Literature Mining For Common Statistical Errors

Randy Ryker
Department of Computer Information Systems and Business Administration

Chuck Viosca
Department of Management and Marketing

Nicholls State University
Thibodaux, LA 70310

Abstract

Researchers from most academic disciplines use statistical tools to analyze experimental and observational data. Unfortunately there is evidence that a disturbing number of peer-reviewed articles that use statistical methods use them incorrectly. Prior research has found this to be a problem in journals dedicated to a variety of fields including medicine, psychology, business and education. Currently, the identification of common statistical errors in journal articles involves a manual and thus very labor-intensive process. As a result, only a small number of articles in a few disciplines from a limited set of journals have been assessed. In this paper we propose developing text mining models to partially automate the identification of common statistical errors in journal articles. We use the term literature mining to refer to the application of text mining tools to academic journals. Once developed, such models can be used to audit the existing journal-based body of knowledge and rapidly compile lists of articles that likely contain a particular statistical error. Looking forward, the models can also be used by editors and reviewers to screen papers prior to publication. The identification of new research opportunities is also discussed.

Introduction

Modern statistical tools have become a product of science whose influence on public and private life is pervasive. They are used to analyze experimental and observational data in fields ranging from the natural and social sciences, to medicine, education, and business. These modern statistical tools have virtually all been developed during the last century (Porter 1986), and during that time many millions of peer-reviewed studies that use these tools have been published.

Consider for a moment all peer-reviewed journal articles published during the past century, regardless of discipline, year or language. Conceptually, from a big data perspective, one may view this entire collection as a large database, with each article representing a record. As academics, we rely on this database as both a foundation and as a source of building blocks for our research. Practitioners, as consumers of research, also rely on this database and depend on the findings and recommendations being valid.

However, there is a wealth of evidence that a substantial portion of the peer-reviewed articles that use statistical methods use them incorrectly (Glantz 1980, Altman 1998, Shott 2011, Fabrigar, et al. 1999, Schor and Karten 1966, Ryker and Nath 1997, Strasak, et al. 2007, Patil, et al. 2008). Medical researchers in particular have engaged in a manual review process to bring this problem to light. For example, one early study consisted of a review of 295 articles from 10 medical journals. The authors found that only 28% of the papers that used statistical methods used them correctly (Schor and Karetn 1966). Altman (1998) summarized numerous additional studies of this kind. Table 1 is used here only to emphasize that these types of manual reviews have been conducted for decades. For full references to the papers in Table 1, see Altman (1998).

Table 1. Summary of Reviews of the Quality of Statistics in Medical Journals, Showing the Percentage of 'Acceptable' Papers (of those using statistics) Source: Altman, 1998. Note: Full references to papers in this table can be found in Altman, 1998.

Year Published	First Author	Number of Papers	Number of Journals	% Papers Acceptable
1966	Schor	295	10	28
1977	Gore	77	1	48
1979	White	139	1	55
1980	Glantz	79	2	39
1982	Felson	74	1	34
1982	MacArthur	114	1	28
1983	Tyson	86	4	10
1985	Avram	243	2	15
1985	Thorn	120	4	<40
1988	Murray	28	1	61
1988	Morris	103	1	34
1995	McGuigan	164	1	60
1996	Welch	145	1	30

Interestingly, it is not the studies that use the more complex statistical procedures that is the problem. Those papers typically have a professional statistician as a co-author, or at a minimum one was consulted prior to submitting the work to a journal (Schor and Karten 1966). Once received, they are routinely sent to professional statisticians for review and are frequently *not* in need of correction. On the other hand, studies using only a few probability values and lacking statistical jargon are typically not reviewed by statisticians, and it is these articles that, all too often, are published with common statistical errors (Schor and Karten 1966). The focus of our research is on the latter set of studies, those that do not use complex statistical procedures.

In this paper, we propose using the relatively new and flexible tools of text mining to partially automate the identification of common statistical problems in journal articles. The first section below describes the nature of the problem. That is followed by a section which describes what has been recommended and what has been tried to address the problem. The field of text mining is then introduced and is followed by a section on how to use text mining tools to partially automate the identification of specific statistical errors in journal articles. The final sections discuss challenges/opportunities, and conclusions.

