calculated under the null hypothesis, that a test statistic is as extreme or more extreme than its observed value. The null hypothesis is typically rejected — and the finding is declared statistically significant — if the p-value falls below the (current) type I error threshold $\alpha = 0.05$ [3]. More recently, concerns about process and purpose around p-values have highlighted the critical importance of context in interpreting statistical outcomes through this lens [1, 3, 19], further motivating the extraction of a richer set of statistical features from scientific documents.

Sample size. The sample size is the size of the observed dataset, or $|x_{obs}|$. In the social and behavioral science studies that were the focus of our tool during development, this was in many cases the number of participants in the study. The sample size of a study is critically important as an indicator of the power of a study and confidence in study outcomes (see, e.g., [9]).

Number of hypotheses tested. Understanding the number of hypotheses tested, whether or not they are explicitly described in a paper as such, is central to ongoing conversations about correct use of statistical methods and the direct attention being paid to p-hacking [10, 17] and related bad practices. Of particular concern is the use or lack thereof of appropriate tools for correction for multiple comparisons, e.g., Bonferroni [5] or false discovery rate (FDR, [18]). Put simply, the greater number of tests of the same hypothesis, the more likely that one of them will return a positive finding. Significance must be calibrated accordingly.

3 FEATURE EXTRACTION PIPELINE

Our tool represents a statistical feature extraction pipeline, which ingests PDF text, preprocesses that text, extracts statistical information using regular expressions, and synthesizes extracted information into meaningful statistical insights. Our pipeline integrates existing software for text extraction from PDF and sentence tokenization, and builds on initial extraction capabilites in Statcheck [8] to expand the statistical tests considered, add output of the sample size both through derivation and explicit extraction, and report the number of hypotheses tested.

3.1 Conversion to text

A necessary first step is to convert PDF to text, through encoding and decoding individual characters. Tools widely used to extract text from PDF include PDFBox,[3] XpdfReader,[4] PDFMiner,[5] and PyPDF2[6]. We use XpdfReader because it works well on bulk documents stored in a single folder. The conversion process is imperfect. On some occasions, the tool outputs missing characters, mismatches symbols or fails to extract text at all (see Figure 1). A particularly challenging task is the extraction of tables and figures, a problem of significant study in its own right, e.g., [6, 16].

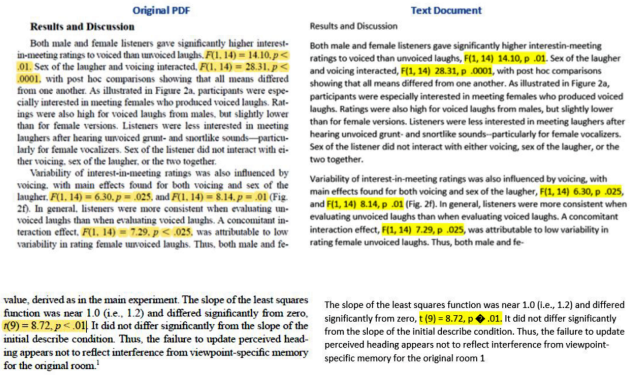

Figure 1: PDF to text conversion resulting in missing and mismatched symbols.

3.2 Sentence tokenization

Sentence tokenization is the process of splitting extracted text into individual sentences. For formal documents, the tokenization algorithms built in to spaCy, NLTK, etc, perform well since the tokenizer is trained on a corpus of formal english text. Many tokenizers perform less well for documents with extensive use of abbreviations, measurements, and other forms not found in standard written English [13, 14]. We have used NLTK [4].

3.3 Statistical feature extraction

As noted, we build on extraction capabilities initially deployed in the Statcheck tool [8]. Specifically, Statcheck uses regular expressions to find statistical results in the following forms:

> t(df) = value, p = value; F(df1,df2) = value, p = value; r(df) = value, p = value; [chi]2(df, N = value) = value, p = value (N is optional, delta G is also included); Z = value, p = value.

All regular expressions take into account that test statistics and p values may be exactly (=) or inexactly (< or >) reported.[7]

[3] https://pdfbox.apache.org/
[4] https://www.xpdfreader.com/
[5] https://pypi.org/project/pdfminer/
[6] https://pypi.org/project/PyPDF2/
[7] See Statcheck documentation: http://statcheck.io/documentation.php.

