CULTIVATING INTERDISCIPLINARY UNDERGRADUATE RESEARCH IN APPLIED STATISTICS

David I Holmes and Elizabeth D Johnson
George Mason University, Virginia, USA
dholmes4@gmu.edu

Undergraduate research is considered to be one of the most beneficial educational practices. George Mason University promotes and supports undergraduate research activities through the Office of Student Scholarship, Creative Activities and Research (OSCAR). In this paper we will describe three ways that students can receive financial support and/or academic credits for projects in addition to developing their professional and analytical skills in areas of statistics. One approach is for students to find a project and a mentor then apply in a competitive field for OSCAR funding. Another is for faculty to advertise for undergraduate research assistantships supported by OSCAR. Finally, receiving research funding for projects integrated into the curriculum such as capstone courses. An interdisciplinary case study involving Statistics and Literature will be discussed.

INTRODUCTION

The importance of engaging students in research is now well understood to be a necessary component in the educational experience of undergraduates. The recently revised Curriculum Guidelines for Undergraduate Programs in Statistical Science stress the need for students to be given opportunities to work with real data and to "apply their knowledge of theoretical foundations to the sound analysis of data." Thus, there need to be avenues to match potential faculty mentors to undergraduate students. At George Mason University, we have three such avenues: the American Statistical Association (ASA) student chapter, the Office of Student Scholarship, Creative Activities and Research (OSCAR) and the senior capstone experience for our Statistics majors. Each of these avenues exposes undergraduate students to faculty research interests. A recent article by Jo Hardin discusses the skills learned during the process of conducting research and the importance of mentor-student matching. Hardin notes that "the most important aspect of successful research is the degree to which you are excited about the project. If you love what you are doing, the student will sense that and be just as engaged. So, if there is something you want to work on, I implore you to assign an undergraduate student to the project, regardless of their background or the curriculum from which they come."

GEORGE MASON UNIVERSITY'S OSCAR PROGRAM

Mason values scholarship as a core characteristic of its graduates and OSCAR helps undergraduate students to find relevant, exciting research and creative projects and supports their work financially through grants and travel funding. Typically, a student would find a project and a faculty mentor and then go through a competitive application process for an OSCAR grant. The grant would provide funding through stipends for the student-mentor team to address a scholarly question during a semester. The student would then showcase his/her work at a professional forum such as Mason's Celebration of Student Scholarship and also be encouraged to travel and present their work at external forums. Students may also receive academic credit for their projects and are expected to attend seminars on research topics. Mason's aim is to have this style of learning permeate all its undergraduate education and the OSCAR program has become one of the jewels in its crown. In 2015 OSCAR won the national award for excellence from the Council on Undergraduate Research.

OSCAR also aids faculty in creating "Research and Scholarship Intensive" courses which are designed around an authentic research or creative project in the context of a course. Such courses provide a unique opportunity for faculty to merge their teaching and scholarship and provide further opportunities for students to articulate a scholarly question and engage in the key elements of the scholarly process.

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

OSCAR also coordinates with the Office of Student Financial Aid in order to allow students to use their federal work-study to work as undergraduate research assistants for assigned faculty members. Job descriptions for each position are posted on the Mason site.

INTERDISCIPLINARY CASE STUDY

During the Fall 2017 semester, I had my first opportunity to mentor a student for the OSCAR program in my field of research, stylometry. The origins of stylometry – the statistical analysis of literary style – date back to 1851 when the English logician Augustus de Morgan suggested in a letter to a friend that questions of authorship might be settled by determining if one text "does not deal in longer words" than another. Since then researchers have been attempting to solve problems of disputed authorship using a variety of statistical techniques.

The growth of computer power and the ready availability of machine-readable versions of literary works have led to almost every conceivable measure being studied for its usefulness in authorship attribution. In a pioneering work first published in 1964, Mosteller and Wallace employed frequencies of function words such as prepositions, conjunctions and articles as discriminators to investigate the mystery of the authorship of the *Federalist Papers*. Their scholarly analysis opened the way to the modern, computerized age of stylometry. The use of non-contextual high-frequency functions words as tools in attributional problems was continued by J. F. Burrows (1992) and since then multivariate statistical analyses involving large sets (50-100) of such words have met with astonishing success. The 'Burrows' approach essentially picks the N most common words in the corpus under investigation and computes the occurrence rate of these N words in each text or text-unit, thus converting each text into an N-dimensional array of numbers. Multivariate statistical techniques, most commonly principal components analysis and cluster analysis, are then applied to the data to look for patterns. The 'Burrows' approach has become the first port-of-call for attributional problems.