Nature of the Problem

The nature of the problem concerns the statistical errors that are most common in journal articles. Although the majority of studies identifying common statistical errors have been conducted in the field of medicine (Glantz 1980, Lang 2003, Prescott and Civil 2013, Strasak, et al. 2007), other fields such as psychology (Hayton, et al. 2004), business (Patil, et al. 2008, Cashen and Geiger 2004, Ryker and Nath 1997) and education (Daniel 1998) have also been examined. Lang articulated one of the strongest descriptions of the problem:

> The problem of poor statistical reporting is, in fact, long-standing, widespread, potentially serious, and almost unknown, despite the fact that most errors concern basic statistical concepts and can be easily avoided by following a few guidelines (Lang 2003).

Common problems range from relatively harmless presentation issues to more substantial concerns, such as the misuse of statistical tests. Numerous presentation problems have been identified and these should be the easiest to identify and prevent (Prescott and Civil 2013). One example of a presentation issue is the over-precise reporting of percentages. For example, if we have 36 responses from a sample of 70 subjects, this is 51.42% and authors that are too precise report it as such. A response from one additional subject would change the percentage by 1.43%. So, when interpreting this percentage, a reader cannot rely on either decimal place. In a paper, it would be best reported as 36/70 (51%) (Prescott and Civil 2013). Another issue with percentages is when authors do not report the actual numbers on which the percentages are based. Similarly, authors are often too precise in reporting means, standard deviations and standard errors. A general rule of using no more than one additional decimal place than in the original measurement will typically suffice, but common sense should prevail (Prescott and Civil 2013). Authors should report only to a precision that means something.

Although presentation problems may seem to be relatively harmless, good research deserves to be well presented and sound presentation should be as much a part of the research as the collection and analysis of the data (Evans 1989). For other common presentation problems see Strasak, et al. (2007) and Lang (2003).

The misuse of statistical tests is a more serious problem. The frequencies of these types of statistical mistakes vary somewhat by discipline. In medicine, for example, the Student's t-test is the most popular statistical procedure (Feinstein 1974), and several researchers have documented problems with its use (Glantz 1980, Lang 2003, Strasak, et al. 2007). On the other hand, several business disciplines and their reference discipline of psychology rely more on survey research. One of the most common statistical mistakes found in these journals involves the inefficient use of Exploratory Factor Analysis (Hayton, et al. 2004, Patil, et al. 2008, Ryker and Nath 1997). A description and discussion of all common statistical mistakes is beyond the scope of this paper. For lists of them see Lang (2003), Glantz (1980), and Shott (2011). The focus of this paper is on using text mining to efficiently identify these errors.

For illustration purposes, we examine the misuse of one statistical procedure: the Student's t-test. The t-test is used to compute the probability of being wrong (the p-value) when asserting that the mean values of two groups are different. A common cut-off point is $p < .05$. Such a cut-off point indicates that less than 5% of the time would a researcher erroneously conclude a treatment had an effect when in reality it did not. Statisticians refer to this type of erroneous conclusion as a Type I error. The test is also widely but inappropriately used to test for differences among more than two groups by comparing all possible pairs of means with t-tests.

For example, suppose a researcher randomly assigned subjects with high blood pressure to three groups: a control group (no drug), a group administered drug A and a group administered drug B. It is common to find examples in the literature where they perform three t-tests on these data: one to compare the control to drug A, one to compare the control to drug B, and one to compare drug A to drug B. It is also common for the researcher to assert that there is a significant difference between any of the groups when $p < .05$. This practice is incorrect because when performing multiple t-tests the probability of a Type I error occurring is considerably higher than .05. The overall error rate is calculated as $1 – (1 - .05)^k$, where k is the number of comparisons (Ott and Longnecker 2001). Therefore, if three comparisons were made the probability of finding a falsely significant result from any one of the comparisons increases from .05 to .14. If more groups are involved, the probability of false significance increases even more. A useful rule of thumb to use when estimating the true p-value is to multiply the reported p-value times the number of possible t-tests (Glantz 1980).