Statistical Test	Pattern	Example	Sample Size
T-Test	t(df) = float, p (<, >, =) float	t (12) = 4.3, p = 0.01	df + 1
F-Test	f(df1, df2) = float, p (<, >, =) float	f(21,30) = 2.3, p < 0.01	df1 + df2 + 1
Correlation	r(df) = float, p (<, >, =) float	r(32) = 12.2, p = 0.05	df + 2
Chi-Square	χ^2(df, N=int) = float, p (<, >, =) float	χ^2(10, 35) = 19.4, p > 0.03	N
Z-Test	Z = float, p (<, >, =) float	Z = 9.0, p = 0.05	-
Q-Test	Q(df) = float, p (<, >, =) float	Q(45) = 3.2, p < 0.01	-
Logistic Regression	OR = float, p (<, >, =) float	OR = 3.0, p = 0.00	-
b-test	b ((<, >, =)) float, p (<, >, =) float	b < 1.3, p < 0.3	-
d-test	d ((<, >, =)) float, p (<, >, =) float	d > 4.3, p > 0.1	-
Hazard Ratio	HR = float, p (<, >, =) float	HR = 4.3, p = 0.01	-

Table 2: List of patterns of reported p-values with test statistics extracted by our tool, and corresponding statistical tests.

Statistical Test	Regular Expression
T-Test	`t\s?(\[|\()\s?\d*\.?\d+\s?(\]|\))\s?[<>=]\s?[^a-z\d]{0,3}\s?\d*[,;]?\d*\.?\d+\s?[,;]\s?(([^-z]ns)|(p|P|Ps|ps|pvalue|p-value)\s*(<|>|=)\s*-?\d*\.*\d+(e(-|−)\d+)?)`
F-Test	`(F|F-change)\s?(\[|\()\s?\d*\.?(I|l|\d+)\s?,\s?\d*\.?\\d+\s?(\]|\))\s?[<>=]\s?\d*\.?\d+\s?[,;]\s?(([^a-z]ns)|(p|P|Ps|ps|pvalue|p-value)\s*(<|>|=)\s*-?\d*\.*\d+(e(-|−)\d+)?)`
Correlation	`r\s?\\(\s?\d*\.?\d+\s?\\)\s?[<>=]\s?[^a-z\\d]{0,5}\s?\d*\.?\d+\s?[,;]\\s?(([^a-z]ns)|(p|P|ps|pvalue|Ps|p-value)\s*(<|>|=)\s*-?\d*\.*\d+(e(-|−)\d+)?)`
Chi-Square	`((\\[CHI\\]|\\[DELTA\\]G)\s?|(\s[^trFzQWBn]\s?)|([trFzQWBn ]2\s?))2?\\(\s?\d*\.?\d+\s?(,\s?N\s?\\=\s?\d*\\,?\d*\\,?\d+\s?)?\\)\s?[<>=]\s?\s?\d*,?\d*\\.?\d+\s?,\s?(([^a-z]ns)|(p|P|Ps|ps|pvalue|p-value)\s*(<|>|=)\s*-?\d*\.*\d+(e(-|−)\d+)?)`
Z-Test	`[^a-z]z\s?[<>=]\s?[^a-z\d]{0,3}\s?\d*,?\d*\.?\d+\s?,\s?(([^a-z]ns)|(p|P|Ps|ps|pvalue|p-value)\s*(<|>|=)\s*-?\d*\.*\d+(e(-|−)\d+)?)`
Q-Test	`Q\s?-?\s?(w|within|b|between)?\s?\\(\s?\d*\.?\d+\s?\\)\s?[<>=]\s?[^a-z\d]{0,3}\s?\d*,?\d*\.?\d+\s?,\s?(([^a-z]ns)|(p|P|Ps|ps|pvalue|p-value)\s*(<|>|=)\s*-?\d*\.*\d+(e(-|−)\d+)?)`
Logistic Regression	`[OR|or|oR|Or]\s?\s?[<>=]\s?[^a-z\\d]{0,5}\s?\d*\.?\d+\s?[,;]\s?(([^a-z]ns)|(p|P|Ps|ps|pvalue|p-value)\s*(<|>|=)\s*-?\d*\.*\d+(e(-|−)\d+)?)`
b-test	`b\s*[=><]\s*\d*\.*\d*\s*,\s*(p|P|Ps|ps|pvalue|p-value)\s*(<|>|=)\s*-?\d*\.*\d+(e(-|−)\d+)?`
d-test	`d\s*[=><]\s*\d*\.*\d*\s*,\s*(p|P|Ps|ps|pvalue|p-value)\s*(<|>|=)\s*-?\d*\.*\d+(e(-|−)\d+)?`
Hazard Ratio	`HR[\s*|=]\d*\.*\d*,\s*(.*,(.*(p|P|Ps|ps|pvalue|p-value)\s*(<|>|=)\s*-?\d*\.*\d+(e(-|−)\d+)?))`

Table 3: Regular expressions representing p-values reported alongside test statistics, and associated statistical tests.

