OZCOTS 2021

Proceedings of the 10th Australian Conference on Teaching Statistics

Editors:
Ayse Aysin Bilgin
Macquarie University
and
Stephanie Budgett
The University of Auckland

Virtual Conference
8-9 July 2021

Held conjointly with the Australian and New Zealand Virtual Statistical Conference (ANZSC), 5-9 July 2021

Statistical Society of Australia

The New Zealand Statistical Association

ISBN: 978-0-9805950-3-1

	FRIDAY 9 JULY 2021 (continued)	
	OZCOTS Session 4 (Session chair: Yan Wang)	
13:30	Ben O'Neill – Online Q&A in Statistics – Using the StackExchange Network	65
13:45	Andrew Zammit-Mangion – Assessment Randomisation in Statistics and Related Disciplines	73
14:00	Tania Prvan, Ayse Aysin Bilgin – Different approaches to project work in statistics classes	79
14:15	Charanjit Kaur, Dr Ainura Tursunalieva – The effects of nudging on in-semester student learning behaviour and emotions: A case study of students at risk	85
14:30	Paul Fijn: Effective Questioning in "Interactive Lectures": An Alternative Approach	91
14:45	Sawsan Al-Shamaa – My experience in teaching statistics to international students in Auckland, NZ	95
15:00	CONFERENCE CLOSE	

PREFACE

OZCOTS 2021

10th Australian Conference on Teaching Statistics

OZCOTS 2021 theme: Statistics education in today's world

OZCOTS 2021 built on the success of the timing and format of OZCOTS 2008, 2010, 2012 and 2016 as a conjoint event with the Australian and New Zealand Statistical Conference (ANZSC). Initially, we planned it to be in July 2020 in Gold Coast, unfortunately due to COVID-19 pandemic, we deferred it to 2021 and then we held it as the first fully online Conference. ANZSC2021 and OZCOTS overlapped by one day on Thursday 8 July.

Every day the teaching and learning of statistics is becoming more important than ever to industry, government, business and for everyone in the society from cradle to nursing home. The roles of statistical understanding and statistical thinking are vital in all disciplines, increasingly driven by big data, evidence-based agendas, and technological advances which generate data as well as enabling more complex problem-solving, data visualisation and analysis. To avoid "lies, lies, big lies and statistics" we need to reach further in the society so that "lies" can be separated from "statistics".

The OZCOTS program included keynote and contributed papers, and discussions on issues across the statistical education spectrum of interest to the whole statistical profession. The program aimed to address challenges of the intersection of data science and statistics across different disciplines and learning strategies. It includes topics ranging across the curricula and technology for teaching introductory and undergraduate statistics; resources and online learning; statistics learning for postgraduates, researchers and workers; and research in the teaching of statistics.

The majority of OZCOTS participants came from Australia (n=80) and New Zealand (n=16) as intended with additional participants from US (n=3), Japan (n=2), and one participant from each of the countries listed: Finland, Austria, China and UK. Most of the participants were academics (n=81), some from government organisations (n=17) and a few from private sector (n=7).

Ayse Aysin Bilgin
OZCOTS Program Chair

OZCOTS 2021 Paper Refereeing Process

Papers referred to in the proceedings as refereed were reviewed and accepted as meeting the requisite standards by at least two referees selected from a panel of peers approved by the OZCOTS 2021 Conference Committee.

The Conference Committee took the view that the review of papers would give conference participants and other readers confidence in the quality of the papers specified as "refereed" in the proceedings. The refereeing process also provided a mechanism for peer review and critique and so contributed to the overall quality of statistics education research and teaching. While the refereeing process essentially relied on subjective judgments, referees were asked to compare the paper being reviewed against the accepted norms for reporting of research. It was expected that each accepted paper would represent a significant contribution to advancement of statistics education and/or the research process in statistical education. Authors verified that the refereed published papers for these proceedings were substantially different from papers that have previously been published elsewhere.

OZCOTS 2021 gratefully acknowledges the following referees for their assistance: Pip Arnold, Manfred Borovcnik, Daniel Frischemeier, Ian Gordon, Rossi Hassad, Rhys Jones, Sibel Kazak, Adam Molnar, Brian Phillips, Susanne Podworny, Alexey Ponomarenko, Tania Prvan, Amy Renelle, Alice Richardson, Eric Sowey.

OZCOTS 2021 – Keynote Speakers biographies

Professor Robert Gould

Robert is a teaching professor and vice-chair of undergraduate studies in the Department of Statistics at UCLA. He has been active in statistics education and data science education since 1994. As lead principal investigator of the Mobilize project, he is the architect of the Mobilize Introduction to Data Science course, a year-long high school course implemented in 16 school districts.

He is the founder of the DataFest, a 48-hour undergraduate data analysis competition sponsored by the American Statistical Association and held at 42 sites around the world. With two-year college professors Colleen Ryan and Rebecca Wong, he co-authored an introductory statistics book published by Pearson Higher Education.

Robert was elected Fellow of the American Statistical Association in 2012 and in 2019 was awarded the CAUSE Lifetime Achievement Award for Statistics Education and the American Statistical Association Waller Distinguished Teaching Career Award.

In: A. A. Bilgin and S. Budgett (Eds.). Proceedings of the Tenth Australian Conference on Teaching Statistics, July 2021 [Virtual]

Professor Helen MacGillivray

Prof Helen MacGillivray was only the second Australian and second female to be President of the International Statistical Institute in its 130 year history. She was an inaugural Australian Senior Learning and Teaching Fellow, first female President and Honorary Life Member of the Statistical Society of Australia, and a past President of the International Association for Statistical Education. She is Editor of *Teaching Statistics*, a Principal Fellow of the Higher Education Academy, and was inaugural Chair of the UN Global Network of Institutions for Statistical Training. She has received many national awards and grants, and published textbooks, chapters, keynotes, invited and refereed papers on authentic statistical learning and assessment, curricula design, learning support and research topics in distributions.

Prof Helen MacGillivray has been on organising committees for many international and national conferences. She has chaired reviews of university departments and centres across Australia and internationally. Her many leadership roles include founding and directing university-wide Maths Access Centres, Symposia in Statistical Thinking, and mentored developmental programs in university teaching. Helen has played key roles in mathematics and statistics school education on curriculum, professional development, resources, assessment and a variety of innovative and successful extension and enrichment programs for thousands of high school students.

Professor Manfred Borovcnik

Prof Borovcnik's scientific career started with the Bayesian controversy. As a consequence his interest grew in a *comparative* study of approaches towards inference. Based on his long-term experience as a statistical consultant, he experimented with students how projects in applied statistics can be used for education.

Prof Borovcnik's concern with distance studies resulted in the design of blended-learning courses in statistics for non-mathematical studies. To find ways to elementarise inference led him to investigate the potential of EDA and resampling in the 1990s. In probability, he was fascinated by ideas of Fischbein with his interplay between intuitions and mathematics. This emerged into projects on key concepts of probability and the twin character of probability and risk, which in turn initiated investigations on risk also related to decisions in health issues.

KEYNOTE 1

DATA EDUCATION IN PRE-COLLEGE. PROMISES AND CHALLENGES

Robert Gould
UCLA
rgould@stat.ucla.edu

Call it data acumen, data literacy, data fluency, data science or statistics. By any name, the time is right to design and implement data education pathways from the earliest ages through graduate school. Currently, in the United States, separate data education curricula are promoted by mathematicians, computer scientists, and statisticians with little coordination and agreement between parties. Based on experience with the Mobilize Introduction to Data Science course, a curriculum designed in partnership with the University of California, Los Angeles Department of Statistics, the UCLA Graduate School of Education and Information Sciences, and the Los Angeles Unified School district, we will discuss the promise of data science education. To fulfill this promise, fundamental challenges must be faced, including professional development and a clear sense of the desired outcomes of a data education

KEYNOTE 2

LEADERSHIP ACROSS THE EDUCATIONAL DIVERSITY OF STATISTICS AND DATA SCIENCE

Helen MacGillivray
h.macgillivray@qut.edu.au

Statistics is inextricably linked to all endeavours involving data, variation and uncertainty across disciplines, business, government and society. Statistics both serves and leads developments, and hence has inherent diversity and diffusion which are simultaneously its strength and vulnerability. Such considerations are not new, but have recently received increased commentary, as has the nature of statistical leadership, particularly in collaborative settings. The increasing spotlight on Data Science delivers opportunities for renewal of advocacy and realisation of the contributions of Statistics in all human endeavour. This is reinforced in general agreement ranging from Data Science CEO's to leading researchers, that Data Science is essentially collaborative, and that the Statistical Sciences are the heart of Data Science.

Nowhere is the above more relevant and important than across all levels of education. The lessons, both good and bad, from Statistics must be heeded in Data Science, and the nature of the Statistical Sciences must be revitalised in tandem with Data Science. This presentation probes past and current developments and challenges within the context of statistical leadership, and of 'greater' Statistics and 'greater' Data Science, and emphasizes how we can all contribute to statistical leadership across the educational diversity of Statistics and Data Science.

KEYNOTE 3

APPROACHES TO ELEMENTARISE STATISTICAL INFERENCE

Manfred Borovcnik
Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria
Manfred.Borovcnik@aau.at

The complexity in the concepts and the difficulties in the individual concept acquisition in statistical inference are well known. That has induced the search for new learning forms, such as the ideas of visualisation and simulation. Computer-intensive methods of the discipline of statistics have also served as incentive for didactic innovations. We compare two ways of elementarisation.

Informal inference may be used as a joint label for endeavours to simplify, visualise , or simulate the hypothetical model behind statistical inference. That means, the statistical model remains in the background but is still the target of teaching so that it forms the background for educational decisions. That also implies that the theoretical character of such models is visualised by simpler means. This way of elementarisation – as a transient stage – should create learning paths to the full complexity of statistical inference.

"Informal Inference" – going back to the computer-intensive methods in statistics such as Bootstrap and re- randomisation – is an educational approach that reduces statistical inference to methods solely based on resampling given data. The approach of "Informal Inference" reduces statistical inference completely to the observed data developing the methods solely based on resampling this data. "Shuffling" the data provides tests of significance of natural null hypotheses and intervals that mimic confidence intervals.

We illustrate both approaches and give a detailed discussion about the relative merits. We develop an analogy to decision making, in medicine and economy to clarify the tight connection of statistical inference to decision making. This analogy helps to understand the meaning and the restrictions of the methods. The abstract quality indices of methods of statistical inference get a natural interpretation within the discussed contexts. The framework of decisions also makes clear that probability points far beyond a simple frequentist interpretation and has to be captured by models though we make extensive use of simulation of the consequences of these theoretical models.

In conclusion, we suggest using resampling (Bootstrap and re- randomisation) as a transient stage to statistical inference but aim at ways of elementarising the full complexity of statistical inference. The examples in the presentation show how to build conceptual understanding and disclose the meaning of concepts by meta-knowledge based on simplifications of – the full complexity of – statistical inference. The advent of Big Data will not decrease the role of statistical inference as it does not replace inference by other methods but uses complex methods of statistical inference in a different way.

THE INFLUENCE OF UNREFERENCED INSTRUCTOR PRACTICES ON STUDENT LEARNING

Ian Gordon
Statistical Consulting Centre
University of Melbourne
irg@unimelb.edu.au

It is a truism that there are many modes of statistical learning. Competent statistical educators understand this and attempt to use a variety of approaches, such as verbal, symbolic and visual. Students learn by listening, seeing and trying things themselves; they learn individually and in groups. These insights are part of foundational and well-known frameworks in the theory of education.

Less well recognized, perhaps, is the role that the habits and practices of instructors play, particularly those that are not explicitly referenced, but which display and perhaps reveal the real thinking of the instructor. In some cases, neither the student nor the instructor may be aware of the learning that is taking place, which is potentially concerning: what if the messages being communicated are unintended and undesirable?

In this paper I explore such communication in statistical education. The simplest examples include the order in which topics are covered, the revealed habits when thinking about a statistical problem, and the "go to" steps in analyzing data. I address the need for a greater reflection by instructors on the impact of unreferenced practices and habits, and the connection to statistical thinking in context.

In: A. A. Bilgin and S. Budgett (Eds.). Proceedings of the Tenth Australian Conference on Teaching Statistics, July 2021 [Virtual]

ENGAGING DIVERSE UNDERGRADUATE COHORTS WITH DATA STORIES: INSIGHTS FROM FIRST-YEAR STUDENT-DRIVEN PROJECTS

Di Warren
University of Sydney
diana.warren@sydney.edu.au

It is widely agreed that real data stories are invaluable in engaging statistic undergraduates, with much literature on the value of contextualised data in domain-specific statistics courses. What remains unclear, however, is what stories to choose for large, diverse first year offerings. What type of data best motivates students to develop statistical literacy, when the cohort involves students from many different backgrounds, majors and courses?

This study focuses on a formative collaborative data project from a new suite of first year data science and statistics units at the University of Sydney, in which students choose their own data. Using the choice of data as a proxy for interest, we analyse the nature of the data over thousands of projects in different units. Interesting themes emerge, regarding source, nature, background, size, and the increasing need for data wrangling skills. We also suggest implications for curricula, including the co-creation of data stories with students.

In: A. A. Bilgin and S. Budgett (Eds.). Proceedings of the Tenth Australian Conference on Teaching Statistics, July 2021 [Virtual]

STATISTICS: YOUR TICKET TO ANYWHERE

Ayse Aysin Bilgin, David Bulger and Thomas Fung
Macquarie University, NSW
ayse.bilgin@mq.edu.au

A wide variety of university degree programs include a mandatory first-year statistics unit. However, students often can't see how statistics relates to their degree and future career, and therefore they are usually disengaged with their learning in such service units. As a result, it is common to see a large body of students in first-year statistics classes and very few in higher-level statistics classes.

In this presentation, we will share our successful initiative, "Statistics: Your Ticket to Anywhere!" We organized an information session for the high-achieving students in the previous year's first-year statistics classes. The main goal was to persuade students to adopt statistics as a second major; that is, not to poach students from other departments, but rather, to persuade students of the relevance of statistics to the careers they were planning. Our recent alumni prepared and delivered the majority of the presentation. They shared their journeys from first year to third year at university and then beyond. The alumni described how statistics improved their employability skills, such as problem solving and critical thinking, and showed how a range of professions depends on statistics to enable better use of the resources available to humanity.

In: A. A. Bilgin and S. Budgett (Eds.). Proceedings of the Tenth Australian Conference on Teaching Statistics, July 2021 [Virtual]

RELATIONSHIP BETWEEN STATISTICS ANXIETY AND FINAL MARKS IN INTRODUCTORY BIOSTATISTICS IN UNDERGRADUATE HEALTH SCIENCES STUDENTS

Marijka Batterham[1] and Pavel N. Krivitsky[2]
[1] Statistical Consulting Centre, University of Wollongong, Australia
[2]School of Mathematics and Statistics, University of NSW, Australia
marijka@uow.edu.au

The aim of this study was to determine if statistics anxiety was associated with performance in an introductory statistics subject in undergraduate health science students. Statistics anxiety was assessed using the statistics anxiety rating scale (STARS). Final mark in the undergraduate subject "Fundamentals of Biostatistics" was recorded. Of the 267 students enrolled, 109 (41%) consented to participate. Significant negative correlations were evident between the final mark and the statistics anxiety subscales for Worth of statistics, and for Computational anxiety, with higher scores in these subdomains are associated with higher anxiety. This indicates those performing better in the subject felt statistics was more worthwhile and had a higher computational self-concept. There was substantial missing data for some of the subscales so relationships were confirmed by multiple imputation. Principal components analysis was used to verify which questions were associated with the subscales. The results suggest that strategies to increase students perceived worth of statistics and their perceived computational skills may result in a better student performance and less anxiety.

INTRODUCTION

Statistics anxiety is defined as "the feelings of anxiety encountered when taking a statistics course or doing statistics analyses; that is gathering processing, and interpreting" (Cruise et al., 1985). There are a number of different scales for assessing statistics anxiety with the STARS, the statistics anxiety rating scale, (Cruise & Wilkins, 1980) being the most widely used. This scale measures six separate components: worth of statistics, interpretation anxiety, test and class anxiety, computational self-concept, fear of asking for help and fear of statistics teachers.

Statistics anxiety has generally been studied in graduate students with limited studies in undergraduate students primarily including assessments of psychology students (Nesbit & Bourne, 2018, Primi & Chiesi, 2018). Previous research in graduate students has shown that statistics anxiety is experienced by 80% of students (Onwuegbuzie & Wilson, 2003). The primary aim of this study was to assess statistics anxiety in undergraduate health sciences students. If statistics anxiety is associated with outcomes in undergraduate health science students and there are identifiable factors associated with this anxiety, course material or extra resources and support can be developed to reduce this anxiety. In addition, examination of the STARS revealed questions that appeared to not be relevant to our student population and therefore a secondary aim of the study was to investigate how students would respond to these questions (would they complete these questions) and whether responses in our sample followed the same subdomain groupings as those developed in the original scale.

METHODS

Students were recruited in 2017 and 2018 in order to meet the required sample size. 382 students enrolled in the subject, 267 were eligible to sit the exam (115 withdrew from the subject). Students were asked to participate in the study during their tutorial session, time was allocated at the end of the tutorial for students to complete the scale. Participation was voluntary. Students completed the scales which were then collected and stored by a staff member not associated with the subject until the completion of semester and release of marks. This allowed students to participate without concern that their responses may influence their marks in the subject. After the session completion subject marks and requested demographic data were released for consenting students by the University of Wollongong Information Management Unit. The study had institutional ethics approval. Multiple imputation (MI) was used to account for the missing data in the subscales. In multiple imputation, multiple datasets (10 in the study) are generated using fully conditional specification (an iterative markov chain monte carlo method) and analysed by standard methods. Estimates were combined using combined using Rubins' rules. To pool the imputed correlation coefficients they are first converted using Fishers Z

transformation and then back-transformed after pooling (Van Buuren, 2018), the pooled variance accounts for both the within and between imputation variability. In order to address the primary outcome assessing statistics anxiety and examining the relationship between statistics anxiety and subject marks and demographic characteristics, descriptive statistics were produced. Correlations, regressions and t tests were performed for both the original dataset and the pooled imputed datasets. The MI results give an approximation of the results of the full dataset and provide a sensitivity analysis for the complete case data. In order to address the secondary question about the appropriateness of the scale for the use in our sample principal components analysis with varimax rotation and extraction set to 6 factors (to be consistent with the original scale) was performed to verify the subdomains. Analyses were performed using IBM® SPSS® Statistics Version 25 (IBM Corp, Armonk NY). Eighty-five students were required to show significance based on a small effect using Cohen's criteria for correlation.

RESULTS

One hundred and nine students consented to participate in the study (41% of those eligible to sit the exam). There was missing data for all fields and numbers available for each variable are indicated in Table 1, Demographic Characteristics.

Table 1 Demographic Characteristics

Variable	Sample size (n) Mean(SD), Median(range) or frequency (%)
age	n=94 21.7(5.4), 20(18,49),
Gender **Male** **Female**	n=95 21 (22%) 74 (78%)
Student type **Domestic** **International**	n=94 91 (97%) 3 (3%)
Main degree type **Nutrition** **Exercise Science** **Public Health** **Medical & Health Science** **Exercise Science and Rehab** **Other Science**	n=94 41 (43.6%) 26 (27.7%) 6 (6.4%) 8 (8.5%) 9 (9.6%) 4 (4.3%)
Mark "Fundamentals of Biostatistics"	n=92, 77.4(10.6), 79 (17,93)
ATAR	n=61, 82.5(10.2), 84.9 (52, 97)

Table 2 contains the primary results for the analysis including the mean scores for the subdomains and correlation coefficients for the association between the subdomains and the subject mark. The total sample size of 91 for imputation reflects those who consented to the study, had an exam mark and completed at least some of the questions. Both complete case and imputed analyses are shown and produced substantially similar estimates with the results between the two differing more in the variables with more missing data as would be expected. There was no significant difference between genders in the scores in any of the subdomains for the raw subdomains, mean differences ranged from -3.48 (-7.03,0.07) for the test subscale to 0.26 (-8.21,8.73) for the worth subscale. There were also no differences in the pooled datasets for gender, mean differences and CIs from -2.53 (05.83,0.77) to 0.18 (-7.30,7.66), P values (0.132-0.962). Age did not correlate significantly with any of the subdomains in the original (r=-0.133, P=0.245 to r=0.034, P=0.776) or pooled data (r=-0.062, P=0.584 to 0.086, P=0.412). Given the small sample sizes in student type and degree type comparisons were not performed. Entry model regression with the 6 subdomains as predictors of mark showed that none of the subdomains were significant predictors, there was evidence of multicollinearity with VIFs around 3-5 for some subdomains (particularly the worth subdomain) for the original dataset and the

imputations, in addition condition indexes > 15 and variance proportions ~0.90 for some subdomains in the different imputed datasets indicated high multicollinearity, forward stepwise regression indicated that Computational subdomain was the only significant predictor of mark (pooled Beta = -0.486, t=-2.821, P=0.005).

