DEVELOPING STATISTICAL UNDERSTANDING AND OVERCOMING ANXIETY VIA DROP-IN CONSULTATIONS

GIZEM INTEPE
Western Sydney University
g.intepe@westernsydney.edu.au

DON SHEARMAN
Western Sydney University
d.shearman@westernsydney.edu.au

ABSTRACT

Students in many Australian universities start their studies mathematically underprepared as there are no prerequisites for mathematics, and assumed knowledge requirements are often overlooked. Many degrees include at least one statistics subject for which students require a reasonable level of mathematical ability to successfully complete. Students' efforts to grasp quantitative skills often lead to feelings of anxiety, stress, and lack of self-confidence. The Mathematics Education Support Hub (MESH) at WSU provides free support in mathematics and statistics to all students, both undergraduate and postgraduate, to increase their engagement, understanding, and abilities in statistics as well as to overcome their anxiety. In this paper we focus on the drop-in consultation service, which provides "just in time" help in campus libraries. Data is collected for every consultation, which enabling an investigation in relation to the mathematics background of students and the problematic topics in statistics. Text mining is used to examine students' queries to identify the topics in statistics subjects that students struggle most with. Outcomes of this analysis can be used by statistics instructors and mathematics support centres to improve students' experience in learning and to help to reduce statistics anxiety in future generations of statistics students.

Keywords: *Statistics education research; Statistics topics; Text mining; Drop-in consultations; Statistics anxiety*

1. INTRODUCTION

Many students currently entering university in Australia are numerically unprepared for their study in mathematics and statistics. A poor background in numeracy and statistics leads to negative effects not only on students' motivation but also on university retention and employability of graduates. Because of the greater proportion doing less mathematics in secondary school, universities needed to redesign subjects by reducing the content, or funding new support services such as mathematics learning centres (Rylands & Shearman, 2015) to account for this lack of numeracy skills. Employers look for confident and versatile graduates to utilize mathematics and statistics in numerous new job opportunities (ACME, 2011). For students, excelling in their degree is becoming more challenging since they do not feel themselves competent in mathematics and statistics subjects (Chew & Dillon, 2014). This lack of confidence creates a fear of failing subjects that require basic understanding of mathematics.

Statistics Education Research Journal, 19(1), 149–166, http://iase-web.org/Publications.php?p=SERJ

Statistics is one discipline that students are often uneasy about and is fundamental not only to STEM disciplines but also in other fields including psychology, health, and business. It is often a subject that students are expected to take in the first-year of their degree. According to a report (McNeilage, 2013), a first-year psychology student at an Australian university said her statistics class was "a slap in the face" because she did not study mathematics or science in the last two years of secondary school.

Many Australian universities have moved from having pre-requisites for a course of study (subjects, which students are required to take at school for admission to that course) to one of assumed knowledge (students are assumed to be familiar with concepts but there is not the compulsion to have formally studied them). Much of this change has been driven by market forces (Shearman, Rylands, & Coady, 2012). All but one of the first-year statistics subjects at WSU has an assumed knowledge of introductory-level Calculus. As a result of this, students are arriving at university less mathematically prepared than they have been in the past (Rylands & Shearman, 2015). Previous research (Zeidner, 1991; Wilson, 1997), which is explained in the literature-review section, has demonstrated a strong connection between students' mathematics background and performance in statistics subjects. Some universities provide free drop-in consultations to support mathematics and statistics content in all subjects and to help students to overcome difficulties they experience in these subjects.

In this study, we examine data from drop-in consultations at one university to investigate:

- statistics topics that students struggle the most with,
- the relationship between students' mathematics background and the drop-in consultation use for those who study statistics,
- how text mining can be used to identify problematic topics in statistics, and
- how tutor staff can be backed up for their service to the students.

Data mining in education is an emerging research area that utilises data generated by educational sources. Little research has been done to date studying the use of text mining data provided by educational environments. This paper is the first example of a study, which applies text mining to data on drop-in support centres.

The structure of the paper is a follows: Section 2 investigates the literature regarding the relationship between statistics anxiety and mathematics backgrounds as well as the educational data-mining literature; Section 3 explains the current state of the students who study statistics at WSU. Techniques to analyse data are explained in Section 4, and results of our analysis are presented in Section 5. Finally, research findings are summarized in Section 6 along with a discussion of our findings.

2. LITERATURE REVIEW

This paper investigates students' difficulties in statistics in regards to their mathematics background and statistics anxiety. We review the relationship between mathematics background and statistics anxiety in Section 2.1. This is followed, in Section 2.2, by an examination of statistical techniques used in text mining to extract information from free text data.

2.1 MATHEMATICS BACKGROUND AND STATISTICS ANXIETY

Despite the necessity of basic numeracy skills and published 'assumed knowledge' requirements for statistics subjects, these requirements are largely ignored or misunderstood by students and create a negative effect on students learning (King & Cattlin, 2015). Students start feeling the pressure of necessary mathematics and statistics skills in their first year of university study as many subjects require basic understanding of mathematics to complete assessment tasks. Statistics is one such subject that is taught in many disciplines such as accounting, economics, psychology, health, business, bioscience, engineering, information technology,

criminology, sociology; the courses in statistics expect an ability to understand basic mathematics to implement a statistical task (Advisory Council on Mathematics Education (ACME, 2011)). For instance, in an introductory statistics course, a very basic calculation of a standard deviation includes understanding and utilizing sigma notation and order of operations, which require basic mathematics skills. Yet, students coming into first-year university programs struggle with these mathematical concepts. Previous research has shown a significant association between mathematics ability and statistics-course performance; furthermore, numeracy skills create a roadblock in undergraduate-statistics subjects for some students (Zeidner, 1991; Rabin, et al., 2018). Taking a statistics course is not a welcome experience for students in non-mathematical disciplines as they often do not expect to take any statistics subject. Most of them do not realize the relevance of statistics to their major (Chew & Dillon, 2014).

Students who study statistics without sufficent mathematics backgrounds can experience serious problems at university. Students' efforts to comprehend quantitative skills and apply them to a quantitative task in their discipline often lead to feelings of anxiety, stress, and lack of self-confidence (Loughlin, Watters, Brown, & Johnston, 2015; Matthews, Belward, Coady, Rylands, & Simbag, 2016; Chamberlain, Hillier, & Signoretta, 2015). In the literature, statistics anxiety has been found to be associated with factors such as mathematical skills, number of mathematics courses completed, major, academic status, perception of past achievement in mathematics courses, time gap between the previous mathematics course, calculator attitude, student learning, ethnicity and the expected grade (Roberts & Saxe, 1982; Zeidner, 1991; Wilson, 1997; Onwuegbuzie, 2004). Students become more anxious about learning statistics, precisely because they feel they have poor mathematical background (Chamberlain, Hillier, & Signoretta, 2015). Students with high levels of anxiety are usually hesitant to ask help from their instructors and friends in understanding the material covered in class (Onwuegbuzie, 2000) since the fear of showing a lack of knowledge or ability impacts negatively on students' willingness to ask questions in the classroom environment (Ryan, Pintrich, & Midgley, 2001; Grehan, Mac an Bhaird, & O'Shea, 2011). These students tend to avoid exposing their own inadequacies to themselves, to lecturers or tutors, and to their peers (Grehan et al., 2011). They prefer a safer place where they feel secure in taking risks and where student thinking is respected in order to prevent or reduce the anxiety (Whyte & Anthony, 2012).

