Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 August 2016 doi:10.20944/preprints201608.0191.v1

Peer-reviewed version available at *Data* **2016**, *1*, 14; doi:10.3390/data1030014

Data Descriptor

688,112 Statistical Results: Content Mining Psychology Articles for Statistical Test Results

Chris H. J. Hartgerink

Department of Methodology and Statistics, Tilburg University, Warandelaan 2,
5037AB Tilburg, The Netherlands; c.h.j.hartgerink@tilburguniversity.edu; Tel.: +31-134-664-126

Simple Summary: A dataset of 688,112 statistical test results reported according to the standards prescribed by the American Psychological Association (APA), mined from 50,845 articles out of 167,318 published by the APA, Springer, Sage, and Taylor & Francis. Mining from Wiley and Elsevier was actively blocked. Metadata for each article are included. All journals included, scripts, etc. are available at https://github.com/chartgerink/2016statcheck_data and preserved at http://dx.doi.org/10.5281/zenodo.59818

Abstract: In this data deposit, I describe a dataset that is the result of content mining 167,318 published articles for statistical test results. As a result of this content mining, 688,112 results from 50,845 articles were extracted. In order to provide a comprehensive set of data, the statistical results are supplemented with metadata from the article they originate from. The dataset is provided in a comma separated file (CSV) in long-format. For each of the 688,112 results, 20 variables are included, of which seven are article metadata and 13 pertain to the individual statistical results (e.g., reported and recalculated *p*-value). A five-pronged approach was taken to generate the dataset: (i) collect journal lists, (ii) spider journal pages for articles, (iii) download articles, (iv) add article metadata, and (v) mine articles for statistical results.

Data Set: http://dx.doi.org/10.17026/dans-zsk-k335

Data Set License: CC-0 1.0 Universal

Keywords: nhst; p-values; apa; content mining; tdm; errors

1. Summary

In this data deposit, I describe a dataset that is the result of content mining 167,318 published psychology articles for statistical test results. I tried to mine the content of HTML articles in all journals published by the six major publishers in psychology, and succeeded in doing so for four major publishers (see Table 1 for descriptives per publisher). This mining was done with the R package `statcheck` [1,2], which automatically mines the content in research articles to extract statistical test results, given that they are reported in the format prescribed by the American Psychological Association (APA). Considering this style is prescribed for psychology, only psychology journals were inspected.

Table 1. An overview of the publishers included accompanied by descriptive statistics per publisher regarding the extracted APA results.

Publisher	Timespan	Nr. of articles	Nr. of articles with results	Result count	Median results per article	Mean reported *p*-value	Mean recalculated *p*-value
American Psychological Association (APA)	1985-2016	74,489	36,662	522,367	9	.073	.098
Sage	1972-2016	13,893	5,118	59,561	8	.101	.110
Springer	2003-2016	53,667	8,333	97,657	8	.097	.113
Taylor & Francis	2003-2016	25,274	732	8,527	8	.118	.133
Total	1972-2016	167,318	50,845	688,112	9	.080	.102

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 August 2016 doi:10.20944/preprints201608.0191.v1

Peer-reviewed version available at Data 2016, 1, 14; doi:10.3390/data1030014

As a result of this content mining, 688,112 results from 50,845 articles (out of the 167,318) were extracted. The resulting dataset contains the extracted statistical test results in long format (i.e., each row corresponds to one statistical result), where all individual variables are available in separate columns for easy analysis. Reported statistical values are simultaneously used to recalculate the p-value for the reported statistical result, which is checked against the reported p-value for (decision) errors. A potential error has occurred when the reported p-value is not congruent with the recalculated p-value, whereas a decision error (or gross error) occurs when the recalculated p-value does not correspond to the reported p-value and alters the significance of the result, assuming $\alpha = .05$. The results of this comparison are available in the dataset. The articles for which no results were found are not included in this dataset (filenames without results available at https://raw.githubusercontent.com/chartgerink/2016statcheck_data/master/no_result.txt).