Table 2. Descriptive statistics between subscales and exam mark

Subdomain	Mean(SD)	Imputed mean(pooled SD)	Correlations for each domain with mark	Correlations Pooled estimate (MI)
Worth of statistics	40.79(14.34) n=60	42.34(18.02) n=91	-0.311 P=0.016 n=60	-0.269, P=0.011 n=91
Interpretation anxiety	27.58(7.33) n=68	27.86(7.30) n=91	-0.152 P=0.215 n=68	-0.099, P=0.358 n=91
Test and class anxiety	27.27(6.81) n=66	28.15(7.40) n=91	-0.019 P=0.877 n=66	-0.062, P=0.564 n=91
Computational self-concept	16.65(5.49) n=79	16.01(6.05) n=91	-0.291 P=0.009 n=79	-0.289, P=0.006 n=91
Fear of asking for help	9.51(3.49) n=84	10.01(3.62) n=91	-0.079 P=0.473 n=84	-0.037, P=0.734 n=91
Fear of teacher	12.47(4.92) n=74	11.69(5.39) n=91	-0.165, P=0.159 n=74	-0.187, P=0.083 n=91

In order to address the appropriateness of the STARS scale in our sample the questions included in the scale, the subdomains they belong to, the number of missing responses and the PCA loading are shown in Table 3.

Table 3. STARS questions, number of missing responses and loading from PCA.

No.	**Question**	missing	Factor
	C= Computational Self Concept sub domain		
C25	I have not done maths for a long time. I know I will have problems getting through statistics	2	C
C31	I cannot even understand high school maths; how can I possibly do statistics?	8	C
C34	Since I have never enjoyed maths I do not see how I can enjoy statistics	6	C
C38	I do not have enough brains to get through statistics	3	C
C39	I could enjoy statistics if it were not so mathematical	4	C
C48	Statistics is not really bad. It is just too mathematical	4	C
C51	I am too slow in my thinking to get through statistics	6	L
	H=fear of asking for help sub domain		
H3	Going to ask my statistics teacher for individual help with material I am having difficulty understanding	4	H
H16	Asking one of your lecturers for help in understanding statistical analyses	4	H
H19	Asking someone in the computer lab for help in understanding output	0	L
H23	Asking a fellow student for help in understanding output	1	L
	I=interpretation anxiety sub domain		
I2	Interpreting the meaning of a table in a journal article	6	I
I5	Making an objective decision based on empirical data	2	I
I6	Reading a journal article that includes some statistical analyses	1	I
I7	Trying to decide which analysis is appropriate for my research project	15	T
I9	Reading an advertisement for a car which includes figures on km/litre, depreciation, etc	4	I
I11	Interpreting the meaning of a probability value once I have found it	1	I
I12	Having to enter data onto a computer package	0	I
I14	Determining whether to reject or retain the null hypothesis	1	I
I17	Trying to understand the odds in a lottery	11	L
I18	Watching a student search through a load of computer output from his/her research	12	L
I20	Trying to understand the statistical analyses described in the abstract of a journal article	2	L
	L=fear of statistics teachers sub domain		
L30	Statistics teachers are so abstract they seem inhuman	10	W
L32	Most statistics teachers are not human	13	W
L43	Statistics teachers speak a different language	4	W
L44	Statisticians are more number oriented than they are people oriented	7	W
L46	Statistics teachers talk so fast you cannot logically follow them	3	T

Table 3 (cont'd). STARS questions, number of missing responses and loading from PCA.

No.	Question	missing	Factor
	T=test and class anxiety sub domain		
T1	Studying for an examination in a statistics course	3	T
T4	Doing the coursework for a statistics course	0	T
T8	Doing an examination in a statistics course	12	T
T10	Walking into the room to take a statistics test	13	T
T13	Finding that another student in class got a different answer than I did to a statistical problem	0	T
T15	Waking up in the morning on the day of a statistics test	13	T
T21	Enrolling in a statistics course	3	L
T22	Going over a final examination in statistics after it has been marked	24	T
	W=worth of statistics sub domain		
W24	I am a subjective person, so the objectivity of statistics is inappropriate for me	22	I
W26	I wonder why I have to do all these things in statistics when in actual life I will never use them	5	W
W27	Statistics is worthless to me since it is empirical and my area of specialization is abstract	10	W
W28	Statistics takes more time than it is worth	4	W
W29	I feel statistics is a waste	7	W
W33	I lived this long without knowing statistics, why should I learn it now?	7	W
W35	I do not want to learn to like statistics	6	W
W36	Statistics is for people who have a natural leaning toward maths	6	C
W37	Statistics is a pain I could do without	2	W
W40	I wish the statistics requirement would be removed from my academic program	5	W
W41	I do not understand why someone in my field needs statistics	7	W
W42	I do not see why I have to fill my head with statistics. It will have no use in my career	6	W
W45	I cannot tell you why, but I just do not like statistics	7	W
W47	Statistical figures are not fit for human consumption	11	L
W49	Affective skills are so important in my (future) profession that I do not want to clutter my thinking with something as cognitive as statistics	11	W
W50	I am never going to use statistics so why should I have to take it?	10	W

No. refers to the question order, W=worth of statistics, I=interpretation anxiety, T=test and class anxiety, C=computational self-concept, H=fear of asking for help and L=fear of statistics teachers. Missing is the number of missing responses out of the 109 respondents and factor refers to the results of the Principal Component Analysis

DISCUSSION

The aim of this study was to assess whether statistics anxiety was related to the final mark in an introductory undergraduate biostatistics subject completed by health sciences students, and to assess if statistics anxiety was related to demographic variables. A secondary aim was to examine how students would respond to questions which may not be appropriate at their degree stage, for example asking questions about a statistics exam when they had not yet sat for a statistics exam.

In addressing the primary aim of assessing whether statistics anxiety was associated with performance significant negative associations were evident between the "worth of statistics" and "computational self-efficacy" subdomains and the final mark. With regard to the relationship between the "worth of statistics" and the final mark. "Fundamentals of Biostatistics" is a second year undergraduate subject and students generally take this subject before they have any exposure to research methods in their chosen discipline. Without a context in which to understand why the subject is necessary it is not surprising to see this relationship between the perceived worth of statistics and the exam mark. While a focus of the subject has been on presenting relevant examples to engage the students, these results indicate that further time should be spent demonstrating the use of statistics in health sciences. The mean score for this subdomain in our study was similar to that previously found in undergraduate Chinese students (40.12 (12.25), n=201 studying education (Liu et al., 2011)). This mean in the Liu et al., study and that in our study are higher than reported means in undergraduate students in London 32.22 (14.94), n=93 (Walsh & Ugumba-Agwunobi, 2002) and the USA 35.33 (13.70) n=191 undergrad/55 graduate students (Baloğlu, 2003), although lower than the mean (46.23 (no SD provided) of a summary of 10 studies in undergraduate psychology students (Nesbit & Bourne, 2018). The authors of the study in Chinese students identified that the higher worth score may also be related to these students having low exposure to empirical research.

Advanced high school mathematics is not a prerequisite for all our health science degrees. Some students struggle with calculator skills, for example being unfamiliar with exponential and logarithmic

functions and factorials. This also makes the sections on probability with hand calculations difficult for these students. Most students entering the subject have not previously used a statistical package, the subject involves a 90-minute practical session each week involving calculations and using SPSS®. Assessment tasks involve some calculations and using SPSS® to perform analysis and interpretation of output. The negative relationship between the subject mark and computational self-efficacy (where a higher score indicates more anxiety associated with computational/mathematical skills) may be explained by these factors and indicates that this is also an area where extra teaching resources should be developed. The computational self-efficacy scores were similar to previous studies in China (17.55 (4.86), n=210), London (14.12 (6.97), n=93) and the USA (15.99 (6.30), n=191 undergraduates/55 graduates) and lower than the mean in the 10 combined psychology studies (21.23, no SD provided).

There were no relationships between statistics anxiety and age or gender in this study. Sample sizes were not adequate to meaningfully investigate relationship between student status (domestic/international) and degree program.

In addressing the secondary aim of investigating the appropriateness of the STARS scale in our sample it should be noted that some of the questions, particularly in the Test and Interpretation subdomains had high non response rates (Table 3). Although the STARS scale has been used in other undergraduate student populations in our subject the students have not previously sat for a statistics exam or necessarily completed a research project, therefore some of the questions in these domains are not relevant for them. The overall response rate was low, with 34% (91) providing a response which could be used for imputation. Although demographic and academic data on the participants who did and did not complete the study are available in the University records the ethics approval for this study did not allow the collection of any information that was not explicitly consented too by each student. Therefore, while the data is available we were unable to conduct any non response analysis to investigate the bias in subject marks or demographics between those who did and did not participate in the study. Given that the missing data were substantial particularly for some subscales the provision of the imputed estimates and the similarity to the original data estimates suggests that the results are robust to the missing data. Providing the multiple imputation estimates allows us to address the uncertainty about the missing data by creating plausible imputed values and combining these to appropriately account for the within and between imputation variance.

Further work could reconsider the STARS with the less relevant questions excluded and perform factor analysis to determine the remaining subdomains. The Statistics Anxiety Scale (SAS) (Vigil-Colet et al., 2008) assesses statistics anxiety without assessing attitudes towards statistics and the Attitudes towards statistics (ATS) Scale (Wise, 1985) assess attitudes toward the field of statistics and the course. Both scales in addition to the STARS have previously been used in psychology undergraduate students in Australia and Singapore (Chew & Dillon, 2014) to assess their validity. The STARS has been shown to have superior validity and reliability and has been validated and used in many different countries (Chew et al., 2018), however it would be worth establishing whether one of these other scales may be more relevant for undergraduate health students.

In conclusion this study demonstrated that final mark in an introductory biostatistics subject for undergraduate health science students was related to their perception of the worth of statistics and their perceived computational and mathematical skills. Developing teaching resources to improve these aspects of the course may result in improved performance and a better student experience. Further research on developing or testing statistics anxiety scales specific for undergraduate students in a first statistics class may undercover relationships that were not apparent in the current research.

REFERENCES

Baloğlu, M. (2003). Individual differences in statistics anxiety among college students. *Personality and Individual Differences,* **34,** 855-865.

Chew, P.K.H. & Dillon, D.B. (2014). Reliability and validity of the Statistical Anxiety Scale among students in Singapore and Australia. *Journal of Tropical Psychology,* **4**.

Chew, P.K.H., Dillon, D.B. & Swinbourne, A.L. (2018). An examination of the internal consistency and structure of the Statistical Anxiety Rating Scale (STARS). *Plos One,* **13**.

Cruise, R.J., Cash, R.W. & Bolton, D.L. (1985). Development and validation of an instrument to measure statistical anxiety. In *Proceedings of the American Statistical Association annual meeting* Chicago Illinois.

Cruise, R.J. & Wilkins, E.M. (1980). STARS: Statistical Anxiety Rating Scale. Berrien Springs, MI: Andrews University.

Liu, S., Onwuegbuzie, A.J. & Meng, L. (2011). Examination of the score reliability and validity of the statistics anxiety rating scale in a Chinese population: Comparisons of statistics anxiety between Chinese college students and their western counterparts. *Journal of Educational Enquiry,* **11,** 29-42.

Nesbit, R.L. & Bourne, V.J. (2018). Statistics Anxiety Rating Scale use in Psychology students: A review and analysis with an undergraduate Sample. *Psychology Teaching Review,* **24,** 101-110.

Onwuegbuzie, A.J. & Wilson, V.A. (2003). Statistics Anxiety: Nature, etiology, antecedents, effects and treatments. A comprehensive review of the literature. *Teaching in Higher Education,* **8,** 195-209.

Primi, C. & Chiesi, F. (2018). The role of mathematics anxiety and statistics anxiety in learning statistics. in *International Conference on Teaching Statistics (ICOTS)* Kyoto, Japan.

Van Buuren, S. (2018). *Flexible Imputation of Missing Data*, Second ed. Baton Roca FL: Chapman & Hall/CRC Press.

Vigil-Colet, A., Lorenzo-Seva, U. & Condon, L. (2008). Development and validation of the statistical anxiety scale. *Psicothema,* **20,** 174-180.

Walsh, J.J. & Ugumba-Agwunobi, G. (2002). Individual differences in statistics anxiety: The roles of perfectionism, procrastination and trait anxiety. *Personality and Individual Differences,* **33,** 239-251.

Wise, S.L. (1985). The Development and Validation of a Scale Measuring Attitudes toward Statistics. *Educational and Psychological Measurement,* **45,** 401-405.

STATISTICAL SOFTWARE FOR NON-STATISTICIANS AND NON-COMPUTER PROGRAMMING STUDENTS IN EDUCATION AND SOCIAL SCIENCE DISCIPLINES: AN EVALUATION OF FOUR CONTEMPORARY TOOLS

Sedigheh Abbasnasab Sardareh[1], Gavin Brown[1] and Paul Denny[2]
[1]Faculty of Education and Social Work, University of Auckland
[2]Faculty of Science, University of Auckland
s.abbasnasab@auckland.ac.nz

Most doctoral students in education and social science disciplines struggle to become proficient in statistical analysis for multiple reasons (e.g., instruction, textbooks, motivation, attitudes, etc.). A factor that seems to have been overlooked is the design of software used for statistical analysis. Most software has not been designed for non-statisticians or students who are not familiar with computer programming. Hence, this paper explores human-computer interaction (HCI) factors in contemporary software tools to identify possible issues that contribute to or hinder successful statistical problem-solving. HCI factors include the range of statistical operations available; technical properties (user interface design, data visualization, entry, and manipulation); and usage properties (speed, ease of used, and efficiency). SPSS, RStudio, R Commander & jamovi software systems were selected for detailed evaluation. Analysis suggests that HCI factors are likely to interfere significantly with the completion of statistical tasks for doctoral students in education and social sciences.

In: A. A. Bilgin and S. Budgett (Eds.). Proceedings of the Tenth Australian Conference on Teaching Statistics, July 2021 [Virtual]

AUTOMATED, RECEPTIVE AND INTERACTIVE: A CLASSROOM-BASED DATA GENERATION EXERCISE

LANGAN, Dean & WADE, Angie
Great Ormond Street Institute of Child Health, University College London, 30 Guilford Street, London WC1N 1EH.
d.langan@ucl.ac.uk

It is easier to engage with statistics training when presented with examples from familiar subject areas. However, when teaching students of varying professional backgrounds, finding relatable examples can be especially challenging. Classroom-based data generation exercises offer a solution with students involved in the process from data collection through to choice and use of appropriate analyses. One such exercise that forms an integral part of an introductory statistics course is based on beermat (coaster) flipping, a popular pub game in the UK. We recently moved the data collection process online allowing students to enter data via smartphones. Furthermore, a web application has been developed using the shiny package in R. This application automizes data analysis and allows students to explore the results interactively and independently. The application comes to life with visual demonstrations of core concepts such as the central limit theorem and bootstrapping. This technology further engages students and the ensuing discussion comparing outputs and interpretation is a welcome addition to classroom interactivity. We present details of this exercise, focussing on use of the web application, example outputs, student feedback and guidance for best practice to maximise learning outcomes.

INTRODUCTION

Teaching statistics has transitioned from focus on probability theory and statistical inference as an abstract concept, to practical application and statistical reasoning in many applied disciplines (Bradsheet 1996, Smith 1998). Topics typically taught in today's classroom include the principles of good research design, data collection and generation of answers to research questions through data analysis. We have witnessed a shift towards cooperative learning; a technique giving students more ownership of their learning by allowing students to experience first-hand a quantitative research process (Garfield 1993). A natural synergy of this transition has been to engage students through group exercises that reflect a real quantitative research.

Smith (1998) states "students are more easily convinced of the power of statistical reasoning if they see it applied to questions that are interesting and real to them". In a homogenous classroom, a meaningful project is naturally one that relates to the group's primary research field. In a heterogenous group, such as those containing students from different programmes or short courses, hands-on data generation exercises are advocated. This type of activity gives students a sense of shared ownership over the data and provides motivation to learn universal statistical concepts applicable to many fields. Lee & Famoye (2006) state "data generated from students themselves tend to draw their attention and motivate interest more than a dataset disconnected from their everyday life".

In this paper, we present a collaborative exercise refined over a decade by the Centre for Applied Statistics Courses (www.uck.ac.uk/stats-courses), who are based in University College London (UCL) (United Kingdom). This exercise is run within a stand-alone statistics course titled "Introduction to Statistics and Research Methods" that typically attracts heterogeneous students. They are asked to play a game in pairs that involves flipping a beermat (known as a coaster in the US and Australia); a game made popular in pubs across the UK. Instructions are provided in *Figure 1* on how to successfully flip a beermat. The main research question explored during this course is – is beermat flipping ability related to height? Previously, this exercise was performed using paper-based data collection forms, and analysis was performed by a statistician (Koutoumanou & Wade 2017). As of 2019, the data is entered by students in an online database and analysis are now automated, receptive and interactive though a new web application developed using the *shiny* package in R (Chang et al. 2021).

We provide further details of the development of this exercise split into two parts; data generation and statistical analysis. We present example output, summarise anonymous feedback from students and reflect on its effectiveness measured in terms of learning outcomes. A teacher may wish to recreate the exercise in their own statistical course or simply gain inspiration for a similar exercise.

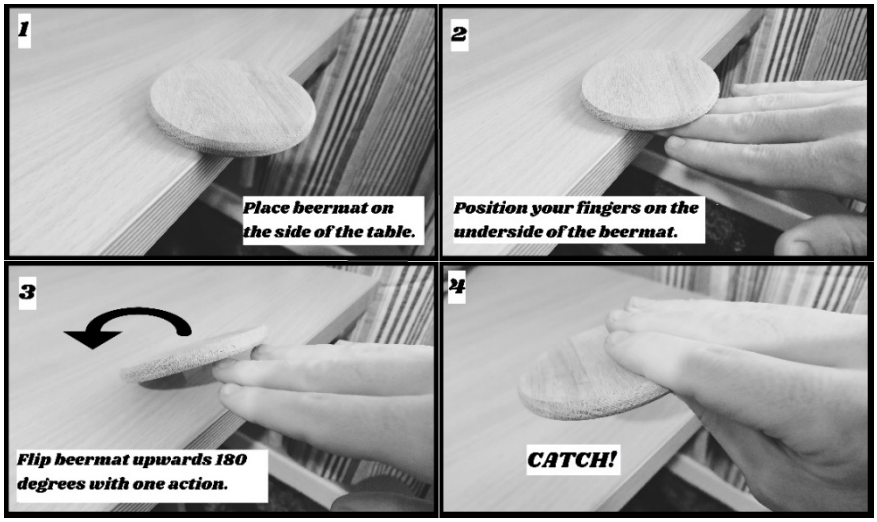


Figure 1: How to successfully flip a beermat

A DATA GENERATION EXERCISE

The development of this data generation exercise falls broadly into three phases: (1) Up to 2018, when data collection was paper-based, and analysis was performed manually through the SPSS software package. (2) In 2019, the data collection and analysis were automated with the intention of running the exercise with students face-to-face in the classroom. (3) From 2020, the exercise was adapted for online teaching due to the COVID-19 pandemic. Given the popularity of online courses, this format is set to continue after students can return to the classroom. In this section we provide details of what the exercise entailed for the student and the teacher at each stage.

1. Manual (2010-2018)

Students are handed a paper form and beermat and arranged randomly into pairs by calling out numbers in the classroom. The tallest person is asked to perform the task first. This is not the optimal study design and may introduce confounding bias (due to a learning effect), which the students are expected to recognise and thus provides a discussion point.