When comparing multiple groups, the use of a generalized analysis of variance is often the best choice (Ott and Longnecker 2001). However, the use of multiple t-tests with a p-value adjustment, e.g. Bonferroni, would still be valid. In the example used in this paper, we propose the development of a text mining model that can identify articles that contain multiple t-tests and further classify those into ones that used a p-value adjustment and those that did not.

What Has Been Recommended/Tried

Some of the researchers who identified common statistical errors in the literature also made recommendations as to how these errors could be prevented. For example, a variety of statistical reviewing checklists and guidelines have been developed (Prescott and Civil 2013, Altman 1998, Schulz, et al. 2010, Moher, et al. 2009, von Elm, et al. 2008). Several of these checklists are specific to certain areas. Guidelines for how to conduct and report randomized controlled trials are covered in CONSORT (Schulz, et al. 2010). Meta-analysis standards are addressed in PRISMA (Moher, et al. 2009) and STROBE is specific to observational data (von Elm, et al. 2008). The Council of Science Editors website is a good reference for these and other guidelines (Council 2014). Although statistical checklists and guidelines have been developed for practical use by researchers, reviewers and editors, they may also serve as a source of patterns to look for when using text mining to search for statistical errors.

In addition to creating new guidelines, the editors of some journals have also revised their review process. Initial reviews are carried out as in the past by subject-area experts. Once a manuscript passes that review and is likely to be accepted, a professional statistician is asked to review the paper to determine if statistics are used properly (Glantz 1980, Strasak, et al. 2007). This prolongs the reviewing process, but the editors believe the delay is justified as it better ensures that proper statistical analysis and conclusions appear in their journals. An important caveat to this approach is that it may not be practical for many journals for two reasons. First, many journals already find it difficult to recruit professional statisticians as reviewers (Altman 1998). Second, most professional (and even amateur) statisticians are not especially interested in reviewing papers just to identify garden-variety statistical errors (Glantz 1980). As proposed in this paper, using text mining models to identify the common statistical errors would be an efficient way to address this problem for many journals and free up professional statisticians to review papers with more complex statistical procedures.

Another recommendation to solve the problem is to improve the statistical education of graduate students. Before advanced topics are covered, students should be thoroughly knowledgeable about how to use and present the most common statistical procedures for their respective disciplines. We do not doubt that some universities are quite diligent with regard to rigorous statistical training. However, there does not seem to be any practical way to set standards for quality statistical education across the wide variety of institutions and disciplines that use statistics. In addition, one set of researchers has suggested that it seems unlikely that departments will be able or willing to invest the resources necessary to substantially upgrade their statistics courses, (Fabrigar, et al. 1999).

Some educators have taken a different approach by writing articles to better inform practitioners about common statistical errors (Shott 2011, Lang 2003). They do this so that practitioners may protect themselves from the harm that may result when invalid study results are accepted and applied (Shott 2011). Shott articulated this approach well:

> Most of the reports published in veterinary journals are not reviewed by statisticians, and veterinary reviewers cannot always determine whether appropriate statistics were used. For these reasons, veterinarians need to critically evaluate the statistics in the reports they read. This may appear to be an impossible (and repellent) task. However, many statistical issues are much simpler than they appear. A reader who knows how to apply a few basic statistical concepts can detect most of the major statistical errors in veterinary reports, (Shott 2011).

Finally, Fabrigar, et al. (1999) urged methodologists to accept a greater responsibility for educating the research community by writing less technical papers that more clearly explain the practical implications of their methods. In addition, they pointed out that editors must be willing to publish such articles in non-quantitative journals that are much more likely to be read by most researchers.

Despite the various recommendations and efforts to improve the quality of statistics in peer-reviewed research, relatively recent evidence indicates that the problem persists (Prescott and Civil 2013, Shott 2011, Patil, et al. 2008, Strasak, et al. 2007, Hayton, et al. 2004, Lang 2003). The proposal in this paper, to use text mining models to identify common statistical errors, addresses the problem in two ways. First, text mining models may be used to audit journals for the presence of such errors. To date, these kinds of audits are rare because they involve a manual, labor-intensive process. Second, such models may be used by reviewers to screen papers prior to publication, and perhaps used by researchers to check their statistics prior to submission. We next present a brief introduction of the rather large field of text mining.