There are only a few studies that investigate how to reduce this statistical anxiety; most studies on attitudes and beliefs focus on the prevalence of anxiety toward statistics. In these studies, immediate feedback, implementing application-oriented teaching methodologies, increasing computer use and group work, using real-world data, instructor immediacy, and being sensitive to students' concerns have been found to be effective interventions in reducing statistical anxiety (Marson, 2007; Pan & Tang, 2004; Forte, 1995; Williams, 2010). Wilson (1997) suggested that institutions should provide additional mathematics support for the students who feel that their mathematics skills are worse than those of their peers in the statistics classroom, to reduce their anxiety. Lalayants (2012)'s study emphasized the importance of teaching strategies that implement practical applications and using connections to students' professional work. Lastly, Chew & Dillon (2014) advocates reducing the emphasis of mathematics, using software rather than manual calculations, introducing weekly quizzes to avoid procrastination, setting up a framework to allow anonymous questions to deal with a fear of asking for help and fear of statistics teachers, and implementing humor in teaching materials.

2.2 TEXT MINING IN EDUCATION RESEARCH

Data mining and text mining are becoming more popular in many fields; their potential as a research tool seems unlimited and useful for many domains. Recently we have seen applications of data mining in education research mostly because of the increasing amount of data generated by the teaching tools and learning environments that support student learning. This field, called educational data mining (EDM), is concerned with the use of data obtained from educational

environments such as learning management systems (LMS) to address essential questions (Reategui, Klemann, Epstein, & Lorenzatti, 2011; Romero & Ventura, 2010). Specifically, if text data is available, text-mining techniques provide a promising option to examine these data. Text mining is considered as an extension of data mining to text data and is an emerging trend in EDM (Feldman & Dagan, 1995; Liu, Cao, & He, 2011). Data-mining techniques are designed to work with structured data from databases, yet text mining can work with unstructured or semi-structured data sets (Gupta & Lehal, 2009).

Although there are many different applications of text mining in the literature, we have found only a few examples in the education domain. Ueno (2004) discussed the need for data mining and text mining for improving the research on collaborative learning in discussion boards in an LMS. Dringus and Ellis (2004) conducted the first study on embedding data-mining and text-mining techniques in a discussion forum in order to provide the instructors with a more efficient tool to evaluate students' participation. Hung (2012) used cluster analysis to identify the trends of e-learning research. He (2013) applied both data mining and text mining to a video-streaming system to identify patterns in students' learning behavior using online questions and chat messages posted by students. Similar research has been conducted by Abdous, Wu, and Yen (2012) to analyse students' chat messages for relationships between students success and their messaging activity. Tobarra et al. (2014) applied the method of topic modeling to establish a network of topics used in the forums of an LMS.

Various sources and learning environments can produce data suitable for EDM. Romero and Ventura (2010) grouped the most important studies in EDM according to the type of data involved, such as traditional environments, web-based education, learning management systems, intelligent tutoring systems, adaptive educational systems, tests and questionnaires, and texts and contents.

If text data is available in the sources mentioned above, there are many techniques that may be applied to analyze it depending on the aim of the research. Clustering models can group the data where data might be students, questions, comments, etc.; these models are helpful to identify the similarities, main themes, patterns, or trends (He, 2013; Dringus & Ellis, 2004; Dhillon & Modha, 2001; Garg & Gupta, 2018). Visualization techniques such as word clouds use word frequencies in the document to discover themes in the data (Brooks, Gilbuena, Krause, & Koretsky, 2014). The goal of topic modelling is to identify the dominant topics in different pieces of data (Blei, Ng, & Jordan, 2003; Blei, 2012). Graphs can be employed to understand the network of the data and the relationships between the variables (Reategui, Klemann, Epstein, & Lorenzatti, 2011).

3. THE PROBLEM

Although mathematics is still an essential subject in school up to year 10 in Australia, many students arrive at Western Sydney University (WSU) without mathematics in the final two years of secondary school. An analysis of 2017 WSU first-year statistics subjects (2514 domestic students) found that 65% of these students had an inadequate mathematics background for their study, 32% of whom having no mathematics in the final two years of secondary school. To help students to close this gap, the university provides free mathematics and statistics support via a centrally organised unit, the Mathematics Education Support Hub (MESH); the support center has seven tutors based on three of the university's seven campuses. This team provides support in a variety of ways, including drop-in consultations held in the university's campus libraries. At these consultations, all students may freely drop in (there are some limitations in schedule when the staff is available) and ask for support in any statistics topic. Students are informed about this support service when they enrol at university.

After each consultation, the tutor records student number, campus, subject name, duration of the consultation, date and the student's query. Most importantly, they record the content of the question in detail. Examples from two enquiries are given below.

> "What is the difference between the chi-squared test statistic and the chi-square critical value? Student had made an error calculating chi-squared test statistic but did understand the process. Student had little understanding of the use of the linear regression model – he preferred to estimate values from the scatterplot rather than substitute them into the equation" (Quantitative Thinking, 2018).

> "Student needed help on the difference between the use of *p* values and test statistics to judge the outcome of a hypothesis test" (Biometry, 2018).

Data is also available on the mathematics background of the students who completed secondary education in the last decade in NSW and undertook first-year mathematics or statistics subjects in the years 2015–18. This data is classified according to Barrington and Brown's (2014) classification system, which categorizes the mathematics subjects of the senior secondary school in the following way: *Elementary subjects* include algebra, some basic statistics, and consumer mathematics; all covered with minimal rigor. *Intermediate subjects* include introductory calculus (but often do not cover any statistics). *Advanced subjects* include more advanced calculus and other topics. As noted earlier, most first-year statistics units specify assumed knowledge to be at the Intermediate level.

In their first-year at the university, many students face statistics subjects. If the student's numeracy skills are insufficient, this creates an additional stress as they fear failure. Some students start feeling stress as soon as they enrol, even before they start studying. A recent email to the support center from a new student illustrates this:

> "I enrolled in Bachelor of Information and Communication Technology first year …. As the course is completely new for me and I am quite nervous about my long forgotten math I did when I was in High School … I would like to deep my head into math and almost starting fresh again. Now, the statistical decision making is in semester one and I am nervous about it already. So, please advise how I can start my prep as soon as possible." (BICT, 2019, Autumn).

MESH also distributes surveys at the end of each semester to students who used drop-in consultations in order to get feedback on the satisfaction about the consulting and to find out if any other means of support have been used. Replies suggest that the students who are embarrassed to ask questions in the classroom appreciate the help provided by MESH more than the help from other students. This could be partly the fear of revealing their statistical incompetence in the classroom environment. Students with poor numeracy skills are usually shy, less confident in the classroom and feel uncomfortable when seeking help (Bledsoe & Baskin, 2014). It is advised to provide a safer environment where their anonymity is secured to ask their questions (Wilson, 1997; Whyte & Anthony, 2012). A reply from the survey supporting this idea is the following one.

> "Excellent service, when in one [o]n one can ask question that may make you feel stupid in a lecture or tutorial situation" (Bachelor of Medical Science; 2016, Autumn).