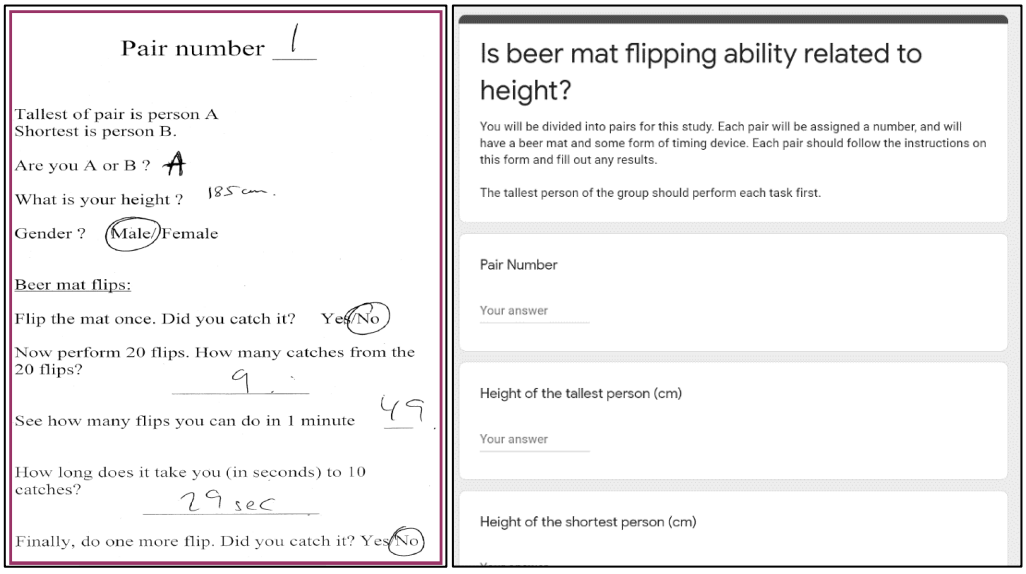


Figure 2: A paper data entry form collected from a student (left) compared against our new online form (right) (note: the questions have changed a little in the process)

The paper form is shown in *Figure 2 (left)*. Questions that students were asked on this paper form include (1) What is your height? (2) Gender? (3) Flip the beermat once, did you catch it? (4) Now perform 20 flips, how many catches from 20 flips? (5) See how many flips you can do in 1 minute. (6)

How long does it take you (in seconds) to 10 catches? (7) Finally, do one more flip, did you catch it? Questions 1-2 provide demographic details of the participants that can be used to explore relationships with subsequent measurements of beermat flipping ability (questions 3 – 7). We gather information in a variety of formats to subsequently demonstrate to students how the type of data impacts approach to analysis. For example, students are asked to perform one flip at the beginning and end of the exercise, providing within-person paired data so that improvement/deterioration can be investigated (questions 3 and 7).

2. Automated (2019)

Most stages of the exercise were automated in 2019, now requiring little time on the part of the statistician. Students can enter data directly into an online spreadsheet through a google form (as shown in *Figure 2, right*). As a side note, we now ask students whether they were born in the UK rather than gender, since this information is less sensitive.

3. Online (from 2020)

In 2020, the core introductory course moved online due to the COVID-19 pandemic, and so did the beermat flipping exercise. Given much of the exercise had been automated in the previous year, this meant coincidently that minimal changes were required to adapt to teaching and performing this exercise online. First, we recognised that not everyone will have access to a beermat in the home, so instead recommend using some form of credit or membership card instead. Arranging students into pairs can no longer be done physically, so instead we make use of the breakout rooms facility in Zoom (zoom.us). This arrangement runs much more smoothly that the equivalent task in the classroom. We found that such interactive exercises that allow students to connect is essential in a virtual environment to avoid disengagement with a lecture-style delivery.

STATISTICAL ANALYSIS

The development of the analysis phase of this exercise falls broadly into two time periods: (1) Before automation up to 2018, when the statistician performed analysis manually and students were provided with static printed results, and (2) After automation from 2019, when analysis is performed through a web application. Little has changed since providing online courses, so we describe the exercise within this setting in phase (2) alongside the equivalent classroom-based exercise. In this section, we provide details of what analysis entails.

1. Manual (2010-2019)

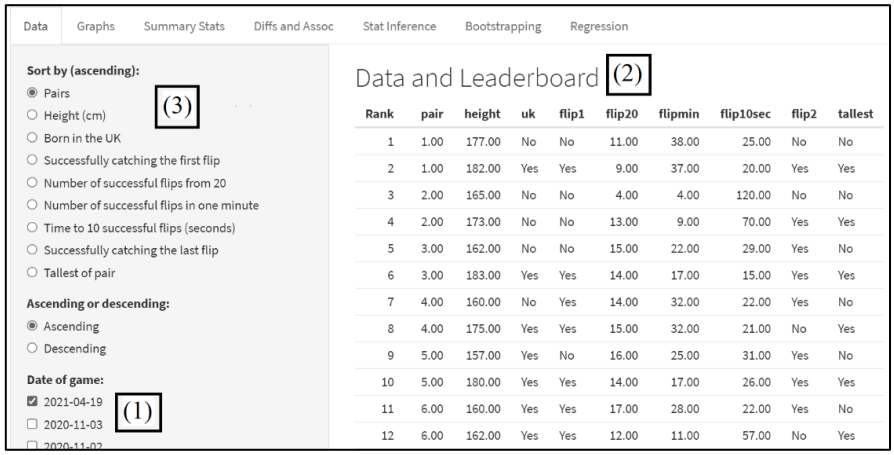
The paper forms are completed by students and entered manually into a spreadsheet by a statistician outside the classroom. SPSS syntax was created so that results can be generated automatically, and a report of the results is produced containing only graphs and tables. The report is printed for the students in time for the next class, where results can be discussed either as a whole class or in smaller groups. This exercise is either for revision in the final afternoon of the course or run in stages to give students a regular break from the core materials.

Students are presented with a series of questions that prompt them to relate the exercise to the course materials and relate readily to the type of questions the student might face in a real research project. These prompts touch on ideas of sampling bias, sample size, the limitations of pairing, appropriate analysis methods and whether the study could have been improved in some way. More specifically, there are questions relating to study design and those that require use of the printed output, such as:

- Could the order of performing the flips, tallest always first, have affected the results?
- How could the outcomes be displayed to show the relationship between ability to flip beermats and height?
- How could you assess whether the ability to catch a single toss improves with practice?
- How would you analyse the within pair differences to see whether there was a consistent tendency for taller individuals to have a better (or worse) ability to flip beermats?

2. Automated (from 2019)

As of 2019, the data (collected electronically) can be immediately incorporated into the analysis and results can be viewed in real-time through a web application developed using the shiny package in R (Chang et al. 2021). The application reads in the data from an online spreadsheet that automatically collects all entries from the data entry form. Screenshots are provided from this application in *Figures 3 and 4*. The application is available through the link tinyurl.com/OZCOTS-app and source code available via tinyurl.com/OZCOTS-code. Similar questions are proposed to the students in the classroom as those before the exercise was automated (see above). The key difference being that students can now access the application, make their own selections for choice of analysis and see the results appear in real time.



Rank	pair	height	uk	flip1	flip20	flipmin	flip10sec	flip2	tallest
1	1.00	177.00	No	No	11.00	38.00	25.00	No	No
2	1.00	182.00	Yes	Yes	9.00	37.00	20.00	Yes	Yes
3	2.00	165.00	No	No	4.00	4.00	120.00	No	No
4	2.00	173.00	No	No	13.00	9.00	70.00	Yes	Yes
5	3.00	162.00	No	No	15.00	22.00	29.00	Yes	No
6	3.00	183.00	Yes	Yes	14.00	17.00	15.00	Yes	Yes
7	4.00	160.00	No	Yes	14.00	32.00	22.00	Yes	No
8	4.00	175.00	Yes	Yes	15.00	32.00	21.00	No	Yes
9	5.00	157.00	Yes	No	16.00	25.00	31.00	Yes	No
10	5.00	180.00	Yes	Yes	14.00	17.00	26.00	Yes	Yes
11	6.00	160.00	Yes	Yes	17.00	28.00	22.00	Yes	No
12	6.00	162.00	Yes	Yes	12.00	11.00	57.00	No	Yes

Figure 3: Data and leader board tab of the web application

Figure 3 presents a screenshot of the 'data and leader board' tab of the application. First, students can select data belonging to any cohort of students from the left-hand side of the application (label 1). The most recent date is selected as a default since this date most likely belongs to the student's cohort, although any combination of dates can also be selected too to increase the sample size. Data collected on the selected dates then becomes the analysis dataset for all other tabs in the application. On this same tab, a leader board is presented (label 2), which can be sorted in order of any of the variables (label 3). The student might choose to order by some variable relating to beermat flipping ability to see how they fared in relation to their peers.

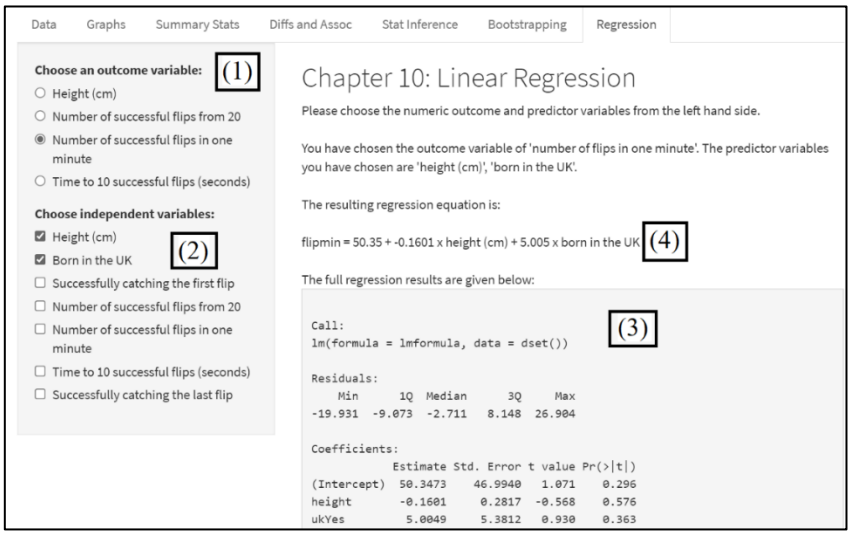


Figure 4: Linear regression tab of the web application

Figure 4 presents a screenshot of the application from the final 'regression' tab. We provide this as a typical example how students can interact and create results with other results tabs being set

up in a similar way. The students are able to select only a numeric variable as the outcome (since only linear regression is taught during the course) (label 1) and one or more independent variables that can take any form (binary or numeric) (label 2). In the screenshot, 'number of successful flips in one minute is chosen as the outcome and 'height' and 'born in the UK' as the two independent variables. Results are shown as they would appear in the R console (label 3) and also equivalent results pasted into an equation format (label 4). One of the strengths of these shiny apps is the ability to merge results into the text, providing students with a fuller explanation than you would typically see in the output for a statistical software package.

FEEDBACK

We present feedback that was collected anonymously via an online questionnaire after an iteration of the course in 2019 within a classroom. 28 of 35 students taking part in the course responded. Students were asked to rate the data generation exercise, the ease of data entry and the web application showing the results on scales of 1 (poor) to 5 (excellent). . There were 23 (82%) respondents that rated the data generation aspect of the exercise a four or above; the equivalent numbers in relation to the data entry form and web application were 23 (82%) and 22 (79%) respectively (*Figure 5*).

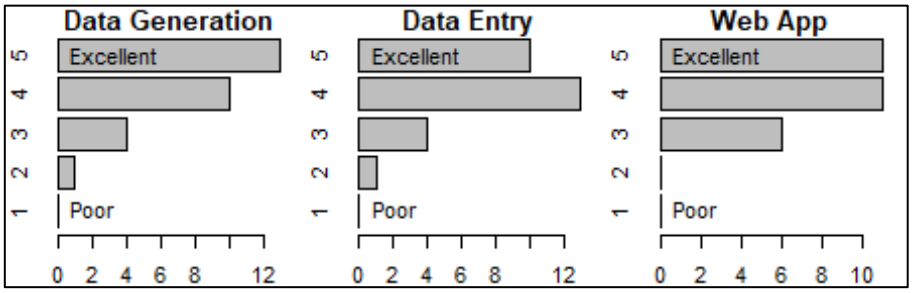


Figure 5: Feedback on data generation, data entry and the web application components of the exercise on a scale of 1 (poor) to 5 (excellent). Frequencies (x-axis) and score (y-axis) presented.

We received 15 free text comments, of which 8 (53%) were solely positive, 6 (40%) neutral and one (7%) negative. Of the positive comments, one student stated, "it was a great ice breaker, and being able to apply what we learned to this data (via [the] app) helped me to understand how to apply the concepts to real data". Another student commented "[I] thought this was really useful and I was really impressed by how well integrated it was throughout the course". The six neutral comments included four students that suggested more time allocated for the exercise, one student felt a simple written example without participation would have personally fulfilled the same purpose, and a final student stated, "I would like to see this being run on SPSS or Stata". Our students have very different preferences for statistical software and varying computer programming skills, but perhaps the application could provide access to the data for download and include the R code used to generate results.

CONCLUSIONS

Data generation exercises such as that demonstrated in this paper give students an opportunity to take ownership over their learning. These exercises give students shared interest, which is particularly challenging in a diverse classroom and facilitates understanding of steps involved in typical research processes. Our exercise encourages aspects of statistical thinking as defined by Wild (1999), closely mimicking a typical interrogative research cycle, where students interpret results and critically appraise them in light of the limitations of the data collection process.

Related classroom exercises that involve students generating their own data can also be found. Zeleke & Lee (2010) suggest research questions such as "is hand size a good predictor of height" and "how many raisins in a 0.5 oz. raisin box". Zetterqvist (1997) gives similar examples of exercises within a chemistry class, such as "determine the concentration of copper in a piece of impregnated wood". Other exercises focus on theoretical understanding. For example, demonstrating the central limit theorem by having students pick random numbers in small groups (Zacharopoulou 2006). These exercises typically involve manual data collection and analysis, setting a clear distinction from the beermat exercise demonstrated in this paper. For example, Zeleke & Lee (2010) ask volunteer students

to collect the paper data entry forms and create appropriate plots, but stated a limitation of their exercises were lack of time and recommend online resources for data collection to speed up the process. Applications developed through the shiny package, much like our application, are commonplace in statistical education as demonstrated in a review by Doi (2016). However, these are generally built to explain concepts such as the central limit theorem, rather than incorporate student-generated data.

Automating data collection and analysis saves time for lecturers and students, providing an opportunity for data collection and analysis to take place within the same class. These stages of the research process don't typically run as smoothly in real life and this point should be emphasised in class. One of the themes coming from feedback was that students like to see how analysis might be carried out in a particular software package; this can be challenging in a classroom filled with diverse student preferences, but highlights how providing the data, code and examples for different packages could further enhance the exercise. This would be a simple process and illustrates the value of student feedback to improve classroom activities and make them more inclusive. The feedback collected for this exercise meant that a full evaluation of the learning outcomes was not possible, but will be a valuable avenue for future research. The online environment has led to the rapid development of this useful tool, the benefits of which will also be apparent with face-to-face courses when these return.

REFERENCES

Bradstreet, T. E. (1996). Teaching introductory statistics courses so that non-statisticians experience statistical reasoning. *The American Statistician*, *50*(1), 69-78.

Chang, W., Cheng, J., Allaire J.J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, D., & Borges, B. (2021). shiny: Web Application Framework for R. R package version 1.6.0. https://CRAN.R-project.org/package=shiny

Doi, J., Potter, G., Wong, J., Alcaraz, I., and Chi, P. (2016), Web Application Teaching Tools for Statistics Using R and Shiny. Technology Innovations in Statistics Education, 9(1), 1–33.

Fawcett, L. (2018). Using interactive shiny applications to facilitate research-informed learning and teaching. *Journal of Statistics Education*, *26*(1), 2-16.

Garfield, J. (1993). Teaching Statistics Using Small-Group Cooperative Learning. *Journal of Statistics Education*, *1*(1).

Koutoumanou, E. & Wade, A. (2017). Students Generating and Using Their Own Data in a 5-day Basic Statistics Course. *United States Conference on Teaching Statistics (USCOTS, May, 2017).*

Lee, C., & Famoye, F. (2006). Teaching statistics using a real time online database created by students. *Proceedings of the Seventh International Conference on Teaching Statistics (ICOTS7, July, 2006).*

Smith, G. (1998). Learning statistics by doing statistics. *Journal of Statistics Education*, *6*(3).

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International statistical review*, *67*(3), 223-248.

Zacharopoulou, H. (2006). Two learning activities for a large introductory statistics class. Journal of Statistics Education, 14(1).

Zeleke, A., & Lee, C. (2010). Teaching introductory statistics using student generated data in a large class. *Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010).*

Zetterqvist, L. (1997). Statistics for chemistry students: how to make a statistics course useful by focusing on applications. *Journal of Statistics Education*, *5*(1).

PERFORMANCE OF ELEMENTARY STUDENTS ON STATISTICAL OPEN-ENDED ITEMS: LARGE-SCALE TEST IN CHINA

Yaoyao Dong and Jian Liu
Beijing Normal University, China
758727149@qq.com

Effective assessment of statistical open-ended items is an important part of Statistics Teaching. Based on the SOLO taxonomy, take the large-scale test data of fourth-grade students in D city in southern China as samples to deeply describe students' performance on the two categories of "pose questions" and "pose strategies" for elementary statistical open-ended items, and further explore the effectiveness of teachers' cognitive stimulation on students' statistical thinking level. Results show that: Compared with the "pose questions" category, students' performance on the "pose strategies" category is slightly inferior; A clear and complete written expression of mathematical language is an obstacle for students with low statistical thinking levels to advance; Most students have difficulty in contacting data context and have weak statistical reasoning ability; A higher level of teacher cognitive stimulation can help students improve their statistical thinking level, especially for students with low thinking levels. On this basis, relevant suggestions are put forward in order to provide enlightenment for the improvement of students' statistical thinking.

INTRODUCTION

The conclusions of statistical items are often probable. How to evaluate statistical open- ended items has become the "short board" and "difficult point" in the field of mathematics assessment. Among them, the fundamental difference between SOLO classification theory and traditional assessment lies in the openness of answers and more attention to the development of students' thinking. Based on SOLO taxonomy, Jones et al. (2000) assessed students' performance on statistical open-ended items, and determined four levels of statistical thinking for elementary students. This framework provides a methodological basis for this study to assess students' performance in answering statistical open-ended items. In the teaching of open- ended items, teachers often need more cognitive support and strategies (Huang et al., 2006). Specifically, teachers' effective cognitive stimulation strategies could help students experience a high-level cognitive learning process (Bao & Zhou, 2010), thereby enhancing students' thinking levels. Then in statistics teaching, to what extent teachers' cognitive stimulation can promote the development of students' thinking becomes another focus point in this study.

To sum up, based on SOLO taxonomy, the study qualitatively classifies and statistically process the answer performance of elementary students on statistical open-ended items, and present typical answer examples of students with different levels of thinking. We use large- scale data mining to diagnose the inadequacy of students' statistical thinking development, and provide strategies for the improvement of assessment results by exploring the role of teachers' cognitive stimulation.

METHODS

The analyses conducted in the current study were based on the data from a large-scale investigation in D city in southern China entitled "Regional Education Monitoring Project (REMP)", conducted in the fall 2020. We used Probability Proportionate to Size Sampling method, and first selected 330 schools and then randomly selected students from each school. After removing the participants with missing data for the relevant variables, the final sample included 30,075 students. Among them there were 16,923 (56.2%) boys and 13,162 (43.8%) girls. In order to effectively obtain the data, the researcher conducted the group test on a class basis. The test subject who voluntarily received the research group training entered the classroom, explained the test purpose and read the instructions. Then, the students began to finish the test.

The study relied on REMP to develop the mathematics test. The development of the test has gone through interviews, a small-scale test, a pre-test with 300 students, and independent review by external professional research group to ensure the quality of the instrument. The propositions in the field of "statistics" in this test adhered to the principle of "literacy-oriented" and focused on real situations. This study selected 3 items, including one "pose questions" item and two "pose strategies" items (see Figure 1).

Teacher's cognitive stimulation strategy was measured with the student's perspective, which was the teacher's cognitive stimulation strategy perceived by the students. The scale was adapted from the PISA 2012 Student Questionnaire, which consisted of 8 items (e.g., “The teacher asked us to explain how we answered the question”). Students responded on a 5-point Likert scale (1=Never to 5=Always).The higher the score, the better the teacher's cognitive stimulation strategy perceived by the individual. And the Cronbach α in the current study was 0.94.