Text Mining

Text mining is a relatively new area of research that is about 20 years old. The field includes many techniques from numerous disciplines that focus on finding patterns in unstructured data. Due to the diverse contributions from various disciplines, text mining can mean different things to different authors, vendors, speakers, and clients. A recent taxonomy developed by Miner, et al. (2012) helps one to visualize the diversity of this new field of research. At the bottom of Figure 1, some of the disciplines that have contributed to the development of text mining techniques are listed. The top part of Figure 1 displays the seven "practice areas" of text mining. They are referred to as practice areas because they are based on the perspective of a text mining practitioner. Coverage of all the techniques and practice areas is beyond the scope of this paper. The one practice area we focus on is Classification.

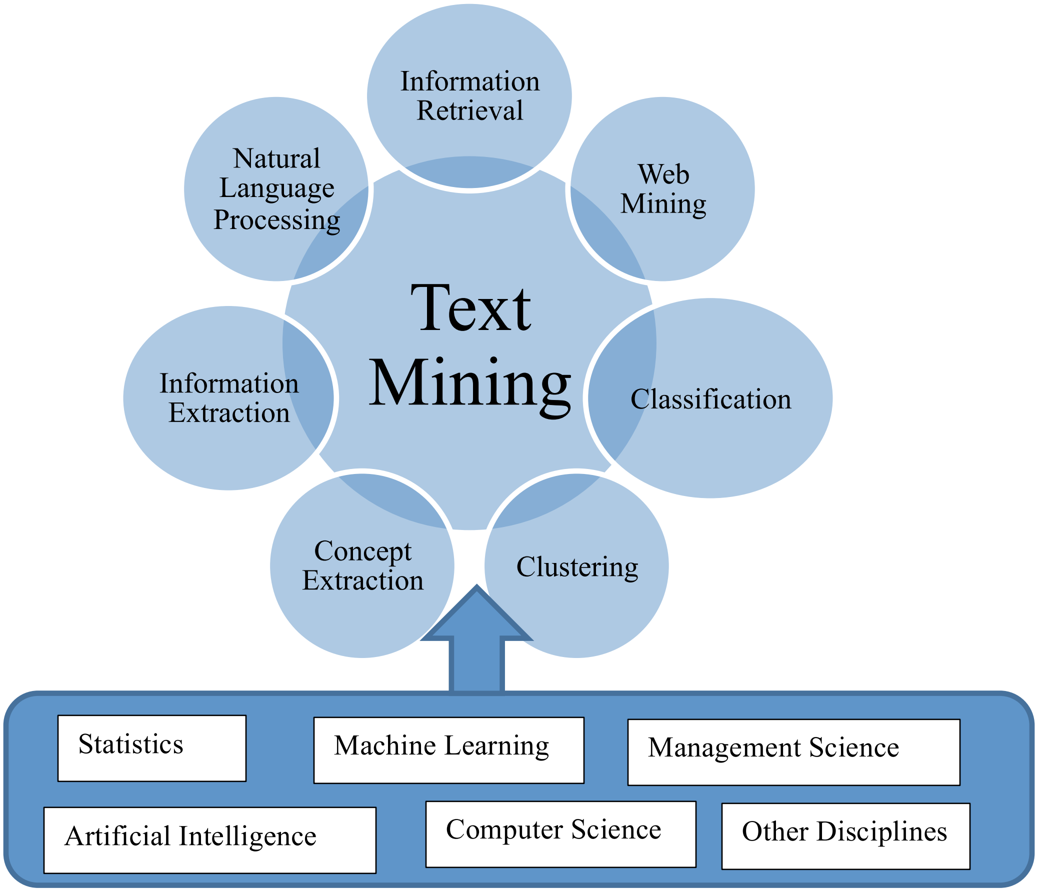

Figure 1. Taxonomy of Text Mining Practice Areas. Source: Miner, et al. 2012.

The use of text mining by business analysts is on the rise. Marketers, for example, use classification techniques to determine which of the thousands of comments on a company's Facebook page are positive versus negative. The process is often referred to as sentiment analysis. Once isolated, the negative comments can then be addressed by customer service representatives. A similar process can be used for other social media data such as the clustering and sentiment analysis of tweets from Twitter.