Drop-in consultations give an opportunity for students to ask their questions in a different environment other than the classroom and help to improve their learning by keeping their identity from their teaching staff and peers. However, some students find any personal interaction daunting as shown by the following:

> "Only problem is, I am usually too embarrassed to ask for help, maybe a form of online help that isn't through a discussion board" (Quantitative Thinking, 2013, Spring).

This feedback shows that some students feel too embarrased if they expose their lack of knowledge to anyone face-to-face. These students prefer online statistics support to a face-to-

face support. MESH already provides online support via a discussion forum. Yet, some students prefer to stay invisible so that other ways to improve the support still have to be found.

One of the goals of this study of the drop-in consultation data is to identify the problematic statistics topics that students struggle the most with. Over the last five years, 5848 consultations have been recorded. This data may provide information about the problems students are experiencing in their subjects and help the team to improve their consultation work. Since it is impossible to read, retain, and understand all queries, algorithms are required to extract suitable information from the data base. Text-mining techniques can analyse textual data and identify most common problem areas by looking through the student queries. Differences and similarities between years of study and different statistics subjects are also examined. This will allow MESH to understand students' problems and build new resources to assist students with the most common problem areas.

4. RESEARCH METHODOLOGY

Data obtained from drop-in consultations are used in this study. Most of the analysis was conducted by text-mining techniques, since the data contains free text in the query section.

4.1 DESCRIPTION OF THE DATA

Of the nearly 6000 mathematics and statistics consultations in the years 2013 to 2018, 1580 involve statistics queries from a wide range of subjects. Of these consultations only those including a statistics related problem were used, leaving 1433 queries. These are analyzed in the following. Figure 1 shows the distribution of the consultations related to a subject. "Other" includes a variety of statistics enquiries from studies as diverse as education, health sciences, environmental science, biology, medical science, and the humanities.

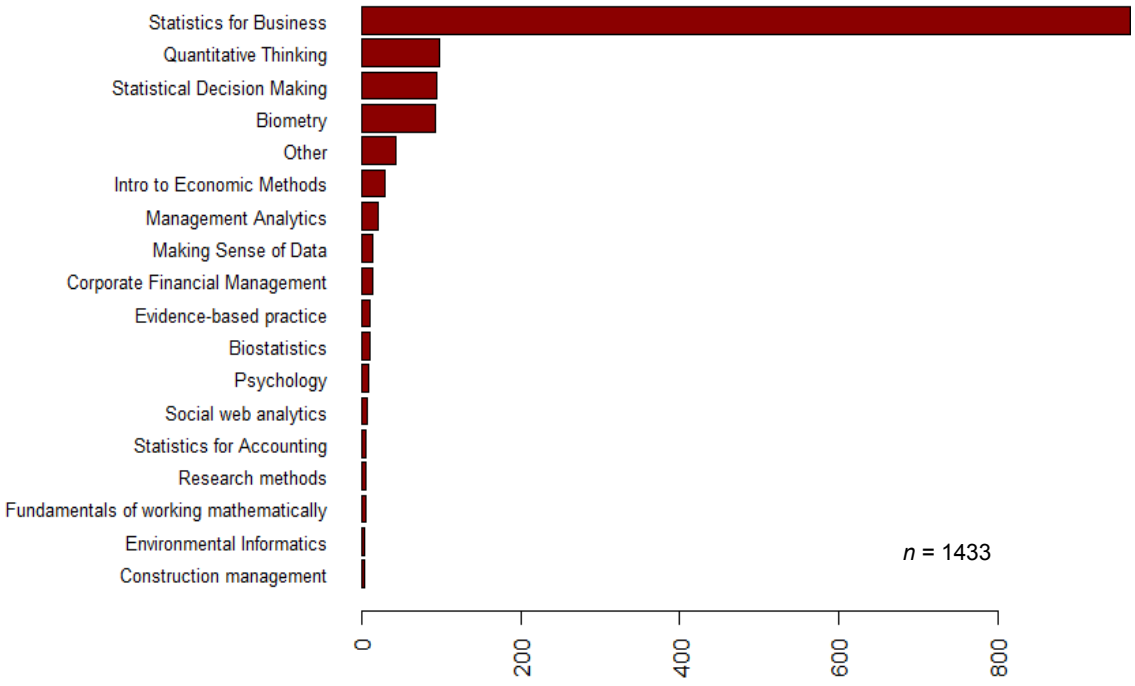

Figure 1. Distribution of statistics queries by the related subject

The distribution over the subjects looks like a Pareto distribution with a steep decrease in frequencies so that the subjects of main influence on the demand for consulting: Statistics for Business, Quantitative Thinking, Statistical Decision Making and Biometry make up nearly 90% of all consultations. The other queries are scattered over more than 13 subjects. The top four subjects are first-year subjects in statistics with the exception of Quantitative Thinking, which is a first-year mathematics subject including basic statistics.

Figure 2 shows the distribution of the year of study of the students who visited drop-in consultations. It clearly shows that this service is mainly used by first-year students.

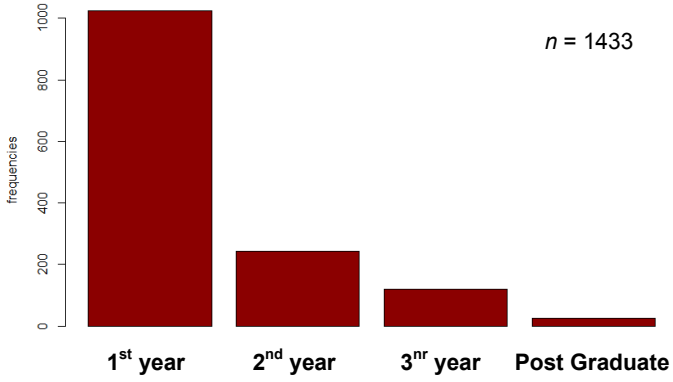

Figure 2. Students' year of study

We reduce the focus of the further investigations to the three introductory statistics subjects (80%): Statistics for Business, Statistical Decision Making, and Biometry. We extracted queries from these subjects for both first-year and non-first-year students. Students who delay taking these subjects and students who repeat the subject in later years are attributed to the Non-first-year statistics group. All other statistics queries from any other subject and from any level are included in the Other statistics group. The description of this partition is as follows:

- *First-year statistics*: queries from first-year students enrolled in a first-year statistics subject,
- *Non-first-year statistics*: queries from non-first-year students enrolled in a first-year statistics subject,
- *Other statistics*: queries from any other statistics subject regardless of the year of study.

The number of consultations in each group is shown at Figure 3. First-year statistics, Non-first-year statistics and Other statistics group form 62%, 20% and 18% of all statistics consultations respectively.