The application of statistical techniques to literature opens avenues of collaboration between statisticians, historians and literary scholars, and projects in stylometry are particularly attractive not only to statistics undergraduates but also to undergraduates in these other disciplines who have taken classes in statistics. See, for example, the investigation into the authorship of the so-called 'Pickett Letters' of the American Civil War (Holmes *et al.*, 2001) conducted by a statistician, military historian and an undergraduate student. Students initially undertake archival research to find authentic works by writers under study along with suitable control texts matched in time and genre, before putting the text corpora into machine-readable form. Next, the use of specialist software to run word-count analyses enhances students' computing skills before statistical skills come into play with the use of multivariate techniques. Not least, undergraduate students learn the art of collaboration with subject specialists in addition to having to disseminate and explain their findings to non-statisticians.

In the Spring 2017, I gave a presentation on my research to undergraduate students in our student chapter of the ASA at George Mason University. One student expressed great interest in joining forces with me on an extension of a project discussed during my talk. Together we applied for and received an OSCAR grant. Evan is an Economics major, a data analysis minor and was in my applied multivariate analysis class. He proved to be a perfect fit.

Our OSCAR project built on the following case study. The book *The Expert at the Card Table* was first published in 1902 and is considered to be one of the most important texts on sleight of hand. For years, fierce debate has raged about the true identity of its author, the mysterious 'S. W. Erdnase'. There is evidence for and against each of the many candidates. For example, early on in the hunt it was pointed out that the surname 'Andrews' is a part reversal of 'S.W. Erdnase', which led to two proposed candidates: a Milton Franklin Andrews and a James Andrews. In addition, newspaper reports from the turn of the last century suggest that Milton Franklin Andrews was clearly knowledgeable about crooked gambling. However, in 1946 the illustrator of *The Expert*, Marshall D. Smith, described Erdnase as a short man, whereas Milton Franklin Andrews was over six feet tall. Several writers have also suggested that *The Expert* may be the work of more than one person. For example, magician Edgar Lawton Pratt claimed to have known people who knew Erdnase, and in 1947 told Martin Gardner that the part of *The Expert* devoted to

cheating with cards was the work of one individual and that the material on magic was the work of another (Ortiz, 1991).

Although Erdnase hunters have examined a vast amount of historical evidence, only a few have paid attention to perhaps the most abundant source of information left by Erdnase, namely, the words used to create *The Expert*. In an initial study in 2011, Wiseman and Holmes split *The Expert* into four main sections and then subdivided each of these sections into textual samples of between 2,000 and 3,000 words. Multivariate analysis using the top 60 most frequently occurring non-contextual function words confirmed that the book may well be the work of two authors, with the section containing card tricks and the sections about sleights (both gambling and magic) clustering separately. The study also showed none of the eight candidates examined were a match to either part of the book.

In early 2017 two further candidates were put forward: (i) one Edward Gallaway, who worked in the printing company believed to have printed the first edition of *The Expert* (James McKinney & Company) and who self-published several books on print estimating in Chicago in the late 1920's, (ii) one Edward D. Benedict of Chicago, a former professional magician who became a book distributor and who, falling bankrupt in 1902, owed McKinney money. By 1903 Benedict mysteriously had his bankruptcy discharged. Evan worked with me to incorporate these additional candidates into the analysis.

Obtaining machine-readable versions of Gallaway's authentic writings, in particular, was time-consuming, but we used the opportunity for this new examination of *The Expert* to clean and re-structure the text files, not least to split the Introduction into new text samples, and revise our list of non-contextual function words. Using his skills acquired in my applied multivariate analysis class, Evan was able to create the dendrogram shown in Figure 1 on our text samples.