In order to provide a comprehensive dataset, the statistical results are supplemented with metadata of the original article. These metadata include the `doi`, the publisher, the publication year, the journal, the author names, the author count, and the publication title. Given that the dataset is in long format, multiple rows can contain duplicate metadata if the results are from the same article.

This dataset of statistical results and accompanying metadata can be used to inspect if specific papers included potential statistical errors and for trends in statistical results over time. Articles based on a similar dataset inspected the degree to which reporting errors occur [1], tried to assess whether such data could be modeled for p-hacking [3], and the degree to which sample sizes and potential false negative results developed over time [4]. This dataset can be used to replicate these findings in a larger dataset and correlate findings with the available metadata. These data can also be used as baseline data to identify extreme statistical results in the literature by determining their percentile score, or to replicate other meta-research. These are only a few examples, and "the best thing to do with [the] data will be thought of by someone else" (quote from Rufus Pollock).

2. Data description

The data are provided in a comma separated file (CSV) in long-format, where each row contains one statistical result. As such, multiple rows can (and most likely will) pertain to the same psychology article and include the same metadata. This information is provided in duplicate because any other file format (wide-format or separate files per article) is unfeasible due to the size of the dataset without increasing the threshold to reuse the data (e.g., in JSON format). Given the size of the full dataset (>200MB), a smaller test dataset is also included to trial run analyses scripts.

For each of the 688,112 results, 20 variables are included, of which seven are article metadata and 13 pertain to the individual statistical results. The variables included in the dataset are explained in Table 2. Two specific sets of variables are worth explaining further. First, only F-values have two degrees of freedom (i.e., *df1* and *df2*). For t-values, the reported degrees of freedom are *df2*, because $t^2(df) = F(1, df)$. For all other test statistics that include degrees of freedom, they are included in *df1* (i.e., χ^2, r; Z contains no degrees of freedom, hence, no *df1* or *df2* is reported). Second, the variable *DecisionError* indicates whether an error results in wrongly concluding statistical significance (report $p < .05$ whereas the recalculated p-value yields $p > .05$, or vice versa). If the variables *OneTail* and *OneTailedInTxt* are `TRUE` (see Table 2), a decision error is reverted to `FALSE`.

3. Methods

A five-step approach was taken to generate the dataset: (i) collect journal lists, (ii) spider journal pages for articles, (iii) download articles, (iv) add article metadata, and (v) mine articles for statistical results. These five steps are specified below. All code and version history is available at https://github.com/chartgerink/2016statcheck_data. Figure 1 gives a flowchart of the different steps in the data collection process.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 August 2016 doi:10.20944/preprints201608.0191.v1

Peer-reviewed version available at Data 2016, 1, 14; doi:10.3390/data1030014

Table 2. Variables included in the dataset and a description of each variable.

Variable	Type	Description
Source	Metadata	Digital Object Identifier (DOI) of the article
publisher	Metadata	Publisher of the article, as available in CrossRef
year	Metadata	Publication year, as available in CrossRef
journal	Metadata	Journal, as available in CrossRef
Statistic	Individual result	Type of statistical test statistic (possible values t, F, r, Z, and χ^2)
df1	Individual result	First degree of freedom of the test statistic
df2	Individual result	Second degree of freedom of the test statistic
Test.Comparison	Individual result	Sign used in reporting of test statistic (>, <, =)
Value	Individual result	Reported value of the test statistic
Reported.Comparison	Individual result	Sign used in reporting of *p*-value (>, <, =)
Reported.P.Value	Individual result	Reported *p*-value
Computed	Individual result	Recalculated *p*-value (two-tailed) based on Statistic and df1, df2
Raw	Individual result	Raw text of extracted statistical result
Error	Individual result	Whether the reported *p*-value differs from recalculated *p*-value
DecisionError	Individual result	Whether the reported *p*-value differs from the recalculated *p*-value AND significance is different ($\alpha = .05$)
OneTail	Individual result	Whether the result would be correct if the *p*-value were one-tailed
OneTailedInTxt	Individual result	Whether the article contains "sided", "tailed", or "directional"
authors	Metadata	Author names, as available in CrossRef
author_count	Metadata	Number of authors
title	Metadata	Title, as available in CrossRef