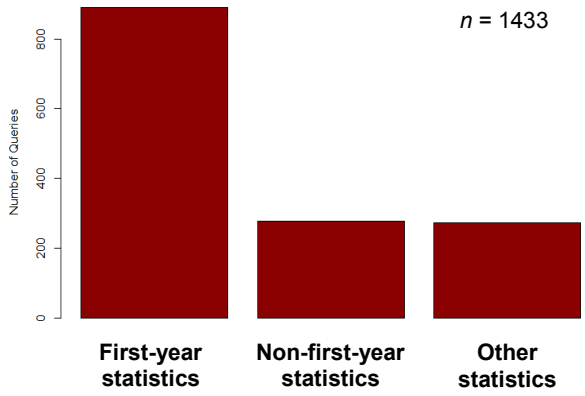

Figure 3. Number of queries in each group

4.2 ANALYSIS OF THE STUDENT BACKGROUND

One of the reasons for difficulties in statistics is students' weak mathematics background. We focus on students' mathematics backgrounds in first-year statistics subjects. Student identifiers were used to link drop-in consultation data with university data, which has a record of secondary school mathematics background, enabling us to determine the proportions for students who took advantage of drop-in consultations. Figure 4 shows the comparison between the mathematics background of the students in statistics drop-in consultations and the 8017 WSU students who study a first-year statistics subject. Students were counted once for each subject they completed. There were 1167 consultations for these subjects which accounted for 464 unique students. Background is categorized by Barrington and Brown (2014)'s classification as mentioned earlier; 'none' corresponds to students who have not included mathematics in their final exam; that is, they have not completed any mathematics course in their final two years at secondary school.

Figure 4 shows that most of the students who study statistics (at WSU) have no senior secondary-school mathematics background or have only an elementary mathematics background. Among these, students with no mathematics background tend to use the drop-in consultation service more than other students. A chi-squared test for independence was performed to examine the relationship between drop-in consultation use and the mathematics background of the students who study statistics: the relation between these variables is statistically significant ($\chi^2 = 22.18$, $p < 0.0001$).

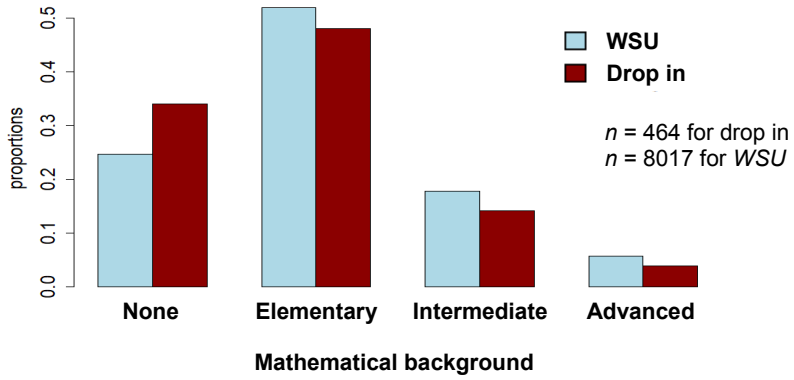

Figure 4. Mathematics background comparison between WSU and consultations

The mathematics background in the studies related to the main three first-year subjects at WSU is not promising but it varies with the subject (see Figure 5).

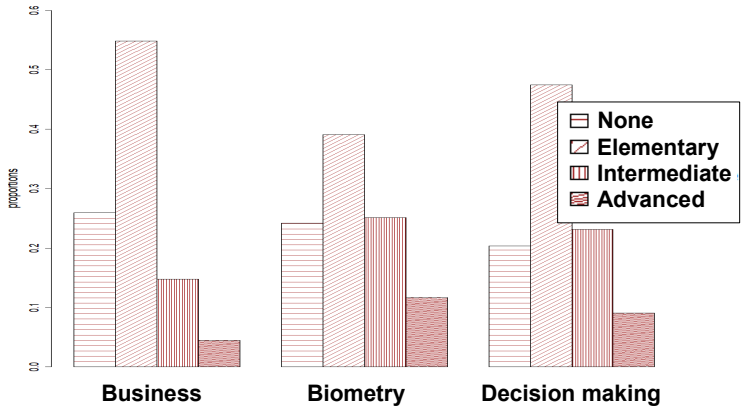

Figure 5. Mathematics background in three first-year statistics subjects at WSU

If the categories of None and Elementary are combined, then one may read from Figure 5 that approximately 70% of the students have at most elementary mathematics background. This raises a problem especially for Statistical Decision Making and Biometry, which demand more mathematics and are offered to computing and science students.

4.3 DATA PRE-PROCESSING

We utilize the language of text mining such as *terms*, *documents*, and *corpus*. *Terms* refer to the words used in queries, *document* refers to each query that includes the details of a consultation, *corpus* refers to the set of text in the whole document collection. The two phases of work are data cleaning and establishing a weighting matrix; the phases are described below.

Phase 1 This is the pre-processing part of the text mining. The purpose of pre-processing is to clean and prepare the text for analysis. This is done through a series of steps: Firstly each *document* (in our case each query) breaks up to a series of *terms* by removing punctuation, white space, and special characters, and converting text to lower case letters. Then, stop words, which are the common words that have limited content information and mostly include prepositions and conjunctions, are removed from the data. Lastly each word is replaced by its stem. The aim of stemming is to reduce different grammatical forms of a word such as its noun, adjective, verb, adverb, etc. to its root form to capture related words together (Jivani, 2011). Although the stemming process can be done automatically in R, we have developed our algorithm manually since the *terms* in the *corpus* consist of mathematical and statistical terms where the stemming function of R does not work sufficiently well.

Phase 2 After cleaning the data, queries from each of the three subjects are extracted and converted to a same data structure, which allows later analysis. The most well-known structure to represent *corpus* is the Bag-of-Words approach, which analyzes the frequency of terms in each *document* but ignores the order of *terms* (Allahyari, et al., 2017). Then each *document* is represented as a vector of *terms* and stored in a *document*-by-*term* matrix. In this matrix, rows correspond to the *documents* and the columns correspond to *terms*. Matrix entries are weightings of each *term* in the corresponding *document*. These weightings may simply be the frequency of each *term* in the relevant *document* or weights determined-by-term frequency and inverse document frequency (tf–df weights), a method, which decreases the weight of *terms* when they occur in (nearly) every *document* and attributes more weight if a *term* occurs in only a smaller fraction of the *documents*. In tf-idf weights, the *term* importance is proportional to the standard occurrence frequency and inversely proportional to the number of *documents* in which the *term* appeared (Salton & McGill, 1986). We use tf-idf weights.

4.4 ANALYSIS OF THE TEXT

The analysis is conducted with R using the packages tm (Feinerer & Hornik, 2019) for text cleaning and processing, tau (Buchta, Hornik, Feinerer, & Meyer, 2019) for bigrams, wordcloud (Fellows, 2019) for generating wordclouds and topicmodels (Grün & Hornik, 2019) for topic modelling. Four different models are applied to interpret the themes in our corpus: word clouds, bigrams, clustering, and topic modelling. These models are briefly explained below.

Word clouds are simple a visual representation of word frequencies (or weights) in a body of text. The words' relative sizes are proportional to their weights in the *document*-by-*term* matrix.

Bigrams are pairs of words that tend to co-occur within the same *document*. These may better represent concepts in the *corpus* as single words often occur in several distinct bigrams.

Topic Modeling automatically discovers topics from a collection of *documents* (Blei, 2012). It interprets patterns in the documents through a small number of *terms* – the topics – while preserving the essential statistical relationships between *terms* and *documents* (Blei, Ng, & Jordan, 2003). Topic modeling estimates the probabilities of each *term* in each topic and the result is the set of *terms* with the highest probability for each topic. Note that a *term* can be included in more than one topic since topic modelling determines set of *terms* within each topic.