16. In order to understand the students' lunch at school, Zhang conducted a survey of the students in Class 4, and the results are shown in Figure 1.

Figure 1. Statistics of lunch leftovers of class 4

A second survey was conducted on the 17 people with leftovers in the picture above. Among them, 5 people have small appetites, 10 people do not like eating, and 2 others, as shown in Figure 2.

Figure 2. Statistics of the reasons for leftovers

M4AS161: In Figure 2, draw the number of people who do not like to eat.
M4AS162: Please pose a question that can be solved with Figure 1 (just pose the question, don't need to answer).
M4AS163: According to Figure 1 and Figure 2, please make a suggestion for the school cafeteria.

16. In 2020, the new crown slips out, and the spread of the epidemic affects everyone. Figure 1 shows the statistical results of 35 new crown incident scenarios in Dongguan where incidents have occurred for 7 consecutive days.

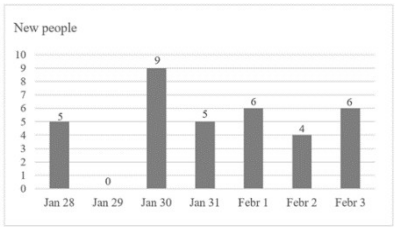


Figure 1. Statistics of 35 new confirmed cases of COVID-19 in Dongguan for 7 days in 2020

Further investigations were conducted on the causes of the 35 cases, and it was found that 8 of them were infected by taking a bus, 16 were infected by eating out together, 9 were infected by traveling, and 2 were infected because of activities in the community. Results are shown in Figure 2.

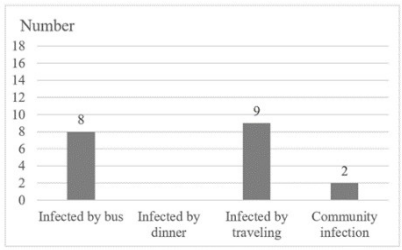


Figure 2. Causes of new cases in Dongguan for 7 days in 2020

M4BS161: In Figure 2, plot the number of cases due to “Infected by dinner”.
M4BS162: Please observe Figure 1 and pose a mathematical problem that can be solved with Figure 1 (just pose the question, don’t need to solve it).
M4BS163: Please combine the specific data in Figure 1 and Figure 2 to provide a basis for the citizens of Dongguan to prevent the epidemic.

Figure 1. Three statistical open-ended items in our study (M4AS162, M4AS163, M4BS163)

Based on the SOLO taxonomy and the statistical thinking level model of Jones et al., a coding analysis of students' answers was shown in Table 1. The test recruited 10 senior undergraduates majoring in science from a university. The online data coding time is 2 days, and the coders formally coded after half a day of training. Finally, we used Excel 2019 and SPSS.22.0 software to enter the coding results and perform descriptive statistical analysis of the data.

Table 1. The level description and examples of the three statistical open-ended items.

item	Idiosyncratic	Transitional	quantitative	analytical
M4AS162 (pose questions)	Blank; irrelevant answer; subjective answer; wrong answer.	Can pose questions, but the presentation is incomplete, or logically wrong.	Can pose simple questions and express clearly.	Can pose questions from a statistical point of view and express clearly.
M4AS163 (pose a strategy from multiple angles)	Blank; irrelevant answer; subjective answer; wrong answer, including not directed to the cafeteria.	The posed strategy is reasonable, but the expression is vague.	Can pose a reasonable strategy from one angle, and express it clearly.	Able to pose a reasonable strategy from two angles, and express it clearly.
M4BS163 (pose a strategy with a basis)	Blank; irrelevant answer; subjective answer; wrong answers.	The posed strategy is inaccurate and no reason is given.	Pose a strategy based on the specific context. The statement is clear, but no reason.	Be able to come up with strategies from a statistical point of view, and express them clearly.

RESULTS

In this part, we analyze the students' statistical thinking level on the three items, and the role of teachers' cognitive stimulation.

Assessment results of statistical open-ended items

Table 2. The level distribution of students on the three statistical open-ended items (%).

item	Idiosyncratic	Transitional	quantitative	analytical
M4AS162	12.6	4.8	57.6	25.0
M4AS163	22.2	19.4	52.5	6.0
M4BS163	17.1	37.7	44.1	1.1

Table 2 presents the level distribution of students on the three statistical items. The context of M4AS162 is the problem of leftovers in the canteen. Approximately 12.6% of the students are at the idiosyncratic level, only 4.8% of the students are at the transitional level, more than half (57.6%) of the students are at the quantitative level, and about a quarter of the students are at the analytical level. Students at the analytical level can establish the information between the data and make a comprehensive summary. Also, they can be aware of the statistical purpose of the graph, and often raise statistical questions from the key information. The following shows the typical answer examples of students at the analytical level:

"How many people have leftovers for lunch?"

"How many people can finish eating?"

"How many more people have leftovers than those who have just finished eating and those who haven't enough to eat?"

The context of M4AS163 is the problem of leftovers in the cafeteria. There are still 22.2% of students at the characteristic level, some students (19.4%) at the transitional level, more than half (52.5%) at the quantitative level, and only 6.0% at the analytical level. Students at the analytical level can understand the information contained in the data in the statistical graphs, consider not only the overall situation of the leftovers, but also the reasons for the leftovers. Also, they can consider the

problems more comprehensively, and can establish the connection between the data. The following shows the typical answer examples of students at the analytical level:

"The school cafeteria should ask for the opinions of students to cook the dishes that students like to eat, and prepare meals according to the appetite"

"I think it should be classified according to the appetite of each student and the food they love together."

The context of M4BS163 is the COVID-19 epidemic. This item requires students to put forward strategies "with a basis", that is, "strategy + reasoning", which is difficult for students. In total 17.1% of students are at the idiosyncratic level, 37.7% are at the transitional level, 44.1% are at the quantitative level, and only a very small number (1.1%) are at the analytical level. Students at the analytical level already have statistical reasoning skills, and they can "justify themselves" based on the data background. The following shows the typical answer examples of students at the analytical level:

Text reasoning: "The possibility of infection at dinner parties is very high. Please try not to dinner together."

Data reasoning: "Between January 28 and February 3, there were 35 new cases. According to the investigation, 33 new cases were infected after going out, so if you can go out, don't go out."

The improvement of teachers' cognitive stimulation

Table 3. Changes in the level of students by different teachers' cognitive stimulation (%).

item	Idiosyncratic		Transitional		quantitative		analytical	
	After 30%	Before 30%	After 30%	Before 30%	After 30%	Before 30%	After 30%	Before 30%
M4AS162	17.9	9.5	5.4	4.5	53.4	61.1	24.3	24.9
M4AS163	41.5	29.9	6.1	6.9	47.4	56.2	5.1	7.0
M4BS163	23.2	12.8	36.7	39.8	39.3	46.3	0.8	1.1

Table 3 presents the differences in the statistical thinking level distribution of the students in the three statistical open-ended items in the lower and top 30% group of teachers' cognitive stimulation. For example, "Before 30%" in the Table 3 means that teachers' cognitive stimulation scores are in the top 30%. On the whole, a good teacher's cognitive stimulation can help students at the idiosyncratic level overcome the subjective response orientation and develop to a transitional level and a quantitative level. It can also allow more students at a low level of statistical thinking to reach the quantitative level, but the proportion of students at the analytical level has not changed much.

In the "pose questions" category, teachers help students overcome obstacles in mathematics language expression, so as to clearly ask simple questions. However, teachers' cognitive stimulation is limited in helping students with high levels of thinking, and it is difficult for students to pose strategies from a statistical perspective. Compared with the "pose questions", teacher cognitive stimulation is more helpful for students to pose strategies, especially for students with idiosyncratic and quantitative levels. In particular, teachers can support students to express their own strategies clearly from one angle, but it is hard to help students put forward statistical strategies from multiple angles, and allow students to state their own reasons on the basis of posing clear strategies.

DISCUSSIONS AND IMPLICATIONS

The study is based on large-scale data in China, and uses SOLO taxonomy to operatively code the performance of students on statistical open-ended items, so as to diagnose students' development dilemma of statistical thinking, and further explore the function of teachers' cognitive stimulation to students' statistics learning. The main conclusions and implications are as follows.

Pay attention to mathematical expression and improve the ability to use mathematical language

The results show that a considerable part of Chinese students is still at a low level of statistical thinking, due to problems about "unclear and vague mathematics register". It means that if students

want to achieve a higher level of statistical thinking, such as quantitative and analytical levels, they must overcome the "mathematics register" problem. In daily teaching, teachers should not only allow students to give answers, but also encourage students to use mathematical language to explain their thinking and own mathematical understanding (Zhao, Li, & Wilkinson, 2018).

Mining the information contained in the data and comprehensively contacting the data context

Through the exploration of our study, it is found that most Chinese students have difficulty in answering statistical open-ended items in connection with the original context of the data. Consequently, they need to have a statistical perspective to think about problems, and be able to respond in the context of data. Data is the "number in context" (Franklin et al., 2007), and it is the context that gives the data meaning (Langrall et al., 2011). Therefore, statistics teaching should be especially aware of the importance of day context.

Develop critical thinking and improve students' statistical reasoning ability

Also, the study finds that students' performance on the "pose strategy" items is slightly inferior, and only a very small number of students can reach the analytical level. Especially in the "statistical reasoning with a basis", more than half of the Chinese students are still at the idiosyncratic level and transitional level, which means that we need to focus on the students' statistical reasoning ability. On the one hand, statistical reasoning requires students to have the awareness of using data and use "data" for reasoning. On the other hand, statistic concerns about uncertain phenomena, and the conclusions obtained by statistical reasoning are subjective and probable, which requires students to reflect and self-criticize.

Pay attention to teachers' cognitive stimulation, discover myths and expand students' thinking

Finally, the data tells that the degree of teacher's cognitive stimulation has a greater impact on students at the idiosyncratic level, and can effectively promote fourth-grade students to reach a quantitative level. It shows that a higher level of teacher cognitive stimulation can help students further improve their statistical thinking level, especially for students with low thinking level. It can be seen that teachers need to pay attention to their own cognitive stimulation of students.

Therefore, teachers can attach importance to the following two points: First, let students explain how they answer questions. This is not only to give students a process of thinking expression, but also an opportunity for the formation of generative educational resources. Teachers can use this to discover students' Myth. Second, encourage students to use a variety of different methods to answer, help students expand their thinking, and enhance mathematical cognition in different methods of communication and learning.

REFERENCES

Bao, J. S., & Zhou, C. (2010). Factor analysis of Affecting Students' High-level Mathematical Ability. *The Monthly Journal of High School Mathematics*, 9, 1–4.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education [GAISE) report*. Alexandria, VA: American Statistical Association. http://www.amstat.org/education/gaise/GAISEPreK-12_Full.pdf

Jones, G. A., Thornton, C. A., & Langrall, C. W., et al. (2000). A Framework for Characterizing Children's Statistical Thinking. *Mathematical Thinking and Learning*, 2(4), 269-307.

Langrall, C., Nisbet, S., Mooney, E., & Jansem, S. (2011). The role of context in developing reasoning about informal statistical inference. *Mathematical Thinking and Learning*, 13(1–2), 47–67.

Zhao, L. Y., Li, M. H., & Wilkinson, L. (2018). The Frontier Research on Mathematical Register in Mathematics Learning: An Interview with Louise Wilkinson. *Journal of East China Normal University(Educational Sciences)*, 36(6), 144-149.

APPROACHES TO ELEMENTARISE STATISTICAL INFERENCE

BOROVCNIK Manfred
Department of Statistics
Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria
Manfred.Borovcnik@aau.at

The difficulties in the concept acquisition in stochastics in general and in statistical inference are well known. The paper has two goals: To illustrate ways of elementarising – in the sense of Felix Klein – statistical inference and to compare two different approaches to elementarisation. Informal inference may be used as a label for endeavours to simplify, visualise, or simulate the model behind inference. That means, the statistical model remains the target of teaching. That implies that the theoretical character of models is visualised by simpler means. The elementarisation is viewed as a transient stage to statistical inference. "Informal Inference" – going back to the computer-intensive methods in statistics such as Bootstrap and rerandomisation – is an educational approach that reduces statistical inference to methods solely based on resampling given data. We illustrate both approaches and give a detailed discussion about the relative merits. The examples used show how to build conceptual understanding and enclose the meaning of concepts by meta-knowledge based on an elementarisation of – the full complexity of – statistical inference.

BACKGROUND

The complexity in the concepts and the difficulties in the individual concept acquisition in statistical inference are well known (Batanero, Chernoff, Engel, Lee, & Sánchez, 2016). That has induced the search for new learning forms, such as the ideas of visualisation and simulation. Computer-intensive methods of the discipline (Lunneborg, 2000; Efron, & Tibshirani, 1993) have served as incentive for didactic innovations. We compare two ways of elementarisation.

"Informal Inference" – going back to the computer-intensive methods in statistics such as Bootstrap and rerandomisation – is an educational approach that reduces statistical inference to methods solely based on resampling given data (Cobb, 2007). This approach reduces statistical inference completely to the observed data developing the methods solely based on resampling this data. "Shuffling" the data provides tests of significance of natural null hypotheses and intervals that mimic confidence intervals.

Informal inference may be used as a joint label for endeavours to simplify, visualise, simulate the hypothetical model behind statistical inference, or embed it into a suitable context. That means, the statistical model remains in the background but is still the target of teaching so that it forms the background for educational decisions (Borovcnik, 2019). That also implies that the theoretical character of such models is visualised by simpler means. This way of elementarisation should create learning paths to the full complexity of statistical inference.

The exposition of "Informal Inference" shows its advantages, as do many papers of the recent past. Yet, after a critical evaluation of the shortcomings of this approach, we suggest using resampling (Bootstrap and rerandomisation) only as a transient stage rather than replacing statistical inference by something new. The examples in this paper show how to build conceptual understanding and disclose the meaning of concepts by meta-knowledge based on an elementarisation of – the full complexity of – statistical inference. The advent of Big Data will not decrease the role of statistical inference as it does not replace inference by other methods but uses complex methods of statistical inference in a different way (Prodromou, 2017).

THE PROBLEM OF COMPLEXITY OF STATISTICAL INFERENCE

Probability without inference is meaningless, statistical inference cannot be understood without a sound comprehension of probability. This view changed curricula in the mid-1980s when after introducing probability, attempts followed to design learning paths towards statistical inference. It soon became clear that statistical inference would widen the focus on probability interpretations. Borovcnik (1996) considered resampling and non-parametrics as an intermediate state for learning paths towards the full complexity of inference. First attempts in the mid-1990s failed because of insufficient computer capacity. This changed after the Millennium. Cobb (2007) suggested replacing

statistical inference completely by resampling techniques grounded on a pure frequentist concept of probability. We describe first the problem of elementarisation in the sense of Felix Klein in general and then various attempts to find tractable approaches towards inference.

The complexity problem as a didactical challenge

Elementarisation is an old idea of mathematics teaching going back to Felix Klein:

> "There is a widespread understanding of the term "elementary", meaning […] something "simple" and not loaded with conceptual dimension – even somehow approaching "trivial". Connected, in contrast, with the notion of element, "elementary" means for Klein to unravel the fundamental conception. What is at stake, hence, is the notion of elements. […] The elements are understood as the fundamental concepts of mathematics, related to the whole of mathematics – according to its restructured architecture." (Klein 1908/2016, p. vi)

This notion of elements corresponds to the first reflections of d'Alembert on the nature of elements undertaken in the wake of Enlightenment how to make knowledge teachable and how to disseminate knowledge (Diderot & d'Alembert, 1751):

> "[d'Alembert] conceptualized in a profound manner […] how to elementarise a science, that is how to connect the elements with the whole of that science. […] to identify the elements of a science, or in other words, have rebuilt it in a new coherent way all parts of a science that may have accumulated independently and not methodically." (Klein 1908/2016, p. vi)

Different ways of tackling the complexity problem in statistical inference

a) Replace statistical inference by a different paradigm of generalisation: EDA (Exploratory data analysis) (Tukey, 1977). This has been perfectly received by the research community in the sense of a hypothesis-generating method; but it is less tractable for hypothesis testing as is required in statistical inference. The idea behind EDA relates to an interactive modeller who adapts the model step-by-step by interpreting results from intermediate analysis by the modeller's knowledge of the context of the problem. The insight into the result from context knowledge justifies the results. The inherent problem of EDA as a form of statistical inference lies in the circumstance that subjective acts of the modeller would be "forced" upon others who – as a usual reaction to it – would then reject the modelling and the result as relevant for them.

b) Teach different views and methods and learn from the differences in the same way as Barnett (1982) tried to evaluate the various schools of inference by his comparative statistics. A parallel approach in teaching was suggested by Vancsó (2009). The Bayesian way focuses on a decision between options rather than on inference and often on a discrete rather than a parametric model (e.g., the normal). The 1997 discussion in *The American Statistician* has been marked fiercely in favour of Bayesian methods (Witmer et al., 1997). Yet it was ended by Moore's (1997) "too difficult". Vancsó's (2009) uses software for the required complex calculations and visual interpretation for (prior and posterior) distributions. Key idea of a parallel approach is to understand the methods better due to the different ways to deal with the inference problem. Stangl (2017) advocates the Bayesian paradigm and gives suitable examples for optimising one-off decisions. The approach extensively uses computer facilities for calculations and graphing and it requires an intuitive understanding of distributions from graphs (Vancsó, 2018). Key are prior and posterior distributions as summary of the status of information: prior to data: qualitative information; posterior to data: prior and data combined. The quality of these distributions reaches far beyond a simple frequentist interpretation of probability.

c) Reduce the complexity of the statistical situation permanently: "Informal inference" (Cobb, 2007). Resampling embraces two different computer-intensive methods: Noether's (1967) non-parametrics that has gained attention because it is easy to simulate from a large number of combinatorially possible cases (rerandomisation). Computer facilities have reinforced the implementation of Bootstrap sampling from the first data set. "Informal inference" is an educational approach that copies the method of resampling and computer-intensive methods from the applications of statistics. Key idea is to simplify inference statistics. The approach has been extended to cover the curriculum of statistical inference over the secondary level and introductory statistics at universities

(Makar & Rubin, 2009; delMas, 2017; Ben-Zvi, Makar, & Garfield, 2018). A theoretical framework was elaborated in Zieffler, Garfield, delMas, & Reading (2008).

d) Informal ways to explore the full complexity of statistical inference. This approach originates from general teaching practice; it has been refined to a didactic position towards teaching statistical inference. Firstly, it comprises the use of analogue contexts or tasks that reveal the purpose of the methods and the character of the concepts: Concepts get a natural interpretation in the analogue (Batanero & Borovcnik, 2016). Medical or economic decisions are embedded in contexts, in which the concepts attain a natural meaning. Secondly, it uses illustrations, materialisations, and visualisations: The simulation shows effects of probabilistic models (regardless of the interpretation of probability) and the consequences of decisions. Thirdly, to use a simplified situation temporarily to pave the way for the full complexity (Borovcnik, 1996).

RESAMPLING AND BOOTSTRAP: „INFORMAL INFERENCE“

Wilcoxon rank test: significance test and p value

Task: Empirical proof of the efficacy of an antihypertensive drug by a placebo-controlled, randomised, double-blind clinical study. Target variable: Intra-individual difference of blood pressure Δ = sysBase – sys4Week [mm Hg]. Large values correspond to a great relief. *Hypotheses* are: Null (H_0): Verum (treatment) = Placebo (control); Alternative: Verum is better (or worse) than Placebo. If Verum is better, large values are expected under treatment compared to Placebo.

We introduce the Wilcoxon rank test or Mann-Whitney for independent samples to replace the usual t test. Let us suppose data for Placebo as 2.5, 0.9, 1.8, 3.6, and Treatment as 3.7, 5.2, 4.8, 6.1. We can rank the joint data from 1 to 8 (we could also use the original data). Rank 1 for the lowest value, rank 8 for the highest. The rank sum for the treatment group is then 26 and for the placebo 10. If we change the labels for treatment systematically, there are ‘8 choose 4’ = 70 selections of four data to form the treatment group with the rest forming the control (Fig. 1). Under the null hypothesis of no difference between treatment and placebo, all these selections (and the related rank sum) have the same justification or probability. The rank sum of the original sample is judged as a random result from all possible selections. From this, one can calculate the *p* value of the observed sample as 2/70 = 0.029 (two-sided).