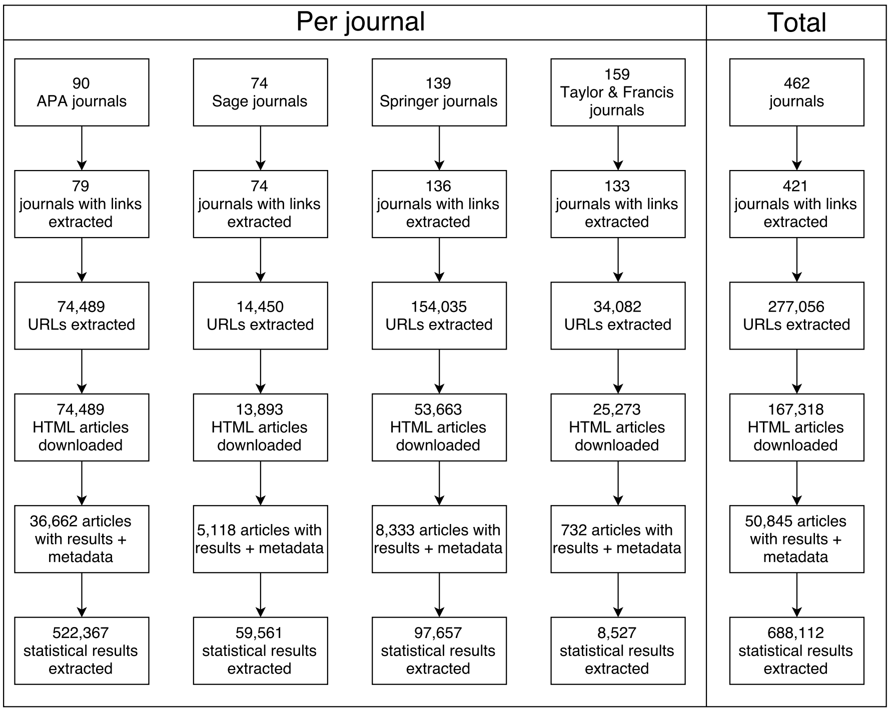
Per journal
Total
90 APA journals
74 Sage journals
139 Springer journals
159 Taylor & Francis journals
462 journals
79 journals with links extracted
74 journals with links extracted
136 journals with links extracted
133 journals with links extracted
421 journals with links extracted
74,489 URLs extracted
14,450 URLs extracted
154,035 URLs extracted
34,082 URLs extracted
277,056 URLs extracted
74,489 HTML articles downloaded
13,893 HTML articles downloaded
53,663 HTML articles downloaded
25,273 HTML articles downloaded
167,318 HTML articles downloaded
36,662 articles with results + metadata
5,118 articles with results + metadata
8,333 articles with results + metadata
732 articles with results + metadata
50,845 articles with results + metadata
522,367 statistical results extracted
59,561 statistical results extracted
97,657 statistical results extracted
8,527 statistical results extracted
688,112 statistical results extracted

Figure 1. Flowchart of the data collection process, with specific.

A list of psychology journals from six major publishers was collected manually. Six publishers were included at the start of this project: Elsevier, Wiley, Sage, Springer, Taylor & Francis, and the APA. Collectively these publishers cover >70% of the published psychology literature [5]. For multidisciplinary publishers, only the journals included in the 'Psychology' or 'Behavioral Sciences' sections were included (as categorized by the publishers themselves). These journal lists were collected in October 2015 and available at https://github.com/chartgerink/2016statcheck_data/blob/master/scraping/journal-spiders/journal_list_old.csv.

The journal list of six publishers was forced down to four publishers, because Elsevier and Wiley prevented me from automatically downloading research articles [6–8]. The library at my university was prompted by these publishers that suspicious downloading activity occurred, which they thought indicated compromised user credentials and theft of copyrighted material. The Tilburg University library services requested me to halt the automated downloading, in light of potential blocks for the entire university. As a result, Elsevier and Wiley were

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 August 2016 doi:10.20944/preprints201608.0191.v1

Peer-reviewed version available at Data 2016, 1, 14; doi:10.3390/data1030014

excluded from the journal list, resulting in a remainder of 461 journals from the original 1011 (this renewed list is available at https://github.com/chartgerink/2016statcheck_data/blob/master/scraping/journal-spiders/journal_list.csv).