We extend these representations as follows. We use regular expressions to extract similar information for the following additional statistical tests: Q-test, Logistic Regression, b-test, d-test and Hazard Ratio. In addition, we build a more extensive list of regular expressions to better capture reporting of the tests described above and part of then original Statcheck tool, namely, t-test, F-test, correlation, Chi-square, and Z-test (see Table 2). A listing of regular expressions used to extract p-values reported alongside test statistics, for each of these tests, is given in Table 3.

Critically, we also consider p-values reported without an accompanying test statistic. Differentiation between these two classes of reported p-values is important for downstream interpretation of extracted information. For example, the presence or absence of a test statistic alongside a p-value directly informs our evaluation of number of hypotheses tested (see below). For the extraction of p-values reported without associated test statistics, we consider expressions of the form: p (>,<,=) float, P (>,<,=) float, ps (>,<,=) float, Ps (>,<,=) float, pvalue (>,<,=) float, p-value (>,<,=) float. Float here includes scientific notation, e.g., p = 1.7e+3.

Sample size. We identify sample sizes, in parallel, in two ways. First, we derive sample size from test statistics where possible, through back-calculation based on degrees of freedom (see Table 2). Second, we search for direct mention of sample size using regular expressions of the form 'n or N = int', following a similar approach to that taken for p-value extraction.

Number of hypotheses tested. We extract the number of hypotheses tested indirectly from the paper by making use of the extracted p-values and associated test statistics when reported (see

Text Extraction	p-value Representation	Count in PDF	Count from Extractor	Accuracy
Manual eval after conversion to text	w/o test statistic	626	573	91.5%
	w/ test statistic	673	561	83.3%
Our model	w/o test statistic	626	565	90.2%
	w/ test statistic	673	532	79.0%
Statcheck	w/o test statistic	626	538	85.9%
	w/ test statistic	673	467	69.3%

Table 4: Accuracy of p-value extractors based on a manually-labelled count in original PDF documents. We compare extraction in three ways: 1. manual extraction of p-values from document after conversion from PDF to text; 2. with the tool presented in this paper; 3. with Statcheck. For additional resolution, we separate analyses of p-values reported along with a test statistic and those without.

	p-value w/o test statistic		p-value with test statistic		sample size	
Test Metric	pdf	text	pdf	text	pdf	text
Precision	0.76	0.74	0.99	0.99	0.66	0.65
Recall	0.93	0.98	0.70	0.93	0.98	1.0
F1	0.83	0.85	.82	0.96	0.79	0.79

Table 5: Performance metrics for the extraction of p-values with and without accompanying test statistics, and sample size on both original PDF and text (converted) documents.

Table 1 for an example of p-values reported with and without an associated test statistic). Specifically, we count the number of p-values reported alongside a test statistic as proxy for the number of hypotheses (statistically) tested in the paper.

4 VALIDATION

To validate our pipeline, we consider a dataset of 300 papers, 30 each randomly selected from prominent journals in the following 10 social and behavioral science fields: Economics; Health; Education; Political Science; Marketing; Criminology; Psychology; Sociology; Management; and Public Administration. We manually label each p-value and sample size reported in each of the 300 papers, and track the number of p-values reported with and without accompanying test statistics. This manual labelling is done, for each paper, for both PDF and converted text documents to facilitate in depth evaluation.

A report of extraction accuracy is provided in Table 4. We compare three approaches: manual extraction of p-values from converted text; our full pipeline model; the Statcheck tool. Accuracy is calculated based on total number of p-values extracted over the dataset using the given approach vs the total number of p-values present in the original PDFs. Our approach meaningfully improves on the Statcheck tool. Accuracy metrics reported on the text-extracted documents indicate the critical importance of the conversion to text process. In particular, we observe during our labelling that statistical information is often captured in tables, where conversion is particularly prone to error.

Table 5 gives the Precision, Recall and F1 performance metrics for our model for both the extraction of p-values (with and without test statistic) and the extraction of sample size from both an original PDF document and text (obtained after conversion). We note that precision of our sample size extractor is relatively lower, indicating that our approach looking for instances of 'n or N = int' is overly inclusive. Somewhat lower precision for p-values reported without test statistics is similarly attributed. Recall was high for all three information categories, but relatively lower for extraction from PDF than from text. These scores also suggest that accuracy of our tool would be improved with more accurate text extraction.