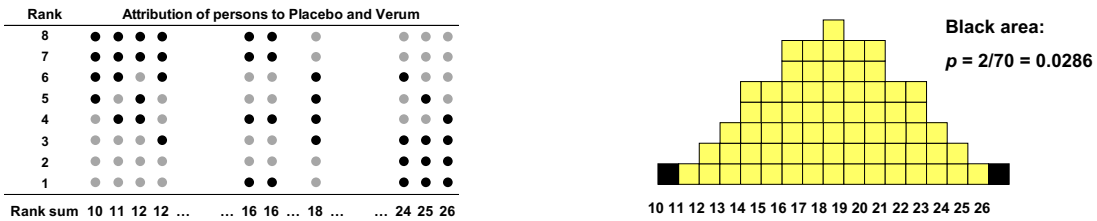


Fig 1. Possible rankings by rearrangement (left) – Probability distribution of rank sums under H_0 (right).

The significance test of the hypothesis of *no difference* between treatment and control leads to a rejection if the size of the test is 0.05 (usually set at 0.05 or 0.01). We conclude that the effect of the medical drug is significant at the 5% level.

The p value: some early concern

Let *p* denote the probability for an observed result (the exact one and ‘more extreme’ results) if H_0 applies. If *p* is smaller than 5%, the null hypothesis is rejected; *p* is the probability for a false positive statement, i.e., the test yields a significant result if the drug is not effective:

$$p = \text{P(Test provides a significant result | Drug is not effective).}$$

We have observed something that has less probability than 5% if H_0 applies (drug not effective). Therefore, we reject the null hypothesis. Yet, we are only interested in this figure:

$$\text{P(Drug is effective | Test significant)}$$

Doctors are not statisticians, but they should know the basics of scientific methods. Neyman and Pearson (1933) note “No test based on probability theory alone can provide valuable proof of the truth or falsehood of a hypothesis.” The following “argument” shows a frequent *misconception* (Borovcnik, 2019). We interchange event and condition and “get”:

p = P(Drug is not effective | Test significant) ⇒ $1–p$ = P(Drug is effective | Test significant)

If p was small, then $1–p$ is large, so the effect would be confirmed by a significant result. Yet, how large this probability is cannot be judged without reference to an alternative hypothesis!

The idea behind "Informal Inference"

The idea behind "Informal Inference" is to extract more information from the first sample by repeatedly taking samples (with replacement) from this given sample mimicking samples from the population. If the mean of each pseudo sample is taken (or any other characteristic), the pseudo-sampling process provides – in some way – an approximation for the sampling distribution of the mean (or this other characteristic). The process is called *Bootstrap* sampling, or Bootstrapping and the Bootstrap distribution is used like a distribution of data in descriptive statistics to deliver a Bootstrap interval that approximates a confidence interval.

As in the blood-pressure example above, rather than bootstrapping a single data set, the attribution of a statistical unit to one of two groups (treatment and control group) may be renewed to cover all possible reattributions. With the null hypothesis of no difference between the two groups, all re-attributions have the same justification or probability; as if there is no difference, one may attribute the labels of treatment and control in an arbitrary manner. That means, the finite set of all reattributions has a uniform probability distribution. As with larger samples, this theoretical universe of reattributions is large, one may take a random sample of it. Practically, this is done by a random reattribution (sampling without replacement) from the first data set. The method is called *resampling*. In the blood-pressure example, one would randomly select the four data for the treatment group (the others form the control) and calculate the related rank sum (or the difference of means if the data is investigated on the original scale). This resampling is repeated very often, which provides an approximation to the distribution of all combinatorially possible reattributions. From this distribution, it is easy to derive a significance test of the null hypothesis of "no difference" between the two groups.

The intention of informal inference is to embed the complex situation in statistical inference in a simple material setting (i.e., the data) leaving out any consideration about hypotheses except the natural null hypothesis of pure random effects on the statistical units. Examples for Bootstrap intervals and for significance tests based on resampling may be found in Borovcnik (2019), delMas (2017), or Stohl, Angotti, and Tarr (2010).

Inference about one "group"

If one data set is to be judged, e.g., for a parameter of location, a Bootstrap interval is provided by repeatedly sampling from the given data (always calculating this parameter). This resampling method provides an empirical basis (data) for the *statistical* measurement of this parameter. If a (hypothesised) parameter value falls outside the Bootstrap interval, it is "rejected" (Engel, 2010). The difficult part from an educational point of view is to justify why one can draw a new "sample" from the existing sample rather than from the population.

Example: We assume data for working hours for a seminar (data see Fig. 2). How accurate is the mean value of the sample as a measure of the population?

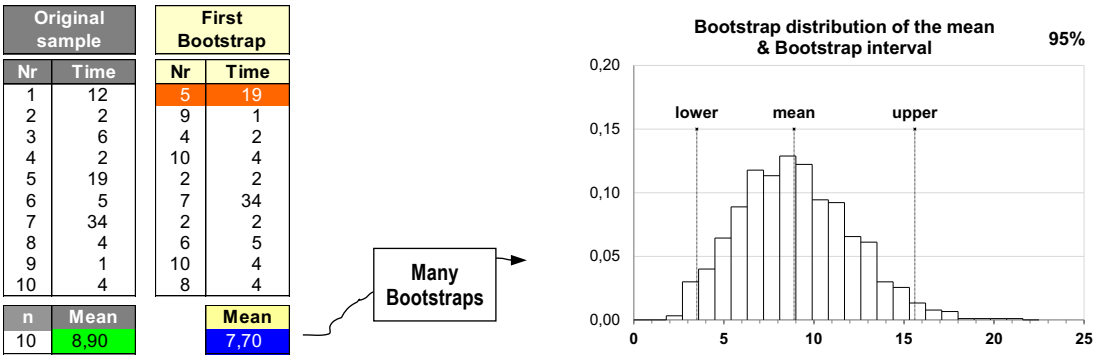


Fig. 2. Original sample for *work time for a seminar* and a first Bootstrap from this data (left) – Histogram of 1000 Bootstrap samples (right).

Instead of drawing from the population from anew, we select from the first sample (with replacement). The first Bootstrap provides a new measure of the population mean value, which does not differ too much from the original sample. We repeat the Bootstrap and get 1000 (or more) artificial measurements. The artificial data generated by this method reflect the *variability of repeated measurements* of the unknown mean of the population. From the Bootstrap distribution for the mean, we can cut the lowest and highest 2.5% to obtain the 95% Bootstrap interval, which is (3.80, 15.10) in our simulation scenario. This can be compared to the classic confidence interval of (2.46, 15.34). We see a good agreement between the two methods. *Yet, the interpretation is different*. The Bootstrap reflects the accuracy of repeated measurements of the population mean, while the confidence interval represents the population mean in 95% of the "repeated" samples. Similarly, Bootstrap can also be used to estimate other parameters. An exhaustive overview on Bootstrap intervals is in Pfannkuch, Wild, and Parsonage (2012).

Inference about two groups

If two data sets are to be compared for a measure of location (or any other parameter), then there are two options: First, resample from the given data on each group separately to derive the Bootstrap interval for this parameter; or, second, rerandomise the attribution of single data to one of the groups by a new random decision. If the null hypothesis of no difference between the two groups applies, then the data can be pooled and from this pool, the data for group 1 (and 2) can be randomly selected so that again an empirical basis of the statistic of interest is generated solely by the given data. The initial random attribution is randomly redone on the existing data, which reflects the natural null effect hypothesis (Stohl, Angotti, & Tarr, 2010).

Is a treatment effective in relation to a target variable? Treatment group receives Verum (TG), control group receives Placebo (CG). The re-randomisation offers an alternative to the two-sample *t*-test. The procedure is similar to the significance test from before. Instead of *ranking* the data, we analyse the *values* of the data here. The procedure is the same, but now we work with the original data and simulate *samples* from all permutations, because otherwise it becomes difficult to determine all permutations even with few data.

Under the null hypothesis of *NO DIFF,* it is intuitive that *any reassignment of people to treatments should have NO impact*. Therefore, we swap people randomly and the next treatment group consists of 8, 2, 3, 12, 7 and 1. The first re-allocation provides a new measurement of the difference in the mean values (as a measure of the treatment effect); the difference between treatment and control group in the original sample is 33.58, while the first reallocation yields a difference of –16.92 (see Fig. 3).

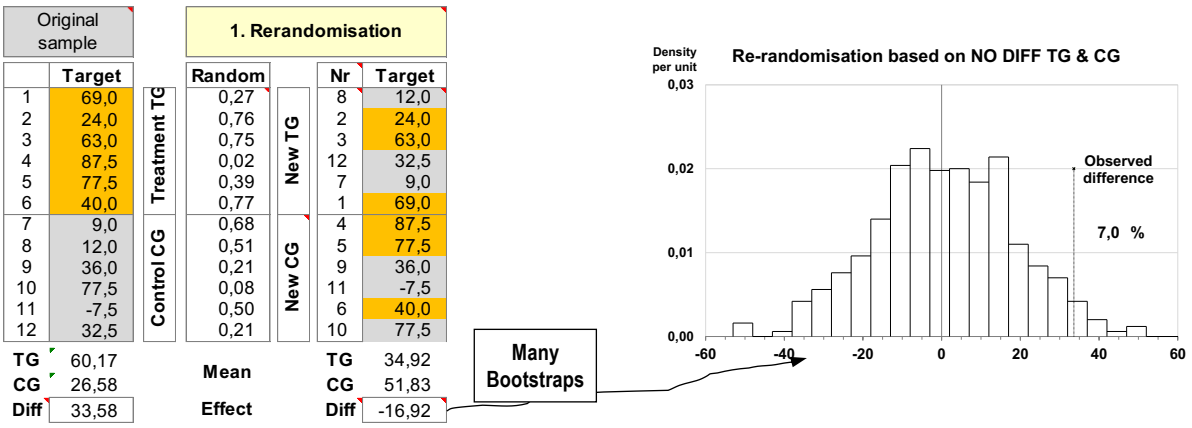


Fig.3. Left: Original sample in treatment and control groups and first rerandomisation of persons to treatment – Right: Histogram with 1000 re-attributions.

The distribution of repeated re-randomisation is shown in Fig. 3 (right); it represents the artificial results based on the *NO DIFF* hypothesis, i.e., the null hypothesis. We can insert the result of the first sample into this distribution and see that the *p*-value is 7.0% (two-sided). The entire simulation scenario can be repeated to show that the result is stable. This result can be compared with

the classic two-sample *t*-test, which gives 2.16, which corresponds to a *p*-value of 5.6% (assuming the same variances in the groups) or 2.16 (!) with 5.9% (for unequal variances). Here again, the similarity of the classic results with the re-attribution test is striking. The method may also be applied to any other comparison.

INFORMAL-INFERENCE EXPLORATIONS BY CONTEXT

We develop an analogy to decision making in medicine and economy to clarify the tight connection of statistical inference to decision making. This analogy helps to understand the meaning of the single elements and the restrictions of the methods. The abstract quality indices of methods of statistical inference get a natural interpretation within the contexts. The framework of decisions also makes clear that probability points far beyond a simple frequentist interpretation and has to be captured by *models* though we make extensive use of simulation of the consequences of these theoretical models.

Analogy to the medical situation

Borovcnik (2019) suggests exploring the situation in medicine, where there is always a decision that may lead to various errors whatever the decision is. A diagnostic test may be compared to a statistical test. A drug experiment is analysed by a statistical test. The analogy serves two directions: Statistical tests get better understandable by the context of medicine. Medical decisions get easier to understand by the superimposed structure of the statistical model.

Example: Diagnosing for a specific disease means to separate the groups of healthy and ill persons by a suitable variable for which the distributions do not overlap so much. Then, to introduce a cutting point (as in Fig. 4), which allows diagnosing a new person either as healthy or ill. The usual statistical key concepts have different names in medicine but are easy to understand and the interrelations between them are easily recognised as antagonistic by shifting the cutting point for the diagnosis. Yet, it becomes very soon clear that these key statistical figures cannot properly describe the risks of diagnosing a person wrongly, which heavily depends on the prevalence, the prior probability of persons that are examined to have that disease or not.

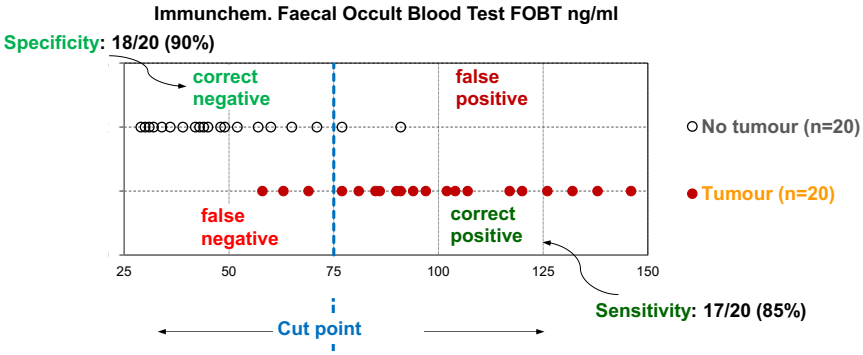


Fig. 4. A cut point separates the groups with diverging quality as measured by sensitivity ($1-\alpha$) and specificity (power, $1-\beta$); diagnosis of colon cancer by FOBT.

Some conclusions of the analogy to medicine are: The *p* value is not easy to interpret in a practical meaningful way. Diagnosing for diseases is a *decision problem*, which compares distributions under the scenario of healthy and ill people. There are always two diverging errors in the play: Diagnose for the disease when the person is healthy. Not recognise the disease despite the fact that the person has it. Several cut points for separating healthy and ill imply different sizes of these errors. There are diseases that are easy to diagnose. *There is a third error*: Whether the decision is a good one, does not only depend on cut points but also on the prevalence of the disease. Thus, we often do not get well interpretable coefficients for the quality of decisions.

Informal explorations of statistical tests or confidence intervals in economic contexts

An analogy to acceptance sampling and statistical process control shows that a frequentist interpretation of the errors of statistical tests is figurative. *The null hypothesis in acceptance sampling is at best a worst case of good quality* and the producer would normally deliver goods at much better

quality. No business partnership would work if 5% of all deliveries were rejected though they meet the quality arrangement between the two partners. Also, the worst-ever scenario that is used as placeholder for bad quality serves just as scenario to find a decision rule that allows an easy-going practice. The related type-II error has no further practical meaning. Details may be found in Batanero and Borovcnik (2016).

COMPARISON OF "INFORMAL INFERENCE" WITH STATISTICAL INFERENCE

Barnett (1982) has investigated the meaning and scope of various schools for statistical inference by a comparative statistics, in which he elaborated key concepts by which the approaches may be differentiated. We follow his criteria and extend the comparison towards the task of education of statistical inference.

There have been several endeavours to compare the various schools of inference starting from Barnett (1982). Key issues are alternative hypotheses and a comparison of different models that are represented by hypotheses. There is no way to introduce alternative hypotheses except by probabilistic assumptions and by simulation (or probability calculations). Any alternative cannot be resampled as is has not been sampled so that resampling fails to analyse alternative hypotheses and type-II errors. Rerandomisation allows only for a test of a null-effect hypothesis, Bootstrap has no direct conceptual link to significance tests. Modelling involves comparing scenarios (described by probability distributions). Hypothesis tests are comparisons of models (possibly restricted by a type of distributions). Thus, statistical inference implies a hypothetical approach. How the judgement of hypotheses is done lies at the core of the single school of inference. Whether it is done by classical or Bayesian methods, there is no link to it from resampling.

It is worthy to note, "There is a considerable body of research documenting students' difficulties understanding the structure of modus tollens and, consequently, interpreting *p*-values […]" (Makar & Rubin, 2018, p. 268). Yet, in the same chapter no mention is made about statistical alternatives and the type-II error, though the authors explicitly aim at "providing students with access to the power of statistical inference" (p. 262). The pure significance test has been disputed right since beginning (see also Hubbard & Bayarri, 2003).

Furthermore, simulation is completely misplaced for the problem of small probabilities. A problem, which is underestimated in statistics education (see Batanero & Borovcnik, 2016). In Bootstrap, a new error is introduced. If the first sample is not big enough, several regions of the distribution cannot be sampled well enough (to have the fine differences represented within the first sample, it takes too much data). If the first sample is big, then anyway the central limit theorem delivers better results. Furthermore, if one resamples, then the additional error is big unless one generates more than 10,000 re-samples. That makes it intractable for teaching.

From a didactical perspective, Biehler (2014) criticises that "[…] formal inferential reasoning as such is controversial itself […] This raises questions with regard to which view of formal […] inference we design […] informal inference activities for." Critique from the discipline comprises that Bootstrap differs from confidence intervals with no guarantee that they "converge" to them; i.e., they have other boundaries and coverage properties (Howell, n.d.; Lunneborg, 2000). Rerandomisation provides no substitute for the power of statistical tests, as there is no way to embed an alternative hypothesis in the method (Borovcnik, 2017). In summary, resampling as a pure approach replacing statistical inference (Cobb, 2007; delMas, 2017) fails to provide the solutions that are promised.

CONCLUSION

"Informal inference" goes beyond informally exploring probabilistic models by simulation; it aims to replace traditional statistical inference (Cobb, 2007). We give reasons why such a radical approach misses to develop the elements of statistical inference and that the full complexity of inference is required to deal with decisions under uncertainty. We see the potential of "Informal Inference" as a transient stage towards statistical inference.

Theoretical and applied concern

With "Informal Inference", it is impossible to address key issues of statistical inference (type-II error). With rerandomisation, we land at a pure significance test, which raises the problems of the interpretation of *p* values (see Hubbard & Bayarri, 2003). With Bootstrap, one provides intervals that mimic classical confidence intervals, which, however, have a different meaning and different

properties. Corrections are complex and destroy the simplicity of the approach (see Howell, n.d.). Furthermore, this approach fails with small probabilities, as small probabilities are not represented in the first sample from where the resampling starts. Overall, "Informal Inference" is NOT an informal approach to what the discipline of statistics calls inference. It presents a restricted approach to inference with no obvious links how to proceed from there to formal inference.

Educational considerations about "Informal Inference"

To ground accessible conceptions of statistical inference (Wild, Pfannkuch, Regan, & Parsonage, 2017) is an essential educational goal. "Informal Inference" seems very convincing but in the end, it leads to a restricted methodology that is a strict subset of statistical inference. "Informal inference" reduces all statistical activities to the data; no hypotheses are any longer involved. This may seem an interesting way to teach inference at first sight. Yet, there are several drawbacks. One is for statistical modelling, that connects data chance and context (Pfannkuch, Ben-Zvi, & Budgett, 2018); modelling provides *hypothetical descriptions* of the real situation that are the result of a modelling process and not the result of shuffling data. The other drawback is that probability is reduced to a pure frequentist concept leaving all Bayesian methods out of reach; a reduction of concepts that may lead to biased understanding as Carranza and Kuzniak (2008) have shown. Spiegelhalter (2014) refers to probability as a metaphoric entity, which goes far beyond a pure frequentist concept of probability as the basis for statistical inference. Related to it, thinking in scenarios (Borovcnik, 2019) is typical for the inference situation; such a way of thinking is precluded by the "Informal Inference" approach. Many didactical issues arise that reduce the value of the approach if taken as a *pure* approach replacing statistical inference. How to continue the curriculum within such a setting? There is neither a path from resampling to decision theory nor to Bayes methods. Furthermore, modelling is absorbed in simulation. This may result in *data as facts* while *models represent a hypothetical way of thinking*. Conceptual understanding differs from easier access and solving of tasks.

"Informal Inference" narrows the focus on probabilistic modelling later. Therefore, we propose to use resampling (Bootstrap and Re-randomisation) only as a transitional phase to statistical inference and focus on ways to appropriately elementarise the complexity of statistical inference. Statistical inference is characterised by thinking in hypotheses. The comparison of assumed scenarios dominates the interpretation of the elements of inference; besides a type-I error, the power is crucial for a proper understanding. Yet, the relevance of the statistical model in the background of inference is best judged with a prior probability of the null hypothesis at stake, which makes it clear that the Bayesian framework is essential for understanding the elements of statistical inference. The scenario character makes it also clear that small probabilities – as are characteristic of many reliability and risk considerations – cannot be judged by resampling techniques, as there are usually not sufficient data to cover such cases for a resampling solution. Such small probabilities have to be modelled by suitable probability distributions, and such models are void of a frequentist interpretation so that a wider conception of probability is required for a conceptual understanding of the elements of inference. We suggest using the potential of "natural" contexts, simulation, illustration of special cases, investigation of dynamic changes in conditions of the model, and visualisation of consequences of decisions.