Another business example involves the field of finance. Analysts have used text mining on Reuters news articles to automatically classify articles as either dealing with earnings or not (Manning and Schutze 2002). This is a not a trivial task given the hundreds of thousands of articles available via Reuters. The general utility of these methods that enables one to automatically classify very large numbers of texts into categories (e.g., of interest or not of interest) is impressive. Once an accurate classification model has been developed, thousands of human work hours may be saved by using text mining.

Scientists are also using text mining with the goal of finding patterns of interest in journal articles. Researchers in the field of bioinformatics in particular have been active in conducting text mining. They look for relationships between genes and diseases, and also the side effects of drugs and new drug applications (Smit and van der Graaf 2011). Scientists often refer to this process as literature mining.

For an expansive list of how these tools are used in various business disciplines see the web sites of the main commercial vendors. The three most common data mining tool sets for business are IBM SPSS Modeler Premium, SAS Enterprise Miner and Text Miner, and STATISTICA Data Miner and Text Miner. There are also open-source text mining tools available. However, those new to text mining will likely find the commercial versions more user-friendly than the freely available open-source tools.

In the section below, we discuss how to use IBM SPSS Modeler Premium to identify articles that contained multiple t-tests and to further classify these into ones that used a p-value adjustment and those that did not.

Building a Text Mining Model

Figure 2 describes both the training corpus and the goals for the model. We begin by describing the training corpus. First, one would collect a set of 100 articles that have been manually reviewed and verified as having used multiple t-tests. These would be labeled as known positives (KP) because they used multiple t-tests. One half of the 100 articles will also be verified as having used a p-value adjustment, e.g. Bonferroni, and be labeled as KP for containing a p-value adjustment. The other half will be verified as not having used a p-value adjustment, and be labeled as KP for failing to use a p-value adjustment. In addition to these 100 articles, 900 other articles will be manually reviewed and verified as not having used multiple t-tests. These will be labeled as know negatives (KN). The total corpus will consist of 1000 articles. The goal is to build a model that can separate the corpus into categories as shown in Figure 2.

Next, we describe how each of the categories can be formed and assessed. The software uses a variety of algorithms including Natural Language Processing to automatically extract features which include "interesting" words and phrases from the full text of the articles (IBM SPSS Modeler 2014). The software terminology refers to these features as concepts.

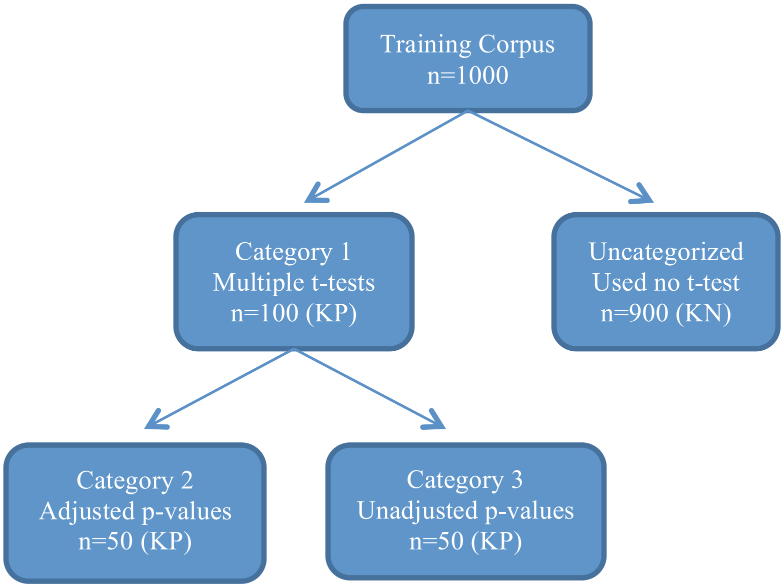

Figure 2. Training Corpus and Goals for the Model

The first step to forming Category 1 is to input the 100 known positive articles that used multiple t-tests and instruct the software to automatically extract concepts. The default setting returns the most frequent 5000 concepts that can be sorted either by frequency or alphabetically. Researchers, using their statistical knowledge of a variety of terms and phrases for which to look, select from the 5000 extracted concepts those which are most likely to appear in articles that used multiple t-tests and through a drag-and-drop interface assign the concepts to Category 1.