Article URLs were collected with a web spider in April 2016. A web spider visits a webpage and collects all or a specific set of URLs included on that webpage. Subsequently, the web spider visits the pages that were referred on the initial webpage and again collects URLs, which it can repeat over and over. For this project, a web spider was developed to extract specific links that referred to full texts (https://github.com/chartgerink/journal-spiders). This web spider produced a set of URLs, which provided direct links to full-text articles in HTML format (all URLs available here: https://github.com/chartgerink/2016statcheck_data/tree/master/scraping/journal-spiders/journal-links). Only those HTMLs that were accessible within the Tilburg University subscription were collected (available journal titles within subscription: https://github.com/chartgerink/2016statcheck_data/blob/master/tilburg_journals.ods?raw=true). The original sample, including Elsevier and Wiley, was ~900,000 articles.

The research articles were subsequently downloaded automatically, with the command-line utilities `wget` (i.e., APA articles) and `quickscrape` (v0.4.6; i.e., Sage, Springer, Taylor & Francis; https://github.com/contentmine/quickscrape). This downloading occurred in April-May 2016 and was done taking into account potential strain on the publisher's servers, restricting downloads to weekends or limiting the download rate to 10 per minute at most (but typically fewer).

Metadata for each article were collected with the Ruby module `terrier` (https://github.com/thewinnower/terrier). This module queries the CrossRef database when provided with a Digital Object Identifier (DOI). If available, it returns the available metadata such as the journal name, publication year, etc. These metadata were collected in April-June 2016 for all included articles. For specific implementations of the `terrier` queries, please view https://github.com/chartgerink/2016statcheck_data/blob/master/scraping/terrier.rb. Not all articles contained a DOI and no metadata could be collected from CrossRef as a result.

Finally, after all HTML files were collected and metadata were added, `statcheck` (v1.0.2; https://github.com/michelenuijten/statcheck; [1,2]) was run in August 2016 to create the final dataset. This `R` package scans the text from an article for APA style statistical results, which it then extracts and processes in order to check whether the reported p-values are equivalent to the recalculated p-value (with a margin of error due to potential rounding). For example, the result $t(85) = 2.86, p = .005$ would be automatically extracted. Version 1.0.2 of `statcheck` is able to mine t, F, r, Z, and χ^2 results, which consequently are the results included in this dataset. Table 2 explains each variable.

4. Usage notes

Usage of the data requires understanding several limitations of the `statcheck` package, in order to provide context for results obtained from this dataset. `statcheck` proved valid for extracting results it is supposed to extract (i.e., APA style reported test results; [1]), but it does not extract results that are not in line with what the APA prescribes. Additionally, `statcheck` only extracts results reported in the text and not those reported in tabular format or in images. As such, statistical results from tables and images are systematically excluded. As a result, any conclusions based on this dataset should not be extrapolated without caution.

Additionally, it is worth mentioning that relatively few articles contained results that were extracted by `statcheck` (~1/3 downloaded articles). This could be due to at least three reasons. First, results are not consistently reported according to the APA format in psychology journals, resulting in fewer extracted results. Second, statistical results might be reported in APA format, but these statistical results are not t, F, r, Z, or χ^2. Third, a considerable part of the literature might not pertain to research applying statistics, such as theoretical papers, case studies, or narrative reviews.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 August 2016 doi:10.20944/preprints201608.0191.v1

Peer-reviewed version available at Data 2016, 1, 14; doi:10.3390/data1030014

The presented data collected with `statcheck` have been deposited in the Dutch Archival Network for the Sciences (DANS) and are available under a public domain license (CC-0). The DANS repository is a trustworthy digital repository and has received the Data Seal of Approval (DSA), the World Data System (WDS) certificate, and the NESTOR-seal. This ensures that deposited data will remain available for a substantial amount of time. All rights to this dataset are waived to the furthest extent possible, such that reuse is maximized. As such, if you have questions about reuse, be reassured that whatever you want to do with the dataset will not be restricted by any kind of copyright, or other Intellectual Property (IP) framework.