REFERENCES

Barnett, V. (1982). *Comparative statistical inference (2nd ed.)*. New York: Wiley.

Batanero, C. & Borovcnik, M. (2016). *Statistics and probability in high school*. Rotterdam: Sense.

Batanero, C., Chernoff, E., Engel, J. Lee, H., & Sánchez, E. (2016). *Research on teaching and learning probability. ICME-13 Topical Surveys*. Cham: Springer International.

Ben-Zvi, D., Makar, K., & Garfield, J. (2018). *International handbook of research in statistics education*. Cham, Switzerland: Springer International.

Biehler, R. (2014). On the delicate relation between informal statistical inference and formal statistical inference. In K. Makar (Ed.), *Proceedings of the Ninth International Conference on Teaching Statistics*. The Hague: ISI.

Borovcnik, M. (1996). Trends und Perspektiven in der Stochastik-Didaktik [Trends and perspectives in the didactics of stochastics]. In G. Kadunz, H. Kautschitsch, G. Ossimitz, & E. Schneider (Eds.), *Trends und Perspektiven* (pp. 39-60). Wien: HPT.

Borovcnik, M. (2019): Informal and "Informal" Inference – Didactic approaches to statistical inference. In C. Batanero, J. Godino (Hrsg.), *Proceedings of the III International Virtual Congress on Statistics Education (CIVEEST)*. Granada.

Carranza, P. & Kuzniak, A. (2008). Duality of probability and statistics teaching in French education. In C. Batanero, G. Burrill, C. Reading, & A. Rossman (Eds.), *Joint ICMI/IASE Study: Teaching Statistics in School Mathematics.* Monterrey: ICMI and IASE.

Cobb, G.W. (2007). The introductory statistics course: A Ptolemaic curriculum? *Technology Innovations in Statistics Education* 1(1).

delMas, R. (2017). A 21st century approach towards statistical inference – Evaluating the effects of teaching randomization methods on students' conceptual understanding. In *Proceedings of the 61st World Statistics Congress*. The Hague: ISI.

M. Diderot, & J. d'Alembert (1751–1780). *Encyclopédie ou dictionnaire raisonné des sciences, des arts et des métiers*. Paris: l'Academie Royale des Sciences.

Efron, B., & Tibshirani, R.J. (1993). *An introduction to the bootstrap*. New York: Chapman.

Engel, J. (2010). On teaching bootstrap confidence intervals. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society.* Voorburg: ISI.

Howell, D. (n.d.). Resampling statistics: Randomization & Bootstrap. *Statistical page Howell*. www.uvm.edu/~dhowell/StatPages/Resampling/Resampling.html.

Hubbard, R. & Bayarri, M.J. (2003). Confusion over measures of evidence (p) versus errors (α) in classical statistical testing. *The American Statistician 57*(3), 171-182.

Klein, F. (1908/2016). *Elementary mathematics from a higher standpoint*. Berlin: Springer.

Lunneborg, C.E. (2000). *Data analysis by resampling*. Pacific Grove, CA: Duxbury Press.

Makar, K. & Rubin, R. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal, 8*(1), 82–105.

Moore, D.S. (1997). Bayes for beginners? Some reasons to hesitate. *The American Statistician, 51*(3), 254-261.

Neyman J. & Pearson E. (1933). On the problem of most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A, 231*, 289-337.

Noether, G. (1967). *Elements of noparametric statistics*. New York: Wiley.

Pfannkuch, M. Ben-Zvi, D. & Budgett, S. (2018). Innovations in statistical modeling to connect data, chance and context. *ZDM Mathematics Education, 50*, 1113–1123.

Pfannkuch, M., Wild, C.J., & Parsonage, R. (2012). A conceptual pathway to confidence intervals. *ZDM Mathematics Education, 44*, 899–911.

Prodromou, T. (2017). *Data visualization and statistical literacy for open and Big Data*. Hershey, PE: IGI Global.

Spiegelhalter, D. (2014). Probabilistic thinking. In: E.J. Chernoff & B. Sriraman (Eds.), *Probabilistic thinking: presenting plural perspectives* (Back cover). New York: Springer.

Stangl, D. (2017). Urging a paradigm change: Why and how to train introductory statistics students in Bayesian thinking? In *Proc. of the 61st World Statistics Congress.* The Hague: ISI.

Stohl Lee, H., Angotti, R.L., & Tarr, J.E. (2010). Making comparisons between observed data and expected outcomes: students' informal hypothesis testing with probability simulation tools. *Statistics Education Research Journal, 9*(1), 68–96.

Tukey, J.W. (1977). *Exploratory data analysis*. Reading: Addison Wesley.

Vancsó, Ö. (2009). Parallel discussion of classical and Bayesian ways as an introduction to statistical inference. *Intern. Electronic Journal of Mathematics Education 4*(3), 291-322.

Vancsó, Ö. (2018, July). How visualisation using software helps understanding classical and Bayesian statistics. *Invited paper "Teaching Probability in School – Understanding and Linking it to Statistics."* ICOTS 10, Kyoto. www.researchgate.net/profile/Oedoen_Vancso.

Wild, C.J., Pfannkuch, M., Regan, M., & Parsonage, R. (2017). Accessible conceptions of statistical inference: Pulling ourselves up by the bootstraps. *International Statistical Review, 85*(1), 84–107.

Witmer, J., Short, T.H., Lindley, D.V. Freedman, D.A., & Scheaffer, R.L. (1997). Teacher's corner. Discussion of papers by Berry, Albert, and Moore, with replies from the authors. *The American Statistician, 51*(3), 262-274.

Zieffler, A., Garfield, J., delMas, R., & Reading, C. (2008). A framework to support research on informal inferential reasoning. *Statistics Education Research Journal, 7*(2), 40–58.

DEFINING RANDOMNESS?

Amy Renelle[1], Stephanie Budgett[1], and Rhys Jones[2]
The University of Auckland[1], The University of Surrey[2]
amy.renelle@auckland.ac.nz

Defining randomness is notoriously difficult – even more so, if attempting to define randomness for high school students. Not only is randomness a strange phenomenon but needing to consider the appropriateness of definitions for the intended audience makes determining a cohesive definition near impossible. So, what definition do New Zealand secondary school mathematics and statistics teachers lean towards? And, more importantly, what are the foreseeable benefits and difficulties with teaching randomness using this definition? Respondents to an online questionnaire were asked to select which one of eight definitions most accurately described how they would define randomness. Of the possible options provided, two were deemed inadequate in the literature. Approximately one-fifth of participants selected one of these inadequate definitions, indicating evidence of misconceptions being held by some teachers. There is therefore potential for misconceptions to be transferred to students. If so, it seems important that a clear definition is used in classrooms, with the potential for tasks exploring the lexical ambiguity of randomness to be created.

BACKGROUND

It is well established that randomness misconceptions, such as the representativeness heuristic, can affect our understanding in statistics (i.e., see Tversky & Kahneman, 1974). High school students often exhibit randomness misconceptions and it is expected that these incorrect intuitions may have multiple origins. Along with possible sources such as biological explanations (i.e., intuition vs. reasoned thinking, see Kahneman, 2011), conflicting experiences (everyday vs. statistics classroom vs. other subject classes, see Pfannkuch & Brown, 1996), and how we learn (constructivism learning paradigm, see McLeod, 2019) (which are beyond this paper), lexical ambiguity and the inherent difficulty of defining randomness are thought to contribute to the presence of these misconceptions. It is also important to note that teachers may also hold randomness misconceptions (see Renelle et al., 2020) and likely have comparable origins for these incorrect intuitions.

As Batanero et al. (2016) stated, "[even] today, we find no simple definition that we can use unambiguously to classify a given event or process as being random or not" (pp. 34 – 35). Similar comments can be seen in papers by Bar-Hillel and Wagenaar (1991), Batanero (2015), and Nickerson (2002). Defining randomness is difficult, with numerous definitions being produced in an attempt to describe randomness succinctly and simply (see Table 1). Even then, it is challenging to choose a definition that can be applied to numerous examples that is written in such a way that is clear to students. In particular, Batanero et al. (2016) and Gougis et al. (2017) posited that participants who selected an Equiprobability Definition or a No-Pattern Definition of randomness likely held randomness misconceptions. As such, these definitions are deemed to be inadequate. The remaining definitions presented in Table 1 are appropriate definitions of randomness, with Predictability Definition 1 preferred for this study as this is promoted by the New Zealand Ministry of Education in the current mathematics and statistics curriculum for secondary school students. Note that the definitions in Table 1 have been edited for comparability as a clear definition is not always stated in the referenced papers.

Part of the difficulty of choosing a definition for randomness is the homogeneity of the term (Kaplan, Fisher, & Rogness, 2009) – it holds more than one meaning. While homogeneity is not uncommon in the English language, Nickerson (2002) noted there is a lot of difficulty caused by the numerous situations that the term randomness can be applied. For example, when discussing the ways in which "random" is used in relation to both a random process (i.e., tossing a coin) and a random product (i.e., a result of a random process), it is suggested that this can cause problems with the way in which randomness is talked about; while something being *more or less random* may make sense when talking about a random product, it would be inappropriate when considering a random process, which is either random or not! Kaplan, et al. (2009) suggest that statistical words with everyday counterparts can endorse incorrect assumptions about statistical concepts. While the term "random" can be used as an adjective to accompany other concepts (i.e., random process, random product), this paper is concerned with defining the noun "randomness" because this is the term introduced at high school.

With a constructivism-led teaching approach, where the emphasis is on student-centred learning, focusing on building knowledge, active-learning and social interactions (McLeod, 2019), the way in which randomness is discussed is important for an understanding of the concept. Students may be used to hearing the word random in an everyday context referring to something that is surprising or unusual – "I saw this random duck in my garden". This is not the same kind of randomness typically considered in statistics classrooms. Kaplan et al. (2009) found that bringing students' attention to the multiple meanings of randomness was necessary (contrasting the colloquial use and statistical use). Pfannkuch and Brown (1996) described encountering "…a clash between [participants'] intuitions and probabilistic thinking" (p. 4) when exploring students' perceptions of probability and randomness. Lexical ambiguity could be a contributing factor– whereby students recognise throwing a die is a random process but feel that an outcome of HTHTH is neither surprising nor unusual. Students' experiences of the randomness in an everyday situation therefore might not align to the use of randomness within a classroom setting.

Table 1. Definitions of randomness adapted from the literature reviewed.

Definition Label	Definition	Reference
Equiprobability Definition	Randomness is where each observation is equally likely to be selected (inadequate definition).	Batanero et al. (2016); Batanero (2015)
No-Pattern Definition	Randomness is where a sequence lacks a discernible pattern (inadequate definition).	Gougis et al. (2017)
Subjective Definition	Randomness is dependent on a person's knowledge.	Batanero et al. (2016); Batanero (2015)
Zero-Correlation Definition	Randomness is where the correlation between pairs of adjacent observations is zero.	Nickerson (2002)
Algorithmic Definition	Randomness is where no algorithm can predict future observations of a sequence.	Batanero et al. (2016); Batanero (2015)
Compressibility Definition	Randomness is where a sequence cannot be compressed or compacted into a shorter form.	Chaitin (1975)
Predictability Definition 1	Randomness is where the outcome cannot be predicted even though the probability of each observation is fixed (curriculum definition).	(New Zealand Ministry of Education, 2012)
Predictability Definition 2	Randomness is where it is impossible to predict when an observation will occur.	(Bennett, 2011)

Furthermore, it is important to note that, in reviewing numerous mathematics and statistics curricula from countries around the world, a definition of randomness is rarely given, let alone highlighted as a lexically ambiguous term. Furthermore, students would likely come across randomness outside of the statistics classroom when learning about, for example, genetic drift or mutations (biology; Martin & Hine, 2008), radioactivity or lasers (physics; Daintith, 2009), dispersion (geology; Allaby, 2013), and diffusion (chemistry; Daintith, 2008). It is feasible that a lack of consistency between school subjects could also be a potential source of randomness misconceptions – especially as randomness appears to be rarely defined in these other fields. For example, although randomness is referenced in relation to different terms within biology, chemistry, and physics dictionaries (Daintith, 2008; 2009;

Martin & Hine, 2008), no definition of randomness is presented. With so many different meanings, how can we expect students to recognise which definition of randomness is appropriate for these contexts?

Considering Table 1, it seems apparent why the New Zealand Ministry of Education selected Predictability Definition 1 as preferred for secondary school students. The definition acknowledges that probabilities are fixed but that the outcome is still unknown. By comparison to the Equiprobability Definition, the curriculum definition allows for different events to occur with different or unknown probabilities. For example, while flipping a fair coin has a 50:50 chance of heads and tails, the Equiprobability Definition is inadequate for asymmetrical devices where one face of an object may not have the same probability of occurring as another face. Predictability Definition 1 can also be applied to short-run or long-run trials. While the No-Pattern Definition may be acceptable for long-run sequences, whereby we would be surprised by a perfectly alternating sequence of 100 heads and tails, it is unsuitable for short-run trials that can often generate a patterned sequence. Hence, the Equiprobability and No-Pattern Definition are deemed "inappropriate" or "inadequate" for describing randomness (Batanero et al., 2016; Gougis et al., 2017). While the other definitions in Table 1 are acceptable, it would be fair to suggest that Predictability Definition 1 is more consistent with other New Zealand statistics curriculum content and would require introduction to fewer new ideas than for other definitions such as algorithms and matrix correlation.

Predictability Definition 1 appears most suitable for a high school audience but the proposal of a fixed probability brings to question whether there are circumstances where randomness is present without fixed probabilities or perhaps with unknown probabilities that could be fixed or could be changing. Currently not promoted in online resources for New Zealand high school classrooms (New Zealand Ministry of Education, n.d.), it may be beneficial for students to experience examples of randomness where the probabilities are unknown. Such examples could then lead onto discussions of whether probabilities are constant or changing in various scenarios and connect to simulation-based exercises.

Hence, the foreseeable benefits to teaching randomness using Predictability Definition 1 greatly outweighs the narrowness and barriers arising from using the Equiprobability or No-Pattern Definition of randomness. The limited view of randomness presented by the Equiprobability and No-Pattern Definition could potentially contribute to students' misconceptions of randomness. Teachers using these inadequate definitions could limit students' acceptance of randomness when outcomes are inconsistent with expectations. Not only is there a disconnect between colloquial uses of the term random but use of an inadequate definition may result in some random scenarios being dismissed by students if they are unrepresentative of expectations generated by these definitions. If teachers introduce inadequate definitions of randomness, it is likely that their students' perceptions of randomness may be restricted and impact on their statistical understanding. As a result, we suggest that, if teachers have a preference for flawed definitions, it may be reasonable to expect students would also be using these definitions of randomness.

STUDY

Using Qualtrics, an anonymous online questionnaire was sent to a sample of New Zealand mathematics and statistics secondary school teachers in October 2019. The volunteer sample was recruited via the mailing lists of some New Zealand mathematics and statistics associations, and, after data cleaning, there were 150 participants included in the sample. The research question we are interested in investigating is: *What definition of randomness is most commonly selected by participants and how many select an inadequate definition?*

The participants were presented with the definitions presented in Table 1 (the middle column, excluding the inadequate and curriculum definition labels) and were asked to: *Select one of the following definitions that most accurately describes how you would define randomness.*

RESULTS

The majority of participants (n = 80) selected the Predictability Definition 1 (Figure 1). This is unsurprising as it is expected to be most familiar to participants given this is the definition specified in the New Zealand mathematics and statistics curriculum. Concerningly, 36 participants (almost a quarter of participants) selected an inappropriate definition – either the Equiprobability Definition or No-

Pattern Definitions. The remaining participants selected one of the other definitions, mostly preferring the algorithmic definition.

Participants were also asked if they used the curriculum definition of randomness when teaching. They were not informed of the curriculum definition to allow for comparison to the definition they selected. Just over a third of participants (n = 53) suggested they neither agreed nor disagreed that they use the curriculum definition of randomness, which may suggest either unfamiliarity with the definition, that the participant chooses not to present a definition in their class, or that they don't teach randomness. The frequency of responses to this statement can be seen in Figure 2. Interestingly, some participants who had strongly agreed that they use the curriculum definition in their classrooms then selected a different definition. Of particular concern, some of these participants selected an Equiprobability Definition or a No-Pattern Definition of randomness, which may mean they believe these are the definitions supported by the curriculum.

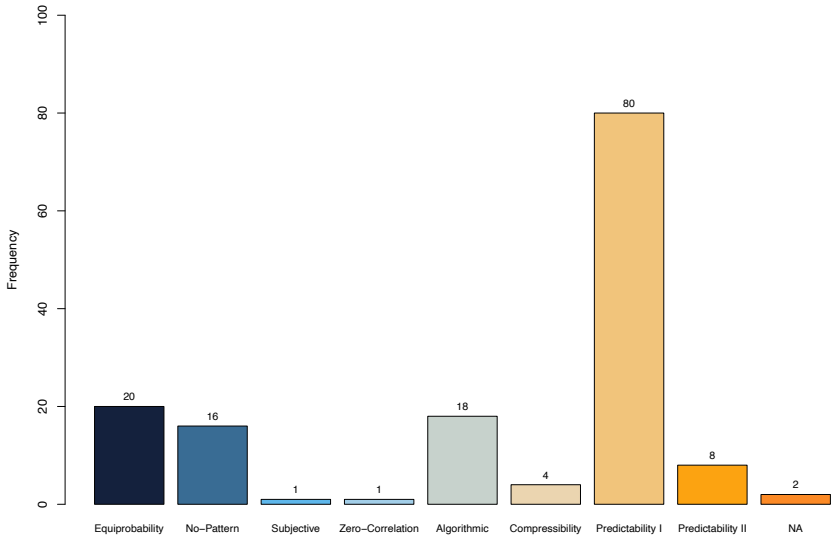


Figure 1. Frequency of Responses to *Select one of the following definitions that most accurately describes how you would define randomness.*

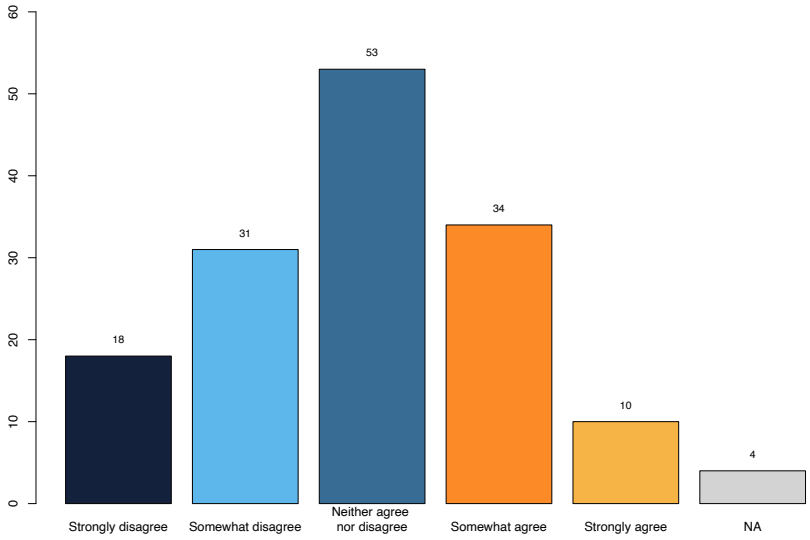


Figure 2. Frequency of Responses to *The definition of randomness provided by the curriculum glossary is what I use in my classroom.*

DISCUSSION

While the majority of participants selected the preferred definition for New Zealand secondary school students, many participants selected an inadequate definition. In particular, it is cause for concern that some of these participants felt that an inadequate definition was one promoted by the New Zealand curriculum. Defining randomness is difficult, and the lexically ambiguous term needs to be carefully

discussed with students. As found by Kaplan et al. (2009), it may be necessary to bring students' attention to the multiple meanings of randomness. Furthermore, the current findings suggest some New Zealand secondary school mathematics and statistics teachers may also need attention brought to the difficulties of defining randomness.

Due to the homogeneity of the term, a more globally applicable definition than offered by the Equiprobability and No-Pattern Definitions is necessary to help clarify the ambiguity of randomness. Because it applies to so many situations, randomness and suitable definitions for different examples should be explored in classrooms. Furthermore, it is important that educators are clear about how randomness should be defined in different situations and clarify that one definition may not be applicable in a different scenario. Consistency between subjects where randomness is intended to be discussed in the same way is necessary.