5 CONCLUSION

We have presented a pipeline for the extraction of statistical information from full text PDF of scientific documents, and validated our tool on a set of 300 papers from the social and behavioral science literatures. Motivating this work is ongoing concern about the reproducibility and generalizability of published claims, which emerged in the social sciences but has since left nearly no empirical field untouched [2]. It is clear that meta-reasoning over a body of literature could provide critically important framing for results of an individual study and move the community to more efficient discovery. Yet, manual search, extraction and assembly of statistical information across corpora will not scale. Rather, computational tools to support this process are needed.

Our work points to some specific next steps for extraction of statistical information from scholarly work. We have noted that tools which can better extract information from tables and figures will be particularly useful, as statistical information is often embedded in these formats. In addition, the section of a paper in which statistical information is reported may add relevant context, as may language around the statistical result. Mining text around extracted statistical information is proposed as a valuable future direction, both for the aim of refining statistical information extraction and for supplementing extracted statistics with investigator interpretations.

REFERENCES

[1] Amrhein, V., Greenland, S., and McShane, B. Scientists rise up against statistical significance, 2019.

[2] Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature News 533*, 7604 (2016), 452.

[3] Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., et al. Redefine statistical significance. *Nature human behaviour 2*, 1 (2018), 6–10.

[4] Bird, S. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions* (2006), pp. 69–72.

[5] Bonferroni, C. E. Il calcolo delle assicurazioni su gruppi di teste. *Studi in onore del professore salvatore ortu carboni* (1935), 13–60.

[6] Clark, C., and Divvala, S. K. Pdffigures 2.0: Mining figures from research papers. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, JCDL 2016, Newark, NJ, USA, June 19 - 23, 2016* (2016), pp. 143–152.

[7] Dahiru, T. P-value, a true test of statistical significance? a cautionary note. *Annals of Ibadan postgraduate medicine 6*, 1 (2008), 21–26.

[8] Epskamp, S., and Nuijten, M. statcheck: Extract statistics from articles and recompute p values (r package version 1.0. 0.).

[9] Faber, J., and Fonseca, L. M. How sample size influences research outcomes. *Dental press journal of orthodontics 19*, 4 (2014), 27–29.

[10] Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., and Jennions, M. D. The extent and consequences of p-hacking in science. *PLoS Biol 13*, 3 (2015), e1002106.

[11] Ioannidis, J. P. The proposal to lower p value thresholds to. 005. *Jama 319*, 14 (2018), 1429–1430.

[12] Iwatsuki, K., Sagara, T., Hara, T., and Aizawa, A. Detecting in-line mathematical expressions in scientific documents. In *Proceedings of the 2017 ACM Symposium on Document Engineering, DocEng 2017, Valletta, Malta, September 4-7, 2017* (2017), K. P. Camilleri and A. Bonnici, Eds., ACM, pp. 141–144.

[13] Lobur, M., Romanyuk, A., and Romanyshyn, M. Using nltk for educational and scientific purposes. In *2011 11th international conference the experience of designing and application of CAD systems in microelectronics (CADSM)* (2011), IEEE, pp. 426–428.

[14] Long, A. Benchmarking python nlp tokenizers, Sep 15, 2019.

[15] Mali, P., Kukkadapu, P., Mahdavi, M., and Zanibbi, R. Scanssd: Scanning single shot detector for mathematical formulas in PDF document images. *CoRR abs/2003.08005* (2020).

[16] Siegel, N., Lourie, N., Power, R., and Ammar, W. Extracting scientific figures with distantly supervised neural networks. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, JCDL 2018, Fort Worth, TX, USA, June 03-07, 2018* (2018), J. Chen, M. A. Gonçalves, J. M. Allen, E. A. Fox, M. Kan, and V. Petras, Eds., ACM, pp. 223–232.

[17] Simmons, J. P., Nelson, L. D., and Simonsohn, U. Life after p-hacking. In *Meeting of the society for personality and social psychology, New Orleans, LA* (2013), pp. 17–19.

[18] Verhoeven, K. J., Simonsen, K. L., and McIntyre, L. M. Implementing false discovery rate control: increasing your power. *Oikos 108*, 3 (2005), 643–647.

[19] Wasserstein, R. L., and Lazar, N. A. The asa statement on p-values: context, process, and purpose, 2016.