Hierarchical Clustering establishes a hierarchy of clusters by grouping similar objects into clusters (Murtagh & Contreras, 2017). In our case we are clustering *terms*, which are the columns of the *document*-by-*term* matrix, regarding their appearance in the queries. The similarity between these column vectors is determined by a distance metric; we used the cosine metric (Garg & Gupta, 2018; Gupta & Lehal, 2009).

5. RESULTS

5.1 WORD CLOUDS

Word clouds are useful to summarize data but they do not provide detailed information about the topic and the relationship between words regarding their joint occurence. Figure 6 shows the word clouds of the queries of the different groups. The three word clouds share common topics such as normal distribution, probability, and hypothesis testing. First-year queries are mainly focused on calculating the probability of an event, confidence intervals, interpreting, hypothesis testing, binomial distribution and the use of Excel. Non-first-year queries are very similar to first-year queries. In the Other statistics group, R-commander, empirical rule, and standard deviation are also highly asked for in addition to first-year topics.

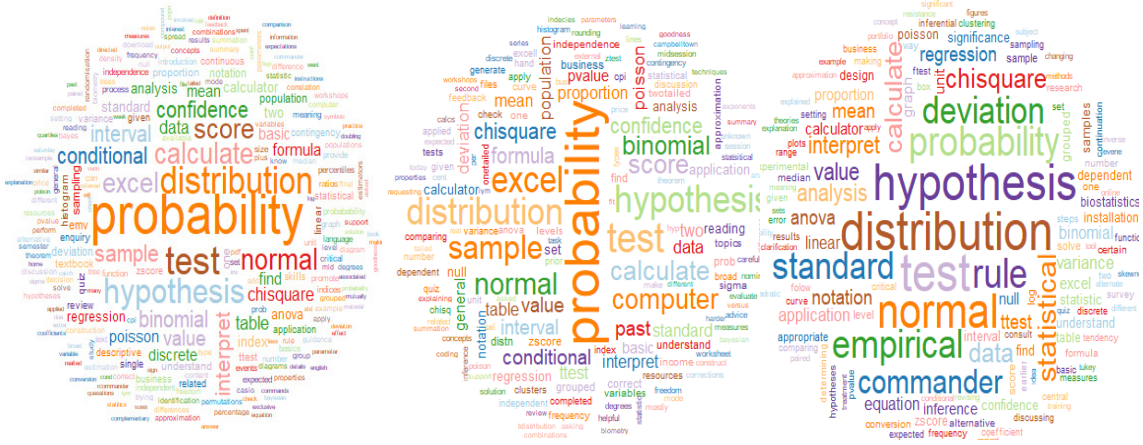

Figure 6. Word clouds of the statistics queries

5.2 BIGRAMS

Phrases rather than words give more information about the content of the query. We counted the word sequences of length 2 (bigrams) regarding their co-occurrence in a significant number of documents. Table 1 shows the most frequent bigrams in the three groups of statistics queries.

Hypothesis test and normal distribution are most asked topics among all statistics queries and chi-squared test is a common problematic topic in all statistics groups. Similar to word-cloud results, first-year and non-first-year students' queries are very similar and dominated by

the main first-year statistics topics such as confidence intervals, *z* score, and conditional probability. Data-analysis questions and Excel use are among top 10 bigrams at first-year statistics queries. Poisson distribution questions only appeared in non-first-year statistics queries. On the other hand, queries related to *p* value and empirical rule are mostly asked by the Other statistics group. Surprisingly, standard deviation is identified as a problematic topic for both Non-first-year statistics and Other statistics group.

Table 1. Most frequent bigrams in the three investigated groups in decreasing order

First-year statistics	*Non-first-year Statistics*	*Other Statistics*
hypothesis test	hypothesis test	hypothesis test
confidence interval	normal distribution	normal distribution
normal distribution	confidence interval	standard deviation
z score	z score	empirical rule
conditional probability	standard deviation	chi-square test
t test	Poisson distribution	*p* value
chi-square test	chi-square test	probability distribution
test excel	conditional probability	linear regression
calculate probability	distribution calculate	test normal
data analysis	calculate probability	test statistic

5.3 TOPIC MODELLING

Topic modeling is utilized to find a small number of words that characterize the topics, which interpret the pattern in the queries. Dominant terms, that means the highest-ranked words associated with each topic, are found for each group. This process does not require that words belong to only one topic as each word has a probability of occurring within each topic, the most probable words are used to define the topic. These terms help us to understand the topic context. We reviewed the terms for all topics to build the meaning of the topics. For all investigated groups, the top terms of each topic and a brief explanation are given in Table 2.

5.4 HIERARCHICAL CLUSTERING

Hierarchical Clustering creates a hierarchical decomposition of the documents or words in a dataset. Unlike topic modeling discussed above, hierarchical clustering places each word in a unique cluster based on its similarity to other words. In this study, we are interested in word hierarchies to understand the topics from the top to the lower levels of the hierarchy. Figures 7 to 9 show term hierarchies obtained from the *document*-by-*term* matrix of each investigtated group.

Hierarcical clustering helps to identify topics in detail since it provides a disaggregation of the topics. The problematic topics among first-year statistics queries are hypothesis testing (especially one or two sample t test) and confidence intervals that include two samples. Students also have difficulty in calculating standard deviation and variance. Calculating conditional probability from a contingency table and understanding proportions, statistical notation and the formula are in same cluster with probability questions. Normal and discrete distributions, chi-squared test, regression analysis, finding test statistics, and computer use to draw a histogram or running an analysis with Excel data analysis add-ins as well as calculator use are the main

problem areas among first-year students. Note that *data analysis* is in same cluster as *Excel*, which helps us to understand that *data analysis* is related to the use of the corresponding add-in of Excel. This was not clear from the previous results. The presence of words such as *calculate*, *find* and *apply* reflect that students at this level may be struggling with the basic calculations associated with each of the topics suggested by the groupings, the presence of a topic comprising *calculator* and *skills* only is also indicative of this.

Table 2. Description of the topic keywords of the queries for each group

Group	*Topic*	*Terms*	*Explanation of the topic*
First-year Statistics	1	*Sample, Population, Confidence, Interval, Basic*	Calculating confidence intervals for population parameters from sample statistics.
	2	*Hypothesis, Test, Two, Score, Interpret, Chi square*	Hypothesis testing; mostly two sample tests. Finding *z* score and interpreting results, chi-squared test.
	3	*Normal, Distribution, Binomial, Poisson, Find, Understand, Value*	Understanding the normal, binomial or Poisson distribution related questions. Finding the value of the test statistic.
	4	*Table, Value, Discrete, Calculate, Conditional, Probability*	Use of discrete distribution tables finding the value from the table. Calculating conditional probabilities from the given (contingency) table.
	5	*Excel, Data, Analysis, Textbook, Formula, Notation*	Using Excel data analysis add-in, understanding the notation of the formulas given in the textbook.
Non First-year Statistics	1	*Hypothesis, Test, Value, Population, Mean, Chi square, Poisson*	Hypothesis testing for population mean, chi-squared test and Poisson distribution.
	2	*Probability, Normal, Distribution, Table, Score, Standard*	Finding probabilities and *z* score from the normal distribution table.
	3	*Excel, Data, Sample, Calculate, Interpret, Confidence, Interval*	Using Excel to calculate confidence interval from a sample dataset, and interpreting results.
Other Statistics	1	*Hypothesis, Test, t Test, Chi square, Sample, Linear, Equation*	Hypothesis testing, t test and chi-squared test, linear regression methods; understanding and using the formulas of these methods.
	2	*Standard, Deviation, Calculate, Calculator, Empirical, Rule, Interpret, Value, Statistic*	Calculating standard deviation with and without a calculator, interpreting test statistics, *p* value and empirical rule.
	3	*Statistical, Significance, Mean, Understand, Data, Research, Regression, Analysis*	Understanding research results, interpreting significance of a statistical test, regression analysis.
	4	*Normal, Distribution, Null, Hypothesis, Find, Probability, Application*	Normal distribution and its applications, finding probabilities from a normal distribution, constructing null and alternative hypothesis.