While participation in this study was limited, it appears New Zealand secondary school teachers generally prefer the curriculum-promoted Predictability Definition 1. However, it is clear that difficulties defining randomness are present and suggest there is the potential for tasks exploring the lexical ambiguity of randomness to be created. Professional development for teachers across different fields is needed to highlight the importance of defining randomness purposefully. Tasks focusing on defining randomness could then be used in classrooms in an effort to contribute to a better understanding of randomness for students. In particular, as Kaplan et al. (2009) suggest, tasks should bring attention to the multiple meanings of randomness, contrasting the colloquial use and statistical use. Further clarification as to why the Equiprobability and No-Pattern Definitions are inadequate are also necessary so a task comparing the benefits and flaws of various definitions may be a suitable starting point for lexical ambiguity to be minimised.

However, extending this to fields other than statistics may be challenging. Current attempts to reach out to biology, chemistry, and physics secondary school teachers in New Zealand has resulted in limited participation. This lack of engagement could indicate that randomness is not considered important which would suggest a potential challenge with incorporating the benefits, barriers, and examples of definitions relating to these fields. Despite this, it seems necessary that the use of the term randomness in these other fields is considered to help promote a consistent definition of randomness. To accomplish this, it may be valuable to reach out to experts in biology, chemistry, and physics to obtain a better idea of how randomness is discussed in these fields. Tasks acknowledging the ambiguity of randomness that could be implemented across the various fields using the term are in development, with further investigation into the impact of clearly defining randomness to be explored.

REFERENCES

Allaby, M. (2013). *A Dictionary of Geology and Earth Sciences*. Oxford Reference. https://www.oxfordreference.com/view/10.1093/acref/9780199653065.001.0001/acref-9780199653065

Bar-Hillel, M., & Wagenaar, W. A. (1991). The Perception of Randomness. *Advances in Applied Mathematics, 12*(4), 428–454. https://doi.org/10.1016/0196-8858(91)90029-i

Batanero, C. (2015). Understanding randomness. Challenges for research and teaching. In K. Krainer & N. Vondrová (Eds.), *Proceedings of the Ninth Congress of European Research in Mathematics Education* (pp. 34–49). https://www.researchgate.net/publication/273213387_Understanding_randomness_Challenges_for_research_and_teaching

Batanero, C., Chernoff, E. J., Engel, J., Lee, H. S., & Sánchez, E. (2016). Research on Teaching and Learning Probability. In *ICME-13 Topical Surveys*. (Research on Teaching and Learning Probability ed., pp. 1–33). Springer, Cham. https://doi.org/10.1007/978-3-319-31625-3_1

Bennett, D. (2011). Defining Randomness. In P. S. Bandyopadhyay & M. R. Forster (Eds.), *Philosophy of Statistics* (First Edition, pp. 633–639). Elsevier B. V. https://doi.org/10.1016/B978-0-444-51862-0.50020-4

Chaitin, G. J. (1975). Randomness and Mathematical Proof. *Scientific American, 232*(5), 47–52. https://doi.org/10.1038/scientificamerican0575-47

Daintith, J. (2008). *A Dictionary of Chemistry*. Oxford Reference. https://www.oxfordreference.com/view/10.1093/acref/9780199204632.001.0001/acref-9780199204632

Daintith, J. (2009). *A Dictionary of Physics*. Oxford Reference. https://www.oxfordreference.com/view/10.1093/acref/9780199233991.001.0001/acref-9780199233991

Gougis, R. D., Stomberg, J. F., O'Hare, A. T., O'Reilly, C. M., Bader, N. E., Meixner, T., & Carey, C. C. (2017). Post-secondary Science Students' Explanations of Randomness and Variation and Implications for Science Learning. *International Journal of Science and Mathematics Education, 15*(6), 1039–1056. https://doi.org/10.1007/s10763-016-9737-7

Kahneman, D. (2011). *Thinking, Fast and Slow*. Penguin Random House.

Kaplan, J. J., Fisher, D. G., & Rogness, N. T. (2009). Lexical Ambiguity in Statistics: What do Students Know about the Words Association, Average, Confidence, Random and Spread? *Journal of Statistics Education, 17*(3), 1–20. https://doi.org/10.1080/10691898.2009.11889535

Martin, E. and Hine, R. (2008). *A Dictionary of Biology*. Oxford Reference. https://www.oxfordreference.com/view/10.1093/acref/9780199204625.001.0001/acref-9780199204625

Mcleod, S. (2019). Constructivism as a Theory for Teaching and Learning. *Simply Psychology*. https://www.simplypsychology.org/constructivism.html

New Zealand Ministry of Education. (n.d.). *NZ Maths*. Te Kete Ipurangi. https://nzmaths.co.nz

New Zealand Ministry of Education. (2012). Randomness. In *Glossary of Statistics Terms for the New Zealand Mathematics and Statistics Curriculum*. https://seniorsecondary.tki.org.nz/Mathematics-and-statistics/Glossary

Nickerson, R. S. (2002). The Production and Perception of Randomness. *Psychological Review, 109*(2), 330–357. https://doi.org/10.1037/0033-295x.109.2.330

Pfannkuch, M., & Brown, C. M. (1996). Building on and Challenging Students' Intuitions About Probability: Can We Improve Undergraduate Learning? *Journal of Statistics Education, 4*(1), 1–15. https://doi.org/10.1080/10691898.1996.11910502

Renelle, A., Budgett, S., & Jones, R. (2020). New Zealand Teachers' Generation Problem Misconceptions. *Teaching Statistics*, 1–6. https://doi.org/10.1111/test.12248

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science, 185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

LEARNING MLE CONCEPTS BY ESTIMATING N IN BINOMIAL DISTRIBUTION

SHANNON Ian
NSW Department of Planning, Industry and Environment
Ian.shannon@environment.nsw.govt.au

Maximum likelihood estimation (MLE) is one of the fundamental concepts taught early in theory of statistics courses. As a foundation constituent of statistical theory, it is important for these beginning students to appreciate the method. Additionally, although not necessarily covered in servicing courses, an appreciation of the MLE method will likely assist students from other facilities to understand statistical methods. This paper uses a practicable environmental issue to illustrate the MLE concept, which could then be followed by a more formal treatment using calculus for mathematically inclined students. Wind turbines often kill endangered bats and birds that fly into the rotating blades. Appropriate remedial action is predicated on knowing the number killed which is estimated using carcass counts beneath the wind turbine. However, many carcasses are scavenged or decay before search parties can count them with probability of discovering the carcass often 10% or lower. To introduce the method of MLE to students this example shows how to estimate the number killed using sample evidence of carcass counts. Other concepts demonstrated include making statistical estimates from a sample of one, non-uniqueness of MLE showing cases of two parameter values which maximise the likelihood function and that MLEs can be biased.

INTRODUCTION

Learning the concept of estimators and maximum likelihood estimators (MLE) may present difficulties for some student who may well be satisfied using the methods of moments. In a beginning statistics course, where students' understanding of fundamental concepts is a necessary goal, appreciation of MLE is just one necessary step in attaining that objective.

An estimator for **n** where counts (**r)** of positives but not of negatives are available is used here to illustrate the MLE concept. Specifically, the duality of pdf and likelihood function and the maximising concept are shown using an explicit set of examples.

BACKGROUND OF EXAMPLE

The Department of Planning, Industry and Environment (DPIE) negotiates with wind farms to firstly monitor and secondly provide a response to bat and bird strikes by wind turbines. If an animal flies into the blades or close enough to be affected by the air pressure gradient of rotating turbine blades the result is fatal. DPIE's concern for flying threatened or endangered species killed by wind turbines means that wind farms have monitoring conditions imposed. As a minimum these conditions require the counting of carcasses beneath and surrounding the base of turbines to be able to estimate a monthly and annual fatality figure for each threatened species.

In NSW the introduced fox is pervasive and with feral cats, wild canids and other carnivores scavenging carcasses below wind turbines often occurs. It has been suggested that the length of time a carcass remains identifiable is between 1 and 8 days depending on carcass size and wind farm location. The fall zone (and hence search area) of animal carcasses can exceed a hectare. The search efficiency is variously quoted at about 0.5-0.7 for humans and 0.8-0.9 for trained dogs. For searches conducted only one or two times a month, the probability of detecting a carcass given a turbine kill is typically a small fraction often less than 0.2. The issue addressed here is to estimate the number of bat and bird deaths given the carcass evidence.

DEVELOPMENT OF MLE

With a total number of deaths n and a probability of carcass detection given a strike, the probability of each of the possible carcass counts (0 to n) is given by the binomial pdf.

$$\Pr(r \mid n, p) = \binom{n}{r} p^r (1-p)^{n-r} \qquad 0 \le \boldsymbol{r} \le \boldsymbol{n}$$
$$= 0 \qquad \text{elsewhere}$$

where

n is total number of strikes $n \in \mathbf{Z}$, $n \geq 0$
r is number of carcasses found,
p is detection probability $0 \leq p \leq 1$,
$\binom{n}{r}$ gives the binomial combinations $\binom{n}{r} = \frac{n!}{r! \times (n-r)!}$

The pdf which is a function of r given p and n, can be considered as a function of p given r and n and relabelled as L(p | r, n). Then by varying p across its range the value of p which delivers a maximum for $L(p \mid r, n)$ is the MLE for p. This is covered in many texts by taking logs of L (to keep the algebra simple), differentiating L w.r.t p and solving the equation for p after setting the derivative to 0.

In the wind turbine case, a value of p is known in advance and n is the unknown. Hence, we investigate the likelihood function L(n | r, p). The example of a wind turbine with typical values of 2 searches per month, 30 days in a month, life span of a moderate sized bird carcass at 4 days and detection probability (using trained dogs) of carcasses at 90%.

p = 2(searches/month) *
4(days carcass survival) *
0.9(carcass/bird strike/search) / 30(days/month)
= 0.24(carcasses found/bird killed)

The crux of MLE estimation is based on the probability statement (a function of r given parameter values n and p). This is then reused and labelled likelihood by using the same formula and values but as a function of the parameter of concern (n) given the observed random variable (r). These relationships are shown in Table 1 and Table 2 for an example that uses p at 0.24.

Table 1 gives the probability mass function, $\Pr(r \mid n, p) = \binom{n}{r} p^r (1-p)^{n-r}$ and Table 2 gives the likelihood function $\mathrm{L}(n \mid r, p) = \binom{n}{r} p^r (1-p)^{n-r}$. These tables are also available as a dynamic self-contained web page where the highlight of rows and columns is changed by mouse clicks.

Table 1 Probability of detecting carcasses for various number of birds killed, using p = 0.24. Columns give the number of carcasses detected (r). Each row applies for one level of bird deaths (n). Alternate rows are colour highlighted to emphasise the probability functions. All rows sum to unity as each row is a probability density function.

	<---------- r ~ Number of carcasses detected ---------->														
n	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	1.0000															
1	0.7600	0.2400														
2	0.5776	0.3648	0.0576													
3	0.4390	0.4159	0.1313	0.0138												
4	0.3336	0.4214	0.1996	0.0420	0.0033											
5	0.2536	0.4003	0.2529	0.0798	0.0126	0.0008										
6	0.1927	0.3651	0.2882	0.1214	0.0287	0.0036	0.0002									
7	0.1465	0.3237	0.3067	0.1614	0.0510	0.0097	0.0010	0.0000								
8	0.1113	0.2812	0.3108	0.1963	0.0775	0.0196	0.0031	0.0003	0.0000							
9	0.0846	0.2404	0.3037	0.2238	0.1060	0.0335	0.0070	0.0010	0.0001	0.0000						
10	0.0643	0.2030	0.2885	0.2429	0.1343	0.0509	0.0134	0.0024	0.0003	0.0000	0.0000					
11	0.0489	0.1697	0.2680	0.2539	0.1603	0.0709	0.0224	0.0050	0.0008	0.0001	0.0000	0.0000				
12	0.0371	0.1407	0.2444	0.2573	0.1828	0.0924	0.0340	0.0092	0.0018	0.0003	0.0000	0.0000	0.0000			
13	0.0282	0.1159	0.2195	0.2542	0.2007	0.1141	0.0480	0.0152	0.0036	0.0006	0.0001	0.0000	0.0000	0.0000		
14	0.0214	0.0948	0.1946	0.2459	0.2135	0.1348	0.0639	0.0231	0.0064	0.0013	0.0002	0.0000	0.0000	0.0000	0.0000	
15	0.0163	0.0772	0.1707	0.2336	0.2213	0.1537	0.0809	0.0329	0.0104	0.0025	0.0005	0.0001	0.0000	0.0000	0.0000	0.0000
16	0.0124	0.0626	0.1482	0.2185	0.2242	0.1699	0.0984	0.0444	0.0158	0.0044	0.0010	0.0002	0.0000	0.0000	0.0000	0.0000
17	0.0094	0.0505	0.1277	0.2016	0.2228	0.1830	0.1156	0.0573	0.0226	0.0071	0.0018	0.0004	0.0001	0.0000	0.0000	0.0000
18	0.0072	0.0407	0.1092	0.1839	0.2177	0.1925	0.1317	0.0713	0.0310	0.0109	0.0031	0.0007	0.0001	0.0000	0.0000	0.0000
19	0.0054	0.0326	0.0927	0.1659	0.2096	0.1986	0.1463	0.0858	0.0406	0.0157	0.0050	0.0013	0.0003	0.0000	0.0000	0.0000
20	0.0041	0.0261	0.0783	0.1484	0.1991	0.2012	0.1589	0.1003	0.0515	0.0217	0.0075	0.0022	0.0005	0.0001	0.0000	0.0000
21	0.0031	0.0208	0.0658	0.1316	0.1870	0.2007	0.1690	0.1144	0.0632	0.0288	0.0109	0.0035	0.0009	0.0002	0.0000	0.0000
22	0.0024	0.0166	0.0550	0.1158	0.1737	0.1974	0.1766	0.1275	0.0755	0.0371	0.0152	0.0052	0.0015	0.0004	0.0001	0.0000
23	0.0018	0.0132	0.0458	0.1012	0.1598	0.1917	0.1816	0.1393	0.0880	0.0463	0.0205	0.0076	0.0024	0.0006	0.0001	0.0000
24	0.0014	0.0105	0.0380	0.0879	0.1457	0.1841	0.1841	0.1495	0.1003	0.0563	0.0267	0.0107	0.0037	0.0011	0.0003	0.0001
25	0.0010	0.0083	0.0314	0.0759	0.1318	0.1749	0.1841	0.1578	0.1121	0.0669	0.0338	0.0145	0.0054	0.0017	0.0005	0.0001
26	0.0008	0.0065	0.0258	0.0652	0.1184	0.1645	0.1818	0.1641	0.1231	0.0777	0.0417	0.0192	0.0076	0.0026	0.0008	0.0002
27	0.0006	0.0052	0.0212	0.0558	0.1056	0.1535	0.1777	0.1683	0.1329	0.0886	0.0504	0.0246	0.0103	0.0038	0.0012	0.0003

Table 2 Likelihood of bird deaths, given the carcass detection, using p = 0.24. Columns give the likelihood for an observed carcass detection (r). Each row gives the likelihood for the level of deaths (n) given at the left-hand side. Alternate columns are colour highlighted to emphasise the likelihood functions. For columns corresponding to r in [0, 6] the maximum value is shown in red. . For columns [7, 15] the maximum occurs below row for n=27 and can't be shown.

	<---------- r ~ Number of carcasses detected ---------->														
n	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	1.0000															
1	0.7600	0.2400														
2	0.5776	0.3648	0.0576													
3	0.4390	0.4159	0.1313	0.0138												
4	0.3336	0.4214	0.1996	0.0420	0.0033											
5	0.2536	0.4003	0.2529	0.0798	0.0126	0.0008										
6	0.1927	0.3651	0.2882	0.1214	0.0287	0.0036	0.0002									
7	0.1465	0.3237	0.3067	0.1614	0.0510	0.0097	0.0010	0.0000								
8	0.1113	0.2812	0.3108	0.1963	0.0775	0.0196	0.0031	0.0003	0.0000							
9	0.0846	0.2404	0.3037	0.2238	0.1060	0.0335	0.0070	0.0010	0.0001	0.0000						
10	0.0643	0.2030	0.2885	0.2429	0.1343	0.0509	0.0134	0.0024	0.0003	0.0000	0.0000					
11	0.0489	0.1697	0.2680	0.2539	0.1603	0.0709	0.0224	0.0050	0.0008	0.0001	0.0000	0.0000				
12	0.0371	0.1407	0.2444	0.2573	0.1828	0.0924	0.0340	0.0092	0.0018	0.0003	0.0000	0.0000	0.0000			
13	0.0282	0.1159	0.2195	0.2542	0.2007	0.1141	0.0480	0.0152	0.0036	0.0006	0.0001	0.0000	0.0000	0.0000		
14	0.0214	0.0948	0.1946	0.2459	0.2135	0.1348	0.0639	0.0231	0.0064	0.0013	0.0002	0.0000	0.0000	0.0000	0.0000	
15	0.0163	0.0772	0.1707	0.2336	0.2213	0.1537	0.0809	0.0329	0.0104	0.0025	0.0005	0.0001	0.0000	0.0000	0.0000	0.0000
16	0.0124	0.0626	0.1482	0.2185	0.2242	0.1699	0.0984	0.0444	0.0158	0.0044	0.0010	0.0002	0.0000	0.0000	0.0000	0.0000
17	0.0094	0.0505	0.1277	0.2016	0.2228	0.1830	0.1156	0.0573	0.0226	0.0071	0.0018	0.0004	0.0001	0.0000	0.0000	0.0000
18	0.0072	0.0407	0.1092	0.1839	0.2177	0.1925	0.1317	0.0713	0.0310	0.0109	0.0031	0.0007	0.0001	0.0000	0.0000	0.0000
19	0.0054	0.0326	0.0927	0.1659	0.2096	0.1986	0.1463	0.0858	0.0406	0.0157	0.0050	0.0013	0.0003	0.0000	0.0000	0.0000
20	0.0041	0.0261	0.0783	0.1484	0.1991	0.2012	0.1589	0.1003	0.0515	0.0217	0.0075	0.0022	0.0005	0.0001	0.0000	0.0000
21	0.0031	0.0208	0.0658	0.1316	0.1870	0.2007	0.1690	0.1144	0.0632	0.0288	0.0109	0.0035	0.0009	0.0002	0.0000	0.0000
22	0.0024	0.0166	0.0550	0.1158	0.1737	0.1974	0.1766	0.1275	0.0755	0.0371	0.0152	0.0052	0.0015	0.0004	0.0001	0.0000
23	0.0018	0.0132	0.0458	0.1012	0.1598	0.1917	0.1816	0.1393	0.0880	0.0463	0.0205	0.0076	0.0024	0.0006	0.0001	0.0000
24	0.0014	0.0105	0.0380	0.0879	0.1457	0.1841	0.1841	0.1495	0.1003	0.0563	0.0267	0.0107	0.0037	0.0011	0.0003	0.0001
25	0.0010	0.0083	0.0314	0.0759	0.1318	0.1749	0.1841	0.1578	0.1121	0.0669	0.0338	0.0145	0.0054	0.0017	0.0005	0.0001
26	0.0008	0.0065	0.0258	0.0652	0.1184	0.1645	0.1818	0.1641	0.1231	0.0777	0.0417	0.0192	0.0076	0.0026	0.0008	0.0002
27	0.0006	0.0052	0.0212	0.0558	0.1056	0.1535	0.1777	0.1683	0.1329	0.0886	0.0504	0.0246	0.0103	0.0038	0.0012	0.0003

OBSERVATIONS ON TABLES

- The columns are truncated at 15 as for all the rows shown the associated probabilities in columns to the right of 15 are either zero or less than 0.00005 and hence would display as 0.0000.
- Both tables have identical values in corresponding positions.
- Tables are blank outside the domain $0 \leq \boldsymbol{r} \leq n$ however the definition above specifically gives a zero value for this otherwise undefined region.
- Using table 2 the maximum of each likelihood function (r ϵ [0, 7]) is easily read off.
- To assist with learning about MLE the use of table 1 and table 2 showing the duality between a probability statement and associated likelihood function can well prompt the student with that *light bulb* moment.
- The tables assist with understanding of the maximising and what it means.
- Each column maximum occurs where n is the integer value within the interval $\left[\frac{r}{p} - 1,\ \frac{r}{p}\right]$.
- It is possible to have more than one value of n as the maximum. This occurs where the value of $\frac{r}{p}$ is an exact integer and both the likelihood values for $\left(\frac{r}{p} - 1\right)$ and $\left(\frac{r}{p}\right)$ are equal.
- If there were to be multiple observations $r_1, r_2, \ldots r_k$ then the columns associated with these r values are extracted and multiplied together elementwise. The resultant column is then scanned for its maximum and the associated value of n becomes the MLE estimate based on the current data vector.
- $\hat{n}$ is biased for n, simply seen in the case where *r*=0, that is no carcasses were found. In that case $\hat{n}$ is zero, however other positive values of *n* are possible meaning the average value of $\hat{n}$ must be greater than zero.