To assess the category, the entire training corpus of 1000 articles is used as input to the model and the results analyzed. The technical term for the training approach used by IBM SPSS Modeler Premium is semi-supervised, because the full training corpus contains both known positives and known negatives. This approach is between a supervised approach that uses only known positives as input, and an unsupervised approach that uses only unlabeled data as input. A semi-supervised approach is useful in a case like this where obtaining a large corpus of only known positive articles can be both time consuming and expensive (Chapelle, et al. 2006). Model performance is based on precision and recall. Precision is a measure of the accuracy of the model and recall is a measure of the inclusiveness of the model.

$$\text{Precision} = TP/(TP + FP)$$

$$\text{Recall} = TP/(TP + FN)$$

Where TP stands for true positive, FP stands for false positive, and FN is false negative.

For example, let's assume we run the corpus of 1000 articles through the model and 90 articles are classified as belonging to Category 1. Of these, 80 were TP and 10 were FP, giving a precision of 89%. The model missed 20 articles which are FN, giving a recall of 80% (Table 2).

KP	KN	Total	Category 1	TP	FP	FN	Precision	Recall
100	900	1000	90	80	10	20	89%	80%

Table 2. Example Data from a Training Session

$$\text{Precision} = 80/(80 + 10) = 89\%$$

$$\text{Recall} = 80/(80 + 20) = 80\%$$

The next step is to analyze the FP and FN articles to identify areas for model improvement. To increase the recall we may need to identify and include additional concepts describing Category 1. To increase the precision we may need to delete one or more concepts from Category 1. Once additions or deletions of concepts for the category are performed, the entire corpus of 1000 is again run through the model and the results analyzed.

With enough iteration it is typical to develop a model that works well with the training corpus, providing a good balance between recall and precision, with high scores for both. The goal is to create models that are general enough to apply to new content almost as well as to the training content.

Once a model demonstrates acceptable performance in classifying articles into Category 1, we can proceed to creating Category 2. The concepts for Category 2 will include all of those from Category 1 plus additional concepts associated with p-value adjustments, e.g. Bonferroni, Tukey, and Scheffe. The concepts from Category 1 and the new p-value adjustment concepts can be combined with a category rule that uses an (AND) Boolean operator. The results in Category 2 should only include articles that used multiple t-tests and also used a p-value adjustment. The training of the model as described earlier would continue until acceptable precision and recall was achieved for both Category 1 and Category 2. Note, the software allows records, in this case articles, to appear in more than one category.

The third category is of course the one we are primarily interested in, i.e. articles that used multiple t-tests but did not use a p-value adjustment. Category 3 will contain all concepts from Category 1 and use a Boolean operator (NOT) to exclude all concepts related to p-value adjustments. Again the iterative training process will be used in an effort to achieve acceptable precision and recall for all three categories.

We suggest setting the acceptable threshold at 80% for both precision and recall. Although precision is set at 80% we would strive for 100% precision for the overall process by performing a quick manual verification that is easily facilitated by the software. Thus we recommend that all Category 3 articles be manually reviewed and verified as having used multiple t-tests without a p-value adjustment. Such manual verification is possible using the software's point-and-click interface. From a list of Category 3 articles, a user can click on a link and the full text of the article will appear in an adjacent window. The terms and phrases that were employed to identify the article will have been automatically highlighted by the software, in a user-specified color. Using this tool, we expect it to require only a few minutes per article to manually verify the use of multiple t-tests and the absence of a p-value adjustment.

The 80% goal for recall is sufficiently high to make the model useful in determining the scope of the incorrect use of multiple t-tests within a very large set of articles that may include decades of research across numerous journals.

Once the model demonstrates acceptable performance with the training corpus, an equivalent testing corpus would need to be procured and used to test the model for precision and recall with the new data.

Assuming the model demonstrates acceptable performance on the new data, the final step is to deploy the model. For example, we can run the model using as input all articles from all journals of interest, and identify those articles that used multiple t-tests without a p-value adjustment and thus may have come to incorrect conclusions.

Challenges and Opportunities

A significant challenge to using text mining tools for literature mining is acquiring a corpus of articles that contain the patterns in which a researcher is interested. This is because doing so is such a labor-intensive manual process that requires a skilled human agent. Cooperation from authors who have conducted manual reviews of articles with statistical problems would be very helpful. If they would, upon request, share lists of specific articles they identified with particular problems, this would save much of the initial research time required to procure the corpus required for model training and testing.