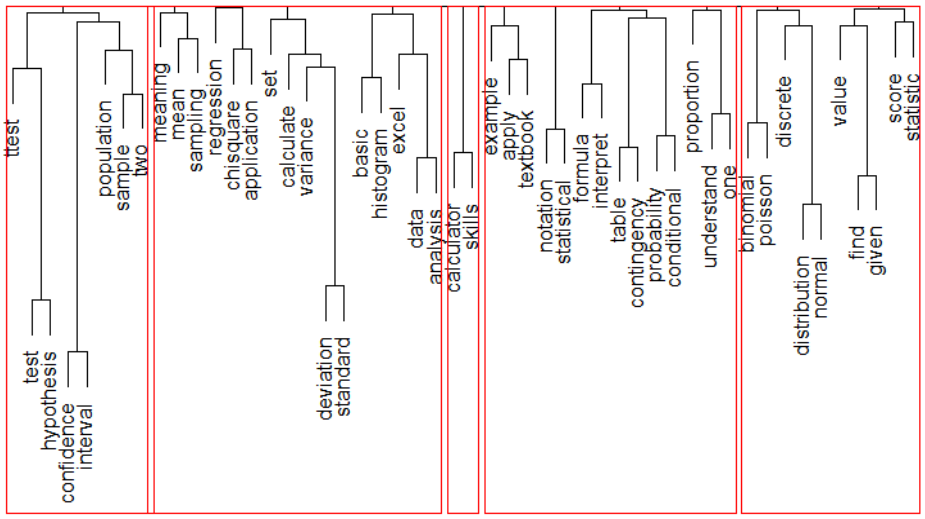

Figure 7. Hierarchy of the most frequent words in first-year queries

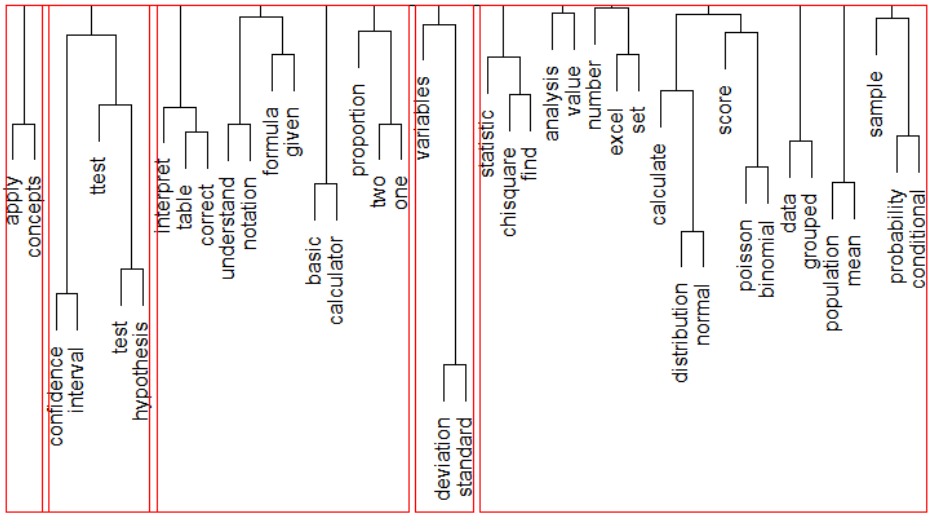

Figure 8. Hierarchy of the most frequent words in non-first-year queries

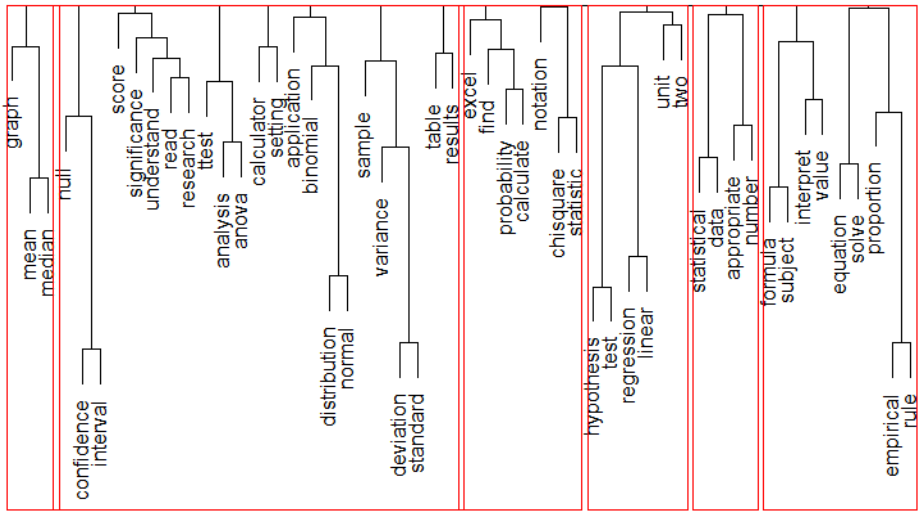

Figure 9. Hierarchy of the most frequent words in other queries

Non-first-year statistics topics are very similar to first-year statistics topics. However one topic comprising *apply* and *concepts* suggests that these students may be moving on from the more fundamental questions asked by first-year students (Figures 7 and 8).

Queries from the Other statistics group show the greatest variability of the three groups (Figure 9). This is not unexpected as questions in this group come from a much more diverse group of students enrolled in subjects ranging from introductory first-year statistics through to students who may be involved in undergraduate research projects. This leads to a range of topics covering everything from basic calculation and interpretation of descriptive statistics (*mean*, *median*, *graph*) to those involving understanding of statistics in research papers.

6. CONCLUSION AND DISCUSSION

We have used a range of techniques in an attempt to identify the key topics that are related to the queries of students in drop-in consultations. Whilst there is broad agreement in the results across these methods about the general topics, it appears that there may be more subtle differences in the specifics of questions depending on the student's position in terms of their course of study. While first-year students are mostly concerned with applying a technique and using the formulation of the method, students who study statistics in the later years are more focused on understanding the method and interpreting the results. By contrast, students who do not study a traditional first-year statistics subject, try to create a basic understanding of the key concepts in statistics.