In addition to preserving the data in the DANS repository, individual reports have been generated for each of the 50,845 articles and posted on PubPeer (https://pubpeer.com/). Appendix A shows a fictitious example of such a report. These reports were generated in order to increase the accessibility of the data for those wanting to investigate a specific paper instead of the entire dataset. Additionally, this increases the discoverability of potential errors by posting them in a central forum of post-publication peer review.

Acknowledgments: No funding was received in creating the materials underlying this article or to cover the publication costs. The publication costs were waived by MDPI.

Conflicts of Interest: C.H.J.H. helped develop the `statcheck` package. The author declares no other competing interests.

Appendix Example of statcheck report for PubPeer

Using the R package statcheck (v1.0.2), the HTML version of this article was scanned on 2016-06-11 for statistical results (t, r, F, Chi2, and Z values) reported in APA format (for specifics, see Nuijten et al., 2015). An automatically generated report follows.

The scan detected 5 statistical results in APA format, of which 3 contained potentially incorrect statistical results, of which 1 may change statistical significance ($\alpha = .05$). Potential one-tailed results were taken into account when 'one-sided', 'one-tailed', or 'directional' occurred in the text.

The errors that may change statistical significance were reported as:

$t(67) = -.436$, $p < .001$ (recalculated p-value: 0.66424)

The errors that may affect the computed p-value (but not the statistical significance) were reported as:

$F(1, 126) = 2.1$, $p > .90$ (recalculated p-value: 0.14978)

$t(67) = -1.02$, $p = .35$ (recalculated p-value: 0.31140)

Note that these are not definitive results and require manual inspection to definitively assess whether results are erroneous.

Reference Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2015). The prevalence of statistical reporting errors in psychology (1985-2013). Behavior Research Methods. http://dx.doi.org/10.3758/s13428-015-0664-2

Bibliography

1. Nuijten, M.B.; Hartgerink, C.H.J.; van Assen, M.A.L.M.; Epskamp, S.; Wicherts, J.M. The prevalence of statistical reporting errors in psychology (1985-2013). *Behavior research methods* **2015**.
2. Nuijten, S.E..M.B. *statcheck: Extract statistics from articles and recompute p values*, 2015. R package version 1.0.2.
3. Hartgerink, C.H.J.; van Aert, R.C.M.; Nuijten, M.B.; Wicherts, J.M.; van Assen, M.A.L.M. Distributions of p-values smaller than .05 in psychology: what is going on? *PeerJ* **2016**, *4*, e1935.
4. Hartgerink, C.H.J.; van Assen, M.A.L.M.; Wicherts, J.M. Too good to be false: Nonsignificant results revisited. osf.io/qpfnw, 2016. Accessed: 2016-6-25.
5. Larivière, V.; Haustein, S.; Mongeon, P. The Oligopoly of Academic Publishers in the Digital Era. *PloS one* **2015**, *10*, e0127502.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 August 2016 doi:10.20944/preprints201608.0191.v1

Peer-reviewed version available at *Data* **2016**, *1*, 14; doi:10.3390/data1030014

6. Hartgerink, C.H.J. Elsevier stopped me doing my research. http://onsnetwork.org/chartgerink/2015/11/16/elsevier-stopped-me-doing-my-research/, 2015. Accessed: 2016-5-17.
7. Bloudoff-Indelicato, M. Text-mining block prompts online response. *Nature News* **2015**, *527*, 413.
8. Hartgerink, C.H.J. Wiley also stopped me doing my research. http://onsnetwork.org/chartgerink/2016/02/23/wiley-also-stopped-my-doing-my-research/, 2016. Accessed: 2016-5-17.

© 2016 by the author; licensee *Preprints*, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (http://creativecommons.org/licenses/by/4.0/).