Publishers of journals must decide how they will address any problems that are revealed. Because editors tend to have ready access to a digital version of their journals, they may decide to be proactive and use a previously developed and published text mining model like the one proposed here to identify potential problems. One suggestion is for them to then provide a page on their websites where corrections and clarifications are published so that these are clearly visible to their readers. Although not involving text mining, this approach to dealing with errors has been recommended for literature in the medical field (Majeed 2012).

Significant additional research opportunities exist that can build upon the proposal in this paper. Once a text mining model has been developed that can identify the incorrect use of multiple t-tests, it can be used to review journals from any discipline that uses statistics. In addition, once articles are identified as containing this particular statistical error, *each* of those articles becomes a potential new study. One may choose to sort such articles in descending order by the number of times they have been cited in other research. Those with the most citations may be viewed as higher priority research opportunities.

Following the approach as presented here, individual text mining models can be developed to identify other common statistical errors. The general process is to first classify articles that use a particular statistical method. One must then further classify those articles into ones that used the method correctly and those which likely used the method incorrectly. The general process depends upon the statistical method of interest being related to unique terms and phrases that describe its use in articles. It further depends upon the correct uses of the method being related to unique terms and phrases. Using these terms and phrases in combination with Boolean operators, the articles can be classified as to whether they used the method correctly or not. One limitation that should be noted is that if researchers do not explicitly mention one or more of these terms and/or phrases in their published articles, then the models may misclassify them.

We see this paper as the first in a stream of research dealing with text mining for common statistical errors. One effort the authors are currently engaged in is using text mining to identify the inefficient use of Exploratory Factor Analysis in academic articles. This too is a long-standing common statistical problem in the literature (Hayton, et al. 2004, Patil, et al. 2008, Ryker and Nath 1997). For lists of other common statistical errors see Lang (2003), Glantz (1980), Patil, et al. (2008), Cashen and Geiger (2004), and Shott (2011).

The proposal described here involves the use of IBM SPSS Modeler Premium. Other commercial software and freeware exists that incorporate a wide range of algorithms and thus capabilities. Perhaps one or more of these other software platforms would prove to be even more useful in classifying statistical errors. We encourage interested researchers to experiment with additional approaches.

Conclusions

We have discussed, from a practical standpoint, the major steps required to develop a text mining model that can be used to find a common statistical error in journal articles. Once developed, such a model can be used to audit the existing journal-based body of knowledge to compile lists of articles that likely contain a particular statistical error. Each article that is found to contain a statistical error has the potential to become a new research opportunity.

Looking forward, the models can also be used by editors or reviewers to screen papers prior to publication. The authors believe the proposal in this paper is a significant contribution in that it is an innovative and potentially very efficient way to address a long-standing problem. We hope that it will reach an audience of interested readers and generate both discussions and contributions on the topic.

This study points the way to improving the statistical quality of the journal-based body of knowledge. One thing is certain: the volume of peer-reviewed research that is published worldwide will continue to grow, in some fields exponentially. Using text mining to partially automate the identification of some of the common statistical errors will help to reassure both academics and practitioners of the quality of peer-reviewed research.

References

Altman, Douglas, 1998. Statistical Reviewing for Medical Journals. *Statistics in Medicine* 17(23): 2661-2674.

Cashen, Luke and Scott Geiger, 2004. Statistical Power and the Testing of Null Hypotheses: A Review of Contemporary Management Research and Recommendations for Future Studies. *Organizational Research Methods* 7(2): 151-167.

Chapelle, Oliver, Bernhard Scholkopf, and Alexander Zien. Semi-Supervised Learning. (Cambridge, MA: MIT Press, 2006).

Council of Science Editors. Accessed from http://www.councilscienceeditors.org/i4a/pages/index.cfm?pageid=3637 on February 7, 2014.

Daniel, L., 1998. Statistical Significance Testing: A Historical Overview of Misuse and Misinterpretation with Implications for the Editorial Policies of Educational Journals. *Research in the School*s 5(2): 23-32.