Topics causing concern across all statistics subjects include standard deviation calculation, use of statistical tables, hypothesis testing, calculator and computer use. The investigation of the mathematics background indicates that the difficulty of understanding such simple topics in statistics may simply be the lack of mathematics background. The mathematics background is also significantly related to use of drop-in consultation: Figure 1 shows that in drop-in consultations, 34% of the students have no secondary school mathematics and 48% have only elementary mathematics background. Feedback, which is shown in the following, indicates that students who do not feel themselves equipped, benefit from drop-in consultations.

> "As I do not have an extensive background with statistics, learning some of the basics was a bonus" (Psychology: Behavioural Sciences, 2017, Spring).

Moreover, students find drop-in consultations more approachable and they feel themselves more comfortable to ask questions. A reply from the survey corroborateing this statement is:

> "The great aspect of the roving MESH-iacs (as I like to refer to them), is that in the library, the same staff can feel so much more approachable, particularly for students who are not confident, or generally shy in nature. It's less intimidating getting help in the library, compared with in lecturer rooms. And with different MESH-iacs on different days, it's even better as you can have a concept that you may be struggling to fully grasp, explained to you in a number of different ways each person" (Introduction to Statistics, 2013, Spring).

The literature gives a positive message about the benefits of learning support outside of the classroom environment. Availability of the support mechanisms helps students to overcome their fears; hence give a chance to improve their skills in statistics. Current design of statistics subjects requires basic understanding of mathematics. This can be seen from the study results as students main concerns are about the calculation part of statistical problems such as formula use and applying techniques. Most statistics subjects and assignments are method based rather than interpretation and understanding, so students' focus is mostly on these topics. In such cases it is very likely for a student to experience statistics anxiety if they have weak mathematics background (Roberts & Saxe, 1982; Wilson, 1997). Students who study advanced statistics

subjects are still struggle with basic concepts. This indicates that they did not develop basic statistical understanding in their first-year statistics subjects. Since these first-year subjects focus mostly theory/techniques, students find it difficult to create connections between theory and applications in later years. However achieving an appropriate balance between theory and application is a challenging task in a first-year introductory statistics subject. When this balance is not developed, statistical thinking is not encouraged effectively. Apart from applying a technique or calculating a result, students need to recognize when they should apply statistical thinking, how they accurately employ the statistical techniques, know when they require additional methods and how to obtain this additional understanding (Ramirez, Schau, & Emmioglu, 2012). Without such knowledge, students' confidence is not developed. Having someone from outside the classroom environment to discuss statistical concepts freely can help students to build understanding and how to approach a question. This is an opportunity to discuss statistical concepts deeply if students are willing. Usually drop-in consultations do not have time limits if no one else is waiting. Once a basic statistical understanding is developed it is much easier to build on it. Personal improvement becomes much easier and they will be able to understand/interpret statistical problems they may experience in later years of their study or at work. Additional learning support helps to increase engagement and interest and in the long term increases the number of individuals who are confident in working with statistics. Maybe the most important fact of the students' statistics experience at university is that they impact life-long perceptions and attitudes towards the value of statistics, and hence many future employees, employers, and citizens (Tishkovskaya & Lancaster, 2012).

Many students do not comprehend the significance of statistical literacy in their profession or working environment and battle to engage with statistics. Academic emotions such as anxiety determines the level of engagement with the subject where engaging more with the subject leads to better understanding and academic results (González, Rodríguez, Faílde, & Carrera, 2016). When anxiety is overcome, these students have chance to excel in statistics and in their profession. A major concern in teaching statistics is to enable students to understand statistical ideas and make them able to employ this knowledge to real-world scenarios (Garfield, 1995). Additional support may help students to overcome their statistical anxiety and become more confident about statistics.

The current state of drop-in consultations at WSU shows that students make use of this service and benefit from it. However service can be improved with regards to study outcomes. MESH is building extra resources to improve students' understanding on the problematic topics that emerged from this research. These resources can be provided to students via the university's learning management system. The development of such resources requires considerable time and hence expense. When common problematic topics are identified, resources can target these areas. The analysis provided here can identify areas where these resources can be most effective in reducing student anxiety about statistics and increase engagement. Students who feel too embarrassed to ask questions make use of online resources as well as other students. Finally, an understanding of the differences in focus of questions arising from first-year versus non-first year students can help support staff to be better prepared in answering student questions and allaying their anxiety.

We see the results of this study as having the potential to help alleviate students' anxiety and increase their engagement from three directions. Two involve the provision of support by MESH: development of new resources and better advising members of the team who are doing the consultations. The third one is by informing lecturers of areas of the curriculum, which create the most anxiety for students. There has been considerable discussion on curriculum review with regard to first-year statistics (Lalayants, 2012; Marson, 2007; Williams, 2010) much of which is around pacing of material and feedback to students.

A big problem with drop-in consultations is the lack of awareness of this service and the lack of personal motivation of students (Grehan, Mac an Bhaird, & O'Shea, 2011). Poor engagement with the learning support is another problem since students do not think that they

need help (Rylands & Shearman, 2015). These services should be promoted by the lecturer and by the university to encourage students to use them.

REFERENCES

Abdous, M. H., Wu, H., & Yen, C. J. (2012). Using data mining for predicting relationships between online question theme and final grade. *Journal of Educational Technology & Society*, *15*(3), 77–88.

ACME (Advisory Committee on Mathematics Education) (2011). *Mathematical needs: Mathematics in the workplace and in higher education*. London: The Royal Society. [Online: royalsociety.org/topics-policy/publications/2011/mathematical-needs-mathematics-in-the-workplace-and-in-higher-education/]

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv.org*. [Online: arxiv.org/abs/1707.02919]

Barrington, F., & Brown, P. (2014). AMSI monitoring of participation in Year 12 mathematics. *Gazette of the Australian Mathematical Society, 41*(4), 221–226.

Bledsoe, T. S., & Baskin, J. J. (2014). Recognizing student fear: The elephant in the classroom. *College Teaching, 62*(1), 32–41.
[Online: www.austms.org.au/Gazette+Volume+41+Number+4+September+2014]

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77-84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3,* 993–1022.

Brooks, B. J., Gilbuena, D. M., Krause, S. J., & Koretsky, M. D. (2014). Using word clouds for fast, formative assessment of students' short written responses. *Chemical Engineering Education, 48*(4), 190–198.

Buchta, C., Hornik, K., Feinerer, I., & Meyer, D. (2019, Mar 4). Package 'tau'. *The Comprehensive R Archive Network*. [Online: cran.r-project.org/web/packages/tau/tau.pdf]

Chamberlain, J. M., Hillier, J., & Signoretta, P. (2015). Counting Better? An examination of the impact of quantitative method teaching on statistical anxiety and confidence. *Active Learning in Higher Education, 16*(1), 51–66.

Chew, P. K., & Dillon, D. (2014). Statistics anxiety update; Refining the construct and recommendations for a new research agenda. *Perspectives on Psychological Science, 9*(2), 196–208.

Daly, N. (2017, October 22). Australian students are turning their backs on maths and science, and experts are worried. *ABC News* [Online: www.abc.net.au/news/2017-10-22/australian-students-turning-their-back-on-maths-and-science/9074114]

Dhillon, I. S., & Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. *Machine Learning, 42*(1–2), 143–175.