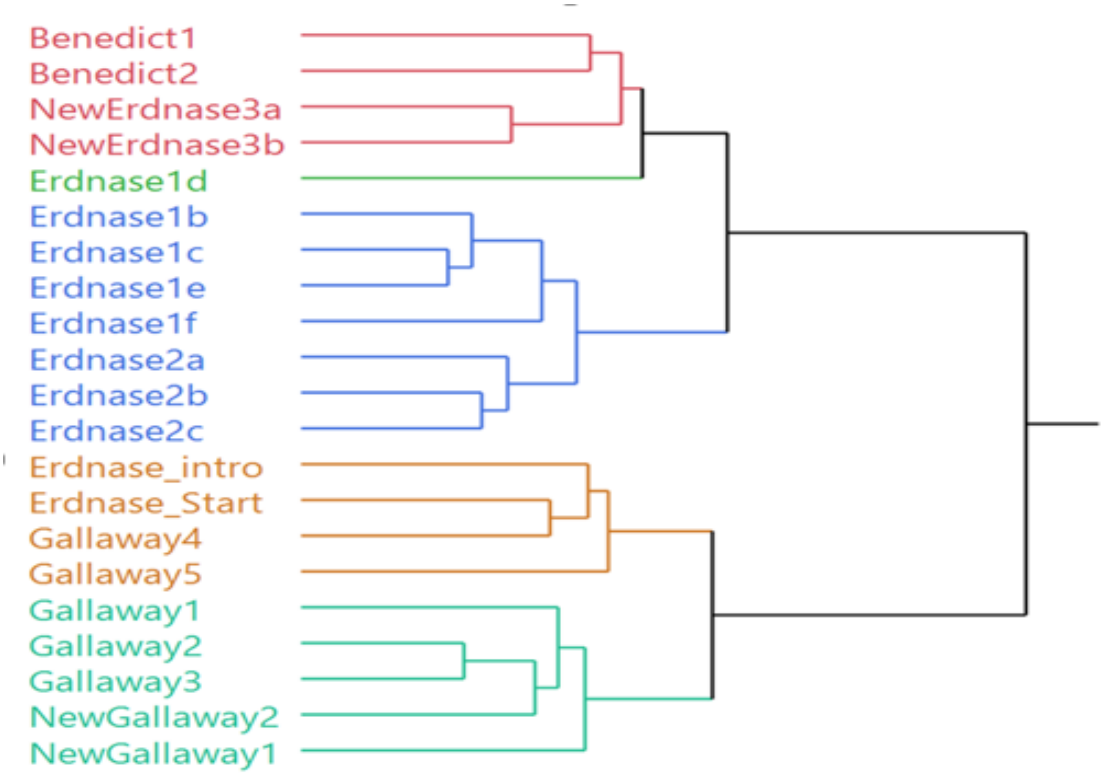

Figure 1. Hierarchical Cluster Analysis of Erdnase, Benedict and Gallaway using Ward's Method on the top 62 most frequently occurring non-contextual words.

This cluster analysis revealed that Gallaway was not a match to the main parts of *The Expert* but, intriguingly, Benedict appeared to match the card tricks section. This discovery has now been passed over to the cadre of professional magicians with whom Evan and I are collaborating. The search continues. Erdnase may have given away his entire repertoire of cheating

but he remains, however, a phantom of the card table, an anonymous figure who plays and wins and then takes his leave. And, as the door shuts behind him, like a great magician, he keeps you guessing.

CONCLUSION

Evan showcased his work on attributing *The Expert* at Mason's annual Celebration of Student Scholarship at Mason. During this event, students, who were representing every school of the university, present talks and posters summarizing their research projects. At the time of writing he has also applied to present at the Undergraduate Research Symposium and Competition Computers & Writing 2018, which invites proposals/abstracts from undergraduate scholars and researchers who work in any area that addresses questions at the intersections of communication and technology. The process of mentoring undergraduate students on research projects will be incorporated in the year-long senior capstone course for our new Bachelor of Science degree in Statistics. For the capstone course, students will work in teams to address research questions related to their area of concentration which could include the areas of mathematical statistics, applied statistics or statistical analytics.

REFERENCES

American Statistical Association Undergraduate Guidelines Workgroup. (2014). *2014 Curriculum guidelines for undergraduate programs in statistical science*. Alexandria, VA: American Statistical Association

Burrows, J.F. (1992). Not unless you ask nicely: the interpretive nexus between analysis and information. *Literary and Linguistic Computing*, *7*, 91-109.

Computers and Writing Conference, May 2018. Undergraduate Research Symposium and Competition, http://candwcon.org/2018/

Hardin, J. (2017). Expectations and Skills for Undergraduate Students doing Research in Statistics and Data Science. *AMSTAT News, September.*

Holmes, D.I., Gordon, L.J. and Wilson, C. (2001). A Widow and her Soldier: Stylometry and the American Civil War. *Literary and Linguistic Computing,16*, 403-420.

Mosteller, F. and Wallace, D.L. (1964). *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. Reading, MA: Addison-Wesley.

Ortiz, D. (1991). Letter: Pratt to Gardner, 21st June 1947. *Annotated Erdnase*. Magical Publications.

Wiseman, R. and Holmes, D.I. (2011). Stylometry and the Search for S.W.Erdnase. *Genii, 74(2),* 70-73.