Evans, M., 1989. Presentation of Manuscripts for Publication in the British Journal of Surgery. *British Journal of Surgery* 76(12): 1311-1314.

Feinstein, A.R., 1974. Clinical Biostatistics XXV. A Survey of the Statistical Procedures in General Medical Journals. *Clinical Pharmacology Therapeutics* 15(1): 97-107.

Fabrigar, Leandre, Duane Wegener, Robert MacCallum and Eri Strahan, 1999. Evaluating the Use of Exploratory Factor Analysis in Psychological Research. *Psychological Methods* 4(3): 272-299.

Glantz, Stanton, 1980. Biostatistics: How to Detect, Correct, and Prevent Errors in the Medical Literature. *Circulation* 61(1): 1-7.

Hayton, James, David Allen, and Vida Scarpello, 2004. Factor Retention Decisions in Exploratory Factor Analysis: A Tutorial on Parallel Analysis. *Organizational Research Methods* 7(2): 191-205.

IBM SPSS Modeler Text Analytics 15 User's Guide. Accessed from ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/en/Users_Guide_For_Text_Analytics.pdf on February 27, 2014.

Lang, Tom, 2003. Common Statistical Errors Even You Can Find. *American Medical Writers Association* 18(2): 67-71.

Manning, C. and H. Schutze. Foundations of Statistical Natural Language Processing, 5th Ed. (Cambridge, MA: MIT Press, 2002).

Majeed, Azeem, 2012. How Should Medical Journals Deal With Errors? *Journal of the Royal Society of Medicine* 105(2): 51-52.

Miner, Gary, John Elder, Thomas Hill, Robert Nisbet, Dursun Delen, and Andrew Fast. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. (Waltham, MA: Academic Press of Elsevier, 2012).

Moher, D., A. Liberati, J. Tezlaff, and D.G. Altman, 2009. The PRISMA Group. *Open Medicine* 3(3): 123-130.

Ott, R. L. and M. T. Longnecker. An Introduction to Statistical Methods and Data Analysis, 5th Ed. (Pacific Grove, California: Brooks/Cole, 2001).

Patil, Vivek H., Surendra N. Singh, Sanjay Mishra, and Todd D. Donavan, 2008. Efficient Theory Development and Factor Retention Criteria: Abandon the 'Eigenvalue Greater than One' Criterion. *Journal of Business Research* 61(2): 162-170.

Porter, T.M., The Rise of Statistical Thinking. (Princeton, NJ: Princeton University Press, 1986).

Prescott, Robin and Ian Civil, 2013. Lies, Damn Lies and Statistics: Errors and Omissions in Papers Submitted to Injury 2010-2012. *Injury* 44(1): 6-11.

Ryker, Randy and Ravinder Nath, 1997. Factor Analysis and the Over-Extraction of Factors: An Empirical Examination. *Proceedings of the Southwest Decision Sciences Institute*, 130-131.

Schor, S. and I. Karten, 1966. Statistical Evaluation of Medical Journal Manuscripts. *Journal of the American Medical Association* 195(13): 1123-1128.

Schulz, K.F., D.G. Altman, and D. Moher, 2010. CONSORT 2010 Statement: Updated Guidelines for Reporting Parallel Group Random Trials. *BMJ*, 340: c332.

Shott, Susan, 2011. Detecting Statistical Errors in Veterinary Research. *Journal of the American Veterinary Association* 238(3): 305-308.

Smit, Eefke and Maurits van der Graaf, 2011. Journal Article Mining: A Research Study into Practices, Policies, Plans…and Promises. *Publishing Research Consortium*. Accessed from http://www.publishingresearch.org.uk/documents/PRCSmitJAMreport2.30June13.pdf on December 18, 2013.

Strasak, Alexander, Qamruz Zaman, Karl Pfeiffer, Georg Gobel, and Hanno Ulmer, 2007. Statistical Errors in Medical Research – A Review of Common Pitfalls. *Swiss Medical Weekly* 137(3/4): 44-49.

Von Elm, E., D.G. Altman, M. Egger, S. Pocock, P.C. Gotzsche, and J.P. Vandenbroucke, 2008. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement. *Journal of Clinical Epidemiology* 61(4): 344-349.