Dringus, I. S., & Ellis, T. (2004). Using data mining as a strategy for assessing asynchronous discussion forums. *Computers & Education, 45*(1), 141–160.

Feinerer, I., & Hornik, K. (2019, Dec 12). Package 'tm'. *The Comprehensive R Archive Network*. [Online: cran.r-project.org/web/packages/tm/tm.pdf]

Feldman, R., & Dagan, I. (1995). Knowledge discovery in textual databases (KDT). In U. Fayyad & R. Uthurusamy (Eds.), *Proceedings of the First International Conference on Knowledge Discovery and Data Mining* (pp. 112–117). Montreal: AAAI Press [Online: dl.acm.org/citation.cfm?id=3001335&picked=prox]

Fellows, I. (2019, January 22). Package 'wordcloud'. *The Comprehensive R Archive Network*. [Online: cran.r-project.org/web/packages/wordcloud/wordcloud.pdf]

Forte, J. A. (1995). Teaching statistics without sadistiics. *Journal of Social Work Education, 31*(2), 204–218.

Garg, N., & Gupta, R. K. (2018). Exploration of various clustering algorithms for text mining. *International Journal of Education and Management Engineering, 8*(4), 10–18.

Grehan, M., Mac an Bhaird, C., & O'Shea, A. (2011). Why do students not avail themselves of mathematics support? *Research in Mathematics Education, 13*(1), 79–80.

Grün, B., & Hornik, K. (2019, Dec 3). Package 'topicmodels'. *The Comprehensive R Archive Network*. [Online: cran.r-project.org/web/packages/topicmodels/topicmodels.pdf]

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence, 1*(1), 60–76.

He, W. (2013). Examining students' online interaction in a live video streaming environment using data mining and text mining. *Computers in Human Behavior, 29*(1), 90–112.

Hung, J. L. (2012). Trends of e-learning research from 2000 to 2008: Use of text mining and bibliometrics. *British Journal of Educational Technology, 43*(1), 5–16.

Jivani, A. G. (2011). A comparative study of stemming algorithms. *International Journal of Computer Technology and Applications, 2*(6), 1930–1938.

King, D., & Cattlin, J. (2015). The impact of assumed knowledge entry standards on undergraduate mathematics teaching in Australia. *International Journal of Mathematical Education in Science and Technology, 46*(7), 1032–1045.

Lalayants, M. (2012). Overcoming graduate students' negative perceptions of statistics. *Journal of Teaching in Social Work, 32*(4), 356–375.

Liu, B., Cao, S., G., & He, W. (2011). Distributed data mining for e-business. *Information Technology and Management, 12*(2), 67–79.

Loughlin, W. A., Watters, D. J., Brown, C. L., & Johnston, P. R. (2015). Snapshot of mathematical background demographics of a broad cohort of first year chemistry science students. *International Journal of Innovation in Science and Mathematics Education, 23*(1), 21–36.

Marson, S. M. (2007). Three empirical strategies for teaching statistics. *Journal of Teaching in Social Work, 27*(3–4), 199–213.

Matthews, K. E., Belward, S., Coady, C., Rylands, L., & Simbag, V. (2016). Curriculum development for quantitative skills in degree programs: a cross institutional study situated in the life sciences. *Higher Education Research and Development, 35*(3), 545–559.

McNeilage, A. (2013, December 14). Maths and science lecturers struggle with ill-prepared university students. In: *The Sydney Morning Herald.* [Online: www.smh.com.au/national/nsw/maths-and-science-lecturers-struggle-with-ill-prepared-university-students-20131213-2zcvq.html]

Murtagh, F., & Contreras, P. (2017). Algorithms for hierarchical clustering: an overview, II. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 7*(6), e1219. [Online: doi.org/10.1002/widm.1219]

Onwuegbuzie, A. J. (2000). Statistics anxiety and the role of self-perceptions. *The Journal of Educational Research, 93*(5), 323–330.

Onwuegbuzie, A. J. (2004). Academic procrastination and statistics anxiety. *Assessment & Evaluation in Higher Education, 29*(1), 3–19.

Pan, W., & Tang, M. (2004). Examining the effectiveness of innovative instructional methods on reducing statistics anxiety for graduate students in the social sciences. *Journal of Instructional Psychology, 31*(2), 149–159.

Rabin, L., Fink, L., Krishnan, A., Fogel, J., Berman, L., & Bergdoll, R. (2018). A measure of basic math skills for use with undergraduate statistics students: The MACS. *Statistics Education Research Journal, 17*(2), 179–195.

Reategui, E., Klemann, M., Epstein, D., & Lorenzatti, A. (2011). Sobek: A text mining tool for educational applications. in R. Stahlbock (Ed.), *Proceedings of the International Conference on Data Mining* (pp. 59–64). CSREA Press.

Roberts, D. M., & Saxe, J. E. (1982). Validity of a statistics attitude survey: A follow-up study. *Educational and Psychological Measurement, 42*(3), 907–912.

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40*(6), 601–618.

Ryan, A. M., Pintrich, P. R., & Midgley, C. (2001). Avoiding seeking help in the classroom: Who and why? *Educational Psychology Review, 13*(2), 93–114.

Rylands, L., & Shearman, D. (2015). Supporting engagement or engaging support? *International Journal of Innovation in Science and Mathematics Education, 23*(1), 64–73.

Salton, G., & McGill, M. J. (1986). *Introduction to modern information retrieval*. New York*:* Mc Graw Hill.

Shearman, D., Rylands, L., & Coady, C. (2012). Improving student engagement in mathematics using simple but effective methods. In *Proceedings of the Joint Australian Association for Research in Education and Asia-Pacific Educational Research Association Conference* (pp. 1–8). Sydney: Australian Association for Research in Education.
[Online: www.aare.edu.au/publications/aare-conference-papers/]

Tobarra, L., Robles-Gómez, A., Ros, S., Hernández, R., & Caminero, A. C. (2014). Analyzing the students' behavior and relevant topics in virtual learning communities. *Computers in Human Behavior, 31*, 659–669.

Ueno, M. (2004). Data mining and text mining technologies for collaborative learning in an ILMS "Ssamurai". *In Proceedings of the IEEE International Conference on Advanced Learning Technologies* (pp. 1052–1053). Joensuu, Finland: IEEE.
[Online: doi.org/10.1109/ICALT.2004.1357749]

Whyte, J., & Anthony, G. (2012). Mathematics anxiety: The Fear factor in the mathematics classroom. *New Zealand Journal of Teachers' Work, 9*(1), 6–15.

Williams, A. S. (2010). Statistics anxiety and instructor immediacy. *Journal of Statistics Education, 18*(2), 1–18.

Wilson, V. (1997, November). Factors related to anxiety in the graduate statistics classroom. *Annual Meeting of the Mid-South Educational Research Association.* Memphis: Educational Resources Information Center (ERIC).
[Online: eric.ed.gov/?id=ED415288]

Zeidner, M. (1991). Statistics and mathematics anxiety in social science students: Some interesting parallels. *British Journal of Educational Psychology, 61*(3), 319–328.

GIZEM INTEPE
Western Sydney University,
Building EQ, Parramatta South Campus,
Locked Bag 1797
Penrith, NSW, 2751, AUSTRALIA