FITTING THIS EXAMPLE INTO A LEARNING STRATEGY

Obviously, an understanding of probability density functions is a prerequisite but the pdf calculations in the example will also engage students with those concepts. Likewise, students need to be familiar with various statistical distributions and specifically the binomial where again the calculations will require students to use this prior knowledge.

With statistics servicing courses, the appreciation that there are other parameter estimation methods apart from the methods of moments will probably be sufficient. Engagement by showing the duality concept using the table with rows emphasised for probabilities and columns for likelihood could better motivate them more than straight formulas.

For mainstream statistics students, where this concept needs to be firmly understood, making the tables or alternatively the web page (where the highlight of rows and columns is changed by mouse clicks) available to students can assist the retaining both important concepts of probability-likelihood duality and likelihood maximisation.

CONCLUSION

The description of an environmental problem and the use of the discrete binomial probability mass functions is conducive to learning and understanding the concept of maximum likelihood.

The use of maximum likelihood estimation as describe here has proven useful to staff in DPIE when assessing wind farm development and operational plans to assist with actions to mitigate negative impact on threatened species.

I would especially thank Dr Samantha Travers and Mallory Barnes for introducing me to the issue requiring better than equating the number of carcasses to kills.

This is another case (see Wikipedia description of German tank problem) where meaningful statistical analysis is possible using a sample size of one.

Calculations were performed using the array programming language J which proved useful for defining the whole table shown above directly from the probability statement without loops.

ADDITIONAL MATERIAL

The web page (an html file) is available from the author at the NSW Department of Planning, Industry and Environment. Ian.Shannon@environment.nsw.gov.au.

REFERENCES

Wikipedia, German Tank problem. https://en.wikipedia.org/wiki/German_tank_problem Retrieved 26 May 2021

JSoftware Inc, Version 903 of J language, http://www.jsoftware.com/download/j903/install/j903_win64.zip Accessed 16 Feb 2021

HIGHLY ENGAGING BAYESIAN DEMONSTRATIONS

Damjan Vukcevic[1,2]
[1]School of Mathematics and Statistics
[2]Melbourne Integrative Genomics
University of Melbourne
damjan.vukcevic@unimelb.edu.au

Introductory statistics teaching typically focuses on the classical/frequentist approach. Bayesian inference usually only gets a short mention, at best. In my experience, students often misunderstand the motivations and differences between the two approaches, especially the distinction between modelling uncertainty (Bayesian/epistemic probability) rather than variation in data (frequency/aleatory probability).

In the second-year undergraduate introductory statistics subject that I teach, we devote a single week (out of 12) to Bayesian inference. To make the most of this, I developed three interactive in-class demonstrations that showcase the key concepts in an engaging and accessible way. These include a 'card trick', a live experiment and interactive use of R. The card trick is a highlight, starting off as a bit of fun but quickly making the students re-think their knowledge of probability. They come to the realisation that they already understand and intuitively use probability in a Bayesian sense, despite having spent the first 9 weeks strictly using frequency probability.

I will describe my demonstrations and explain the many learning points that I take advantage of via strategically prepared commentary. Given the ubiquity of Bayesian techniques in modern-day applications, I hope to leave my students with the best foundation for their developing statistical careers.

In: A. A. Bilgin and S. Budgett (Eds.). Proceedings of the Tenth Australian Conference on Teaching Statistics, July 2021 [Virtual]

LEARNING ANALYTICS TO PREDICT FINAL GRADE PARTWAY THROUGH INTRODUCTORY STATISTICS

MOLNAR, Adam
Oklahoma State University, Stillwater, USA
adam_molnar@yahoo.com

As part of a study about student performance in introductory general-education statistics at a large US university, learning analytics models were developed to prospectively estimate final grades after week 6 of a 16 week term. Models were constructed to predict percentage marks and the dichotomous result of above-average or below-average grade. Available variables for 199 students included grades from four homework assignments and one hour exam, standardized maths test scores and highest maths class completed, college and secondary school grade averages, and demographics including year in university, gender, race, age, university transfer, and first generation college status. Modelling methods included linear regression, logistic regression, LASSO regression, decision trees, and nearest neighbours. Reduced models such as stepwise regression and decision trees had more accurate cross-validation estimates. First exam score, college grade point average, and score on probability homework were part of all best- performing models. Ways to apply model results with students and instructors are discussed.

INTRODUCTION

Many universities in the USA (and some elsewhere) require that instructors provide intermediate grades to students during a course. For example, my current university requires that students receive a grade representing their progress after week 6 of a 16 week term. Some institutions have provided midterm marks since the 1980s. Although provided grades have shown mixed evidence of benefit (Alley, 2002), student advising offices believe strongly in them. The Utah Valley University Retention group (n.d.) offers a representative statement, that "a midterm grade provides [students] with the opportunity to make adjustments while there is still time to achieving a passing grade." Providing assignment scores to students is not enough, though; academic advisors and retention specialists should also have access.

Instructors tend to assign intermediate grades based on completed work, as universities do not generally mandate a grading approach. More data is potentially available, however, and if the goal of intermediate information is to accurately help students judge progress towards a final outcome, a more precise prediction would be more helpful. This paper contains the result of an attempt to find a better grade prediction model than applying earned percentage so far, based on variables available at week 6. The project is a practical example of applying predictive analytics in higher education, often called learning analytics, "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs" (Long & Siemens, 2011, p. 33).

Predicting final grades has been a process for about as long as there have been final grades, with academic papers such as Watson (1988) modeling success in mathematical sciences. Increased computing power and Internet-based courses have made the process easier, with authors such as Wang and Newlin (2000) demonstrating benefits from including online activity measures. Learning analytics has developed into a field with international conferences starting in 2011, and continues to grow as part of data science.

Initial efforts retrospectively analyzed individual courses, but forward-looking analysis would be more useful. An early prospective system was Course Signals at Purdue University. Factors in the Course Signals model included course grades, relative participation in the online course learning management system (LMS), prior academic history, and demographic characteristics. Based on these models and faculty settings, students saw a red, yellow, or green light for each course (Arnold & Pistilli, 2012). These systems proved popular and multiple LMS developers offer automated systems. Course Signals became part of Ellucian's CRM Advise, with other competitors including Retention Coaching by Blackboard (2020). Corporate LMS products cost money, though, and an open source initiative has developed to attempt to reduce costs and providing more transparency (Jayaprakash et al., 2014). Papers

and books have examined benefits, including how instructors perceive these systems (Atif et al., 2020, Ifenthaler et al., 2019).

Model building techniques from statistics, machine learning, and data science have been used in learning analytics. Popular techniques have included decision trees, naïve Bayes classifiers, support vector machines, neural networks, random forests, and logistic regression (Chen & Cui, 2020; Cui et al., 2019; Marbouti et al. 2016). Learning analytics models have used many different types of variables. In a review article, Cui et al. (2019) identified seven categories of potential predictors.

- Performance in the current course.
- LMS activity such as time spent, clicks, and discussion posts.
- Previous academic history such as college GPA and placement test scores.
- Student demographics such as gender, age, and race.
- Student socio-emotional variables such as attitudes.
- Features of the course such as modality and enrollment.
- Instructor variables such as teaching quality and style.

DATA COLLECTION

During northern hemisphere spring (January to May) 2019 and fall (August to December) 2019 semesters, I conducted a study to examine mathematical and other factors related to success in an introductory statistics course. The non-calculus-based introductory statistics course is the largest course in mathematics or statistics at this public university in the south-central part of the USA, with about 1500 enrolled students each calendar year. Multiple instructors offer sections with 50 to 100 students per section, with about 75% of sections in-person and 25% online. Students complete common graded homework assignments and take exams designed to be roughly equal, although not the same exams because sections take exams at varying times. Course staff discuss exam grades; exam marks for sections can be adjusted upward to ensure fairness.

In the spring, four instructors asked students to participate in the study. Students could receive a small amount of extra credit, about 0.5% of the course grade, for participating and completing a mathematics assessment during the first two weeks of the course. Perhaps because of the extra effort required, only 52% of eligible students completed the assignment. To increase fall participation rates, 6 in-person instructors (2 returning, 4 new) decided to gather consent and administer the survey during class time. Even though extra credit was not awarded in the fall, the in-class administration increased participation rate to 88%. 213 students participated in the spring and 400 participated in the fall.

Participating students could choose if their data would be available for external publication or internal use only. Furthermore, not all instructors provided all homework and exam scores necessary for this prospective analysis. The sample for this paper consisted of 199 students who gave external consent: 26 from one section taught by instructor A in the fall, 102 from four sections taught by instructor B across both semesters, and 71 from three sections taught by instructor C across both semesters. One section from instructor C was online; the remainder were in-person.

VARIABLES

Outcome variables of interest were percentage grade in the course and letter grade awarded. Because 12 of the 199 students withdrew from the course after week 6, these students have a letter grade of W and no percentage grade. Additionally, a few students stopped participating and had very low final percentages. To reduce potential influence, grade percentages below 40% (including withdrawn students) were raised to 40%. This made the range 40 to 107, with mean of 84.07 and standard deviation 17.25.

For ordinal letter grade, I decided to collapse letter grades into two categories, strong grades of A and B, and weaker grades of C, D, F, and W. This course uses a common US system where a percentage of 90 is needed for an A, 80 for a B, 70 for a C, and 60 for a D. A and B letter grades make up a small majority of overall course grades, so this categorization into good and not-so-good grades provides two groups of roughly equal size. Students who gave external consent tended to have higher grades, so the sample contains 144 AB and 55 CDFW.

Potential predictors were taken primarily from the first, third, and fourth categories described by Cui et al. (2019) – course performance, academic history, and demographics. Thinking about the other categories, LMS activity was not available. No socio-emotional surveys were taken to keep student time commitment low. Section enrollment counts were similar and thus not included, but online modality was noted. Instructor was included as a categorical variable, but no teaching quality or style variables were available at the individual instructor level.

Course performance variables were grades from the first four homework assignments and the first exam, all set on a 0 – 1 scale. The class has a total of 350 available points. The first four homework assignments are each worth 10 points, while the first exam is worth 50. Thus, at this point, slightly over one-quarter of course points (90/350) have been earned. Course performance variables were expected to be predictive, given a partial direct relationship to the outcomes.

Academic history variables included the following:

- Score on 19 question algebra and arithmetic exam created to test skills used in the class (Molnar & McDonald, 2019).
- Mathematics score from the ACT or SAT, nationwide college entrance exams. SAT scores were converted to the ACT scale of 1 to 36; students who took both got the higher score.
- If the student took the university's math placement exam, 60% had done so. Doing so generally indicates that the student took a prior math course at this university, because students tend to avoid placement exams unless necessary.
- Highest college-level math course successfully completed, in four categories as none (21% of students), college algebra (48%), precalculus (16%), calculus or higher (15%).
- If the student had a previous attempt in this course. Only 11 students did, 5.5% of the data.
- If student enrolled in a corequisite section with extra math support. 25 students (13%) did.
- Total college credit hours completed. In the US college system, graduation requires about 120 credit hours, with a full semester load around 15 hours.
- College credit hours attempted in that term.
- College Grade Point Average (GPA) on a 0 – 4 scale, with A worth 4, B 3, C 2, D 1, F 0. College GPA had minimum 1.645, first quartile 2.952, median 3.429, third quartile 3.815, and maximum 4.0, with a mean of 3.346. The mean GPA for the course under study is generally around 2.8, making the course more difficult than average.
- High school GPA on a 0 – 4 scale. About 15% of students had a 4.0 GPA and very few were less than 3.0.

Demographic variables included the following:

- Tuition status to indicate residency, either in state (72%) or out of state (28%).
- Self-reported gender. 129 of the 199 students (65%) were female. Fewer than 5 students selected a non-male, non-female gender; due to small size, they were combined with males.
- Age in years. Almost all students had age between 18 and 24; only 4% of students were age 25 or older.
- Self-reported race, as categorized by the University office. Due to multiple small racial group sizes, race was dichotomized into White (67%) and Nonwhite (33%).
- If the student was a transfer student from another university. 18% had transferred.
- If the student self-reported first generation college student status. 18% reported first generation status.

Overall, there were 5 course performance variables, 10 academic history variables, 6 demographic variables, online modality, and instructor, a total of 23 potential predictors.

RESULTS

Models were constructed in R (R Core Team, 2020). Variables were added in three groups. The simplest models used only course performance variables, four homework scores and first exam. Next, the 10 academic history variables, modality, and instructor were added. These models are labeled as "Course + Academic" in the tables below. Finally, the six demographic variables were added and all 23 potential predictors were used.

MODELS PREDICTING A OR B GRADE

To model the categorical AB-vs-CDFW outcome variable, I tried logistic regression with three variable selection methods – all predictors, stepwise regression based on the Bayesian Information Criterion BIC (Schwartz, 1978), and LASSO –, classification trees, and K-nearest neighbors. Misclassification rates in Table 1 were measured using leave-one-out cross validation. Adding demographic variables never improved the misclassification rate. Furthermore, adding demographic variables actually increased misclassification rate in three of the five models, evidence of overfitting.

Table 1. *Misclassification rates for categorical models by data used.*

Method	Course Variables	Course + Academic	All Variables
Logistic regression, all predictors	14.6%	10.1%	13.1%
Logistic regression, stepwise BIC	14.1%	9.5%	9.5%
Logistic LASSO regression	12.6%	9.5%	10.6%
Decision Tree	12.6%	7.5%	7.5%
K-Nearest Neighbors	13.1%	15.6%	19.1%

The best-performing model was a decision tree. Students were classified as receiving a C, D, F, or W if their homework 4 grade was less than 75%, or college GPA was less than 3.1 and exam score was less than 70%, or college GPA was less than 3.1 and exam score was less than 83% and homework 2 grade was less than 94%. The second best model, stepwise logistic regression with BIC, included homework 4 grade, exam grade, and college GPA – three of the four variables in the decision tree.

MODELS FOR PERCENTAGE

To model percentage grade, I tried linear regression with three variable selection methods – all predictors, stepwise with BIC, and LASSO – regression trees, and K-nearest neighbors. Mean absolute error (MAE) was selected as the judgment criteria because of its natural, unambiguous ease of interpretability across models (Willmott & Matsuura, 2005). Table 2 contains leave-one-out cross-validation MAE results for candidate models and variable sets.

Table 2. *Mean absolute error for numeric models by data used.*

Method	Course Variables	Course + Academic	All Variables
Linear regression, all predictors	6.82	5.53	5.74
Linear regression, stepwise BIC	6.97	5.41	5.25
Linear LASSO regression	6.82	5.35	5.26
Decision Tree	6.05	5.56	5.56
K-Nearest Neighbors	6.65	6.90	8.05

As in the categorical case, more complicated models sometimes exhibited overfitting. The best-performing model using stepwise BIC included homework 2 grade, homework 3 grade, homework 4 grade, exam 1 score, math placement test score, college GPA, and transfer status. The second best model from LASSO retained a similar set of coefficients that were not reduced to zero, only replacing homework 2 grade by completion of calculus.

DISCUSSION

Three variables appeared in both sets of best and second-best models – exam score, college GPA, and homework 4 score. Exam score is the largest component of the grade after 6 weeks; not seeing it would be surprising. It is also not surprising to find that overall college GPA is related to performance in this general education introductory course. Regarding homework 4, in the course textbook by Bluman (2018), Chapter 4 covers probability, a more challenging topic than the first three chapters on data collection, visual and table representation, and numeric summaries. As an additional measure of ability, homework 4 makes sense in the models. Interestingly, the other homework scores did not always appear.

Overfitting occurred frequently. In classification, adding demographic variables into logistic regression, LASSO regression, and nearest neighbors increased the validation misclassification rate. In percentage prediction, nearest neighbors and linear regression had a higher mean absolute error with more predictors in the model. Overfitting is an important topic in data science; as Hosseini et al. (2020) and others have written, haphazardly trying a bunch of things without validation frequently leads to dangerous situations. Although I would not cover overfitting in a general introductory course, this dataset is appropriate for a data science course. Having more data – here demographic variables – does not always make prediction better.

With graduate students training to be instructors, these models can also serve as a basis for discussion about factors affecting student performance and how instructors talk about grades with students. Speaking with students about grade prediction involves psychological and ethical considerations. Adults often speak about how a particular teacher encouraged or discouraged their school performance. Self-efficacy, an individual's confidence in that person's ability to successfully accomplish a task, has an established positive effect on statistics course performance (Finney & Schraw, 2003). Would a student make positive adjustments after receiving a poor projection as suggested by Utah Valley University advisors, or reduce effort due to lower self- efficacy? Tone likely matters, and balancing statement of grade facts and projects with encouragement to work well can be very challenging. Having a well-developed model based on the generalized results of hundreds of students could simplify the statement of fact and allow more space for instructor support.

There are also concerns about the models. Authors such as Slade and Prinsloo (2013) have pointed out concerns related to informed consent, data management, and the role of power. All participants gave consent for this study, but does informed consent apply in a system applied to all students? Using data from outside the class might also lead to ethical concerns. Demographic variables such as race, gender, and age did not appear in the best-performing models, but what if they did? Is it appropriate to consider transfer status, which did appear? Learning analytics researchers often presume that information benefits students and instructors. For further research, it might be possible to design an experiment randomized by class to evaluate the effects of projection, similar to Alley's 2002 effort, and applying knowledge learned in these models to test for benefit.

ACKNOWLEDGEMENTS

I would like to thank my research assistant Siyu An who contributed much of the data cleaning and some initial models to this project.

REFERENCES

Alley, V. M. (2002) Midterm grade reports: Are they effective? *Research and Teaching in Developmental Education*, *19*(1), 14–24. https://www.jstor.org/stable/42802148

Arnold, K. E., & Pistilli, M. D. (2012). Course Signals at Purdue: Using learning analytics to increase student success. In S. Buckingham Shum, D. Gašević, & R. Ferguson (Eds.), *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (LAK '12), 29 April–2 May 2012, Vancouver, BC, Canada (pp. 267–270). New York: ACM. https://doi.org/10.1145/2330601.2330666

Atif, A., Richards, D., Liu, D., & Bilgin, A. A. (2020). Perceived benefits and barriers of a prototype early alert system to detect engagement and support 'at-risk' students: The teacher perspective. *Computers & Education, 156*. https://doi.org/10.1016/j.compedu.2020.103954

Blackboard. (2020, July 22). *Blackboard launches proactive and scalable student retention solution* [Press release]. https://www.prnewswire.com/news-releases/blackboard-launches-proactive-and-scalable-student-retention-solution-301097920.html

Bluman, A. (2018). *Elementary statistics: A step by step approach* (10th ed.). McGraw-Hill Education.

Chen, F., & Cui, Y. (2020). Utilizing student time series behaviour in learning management systems for early prediction of course performance. *Journal of Learning Analytics*, 7(2), 1– 17. https://doi.org/10.18608/jla.2020.72.1

Cui, Y., Chen, F., Shiri, A., & Fan, Y. (2019). Predictive analytic models of student success in higher education: A review of methodology. Information and Learning Sciences, 120(3/4), 208–227. https://doi.org/10.1108/ILS-10-2018-0104

Finney, S. J., & Schraw, G. (2003). Self-efficacy beliefs in college statistics courses. *Contemporary Educational Psychology, 28*(2), 161–186. https://doi.org/10.1016/S0361-476X(02)00015-2

Hosseini, M., Powell, M., Collins, J., Callahan-Flintoft, C., Jones, W., Bowman, H., & Wyble, B. (2020). I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data. *Neuroscience & Biobehavioral Reviews, 119*, 456–467. https://doi.org/10.1016/j.neubiorev.2020.09.036

Ifenthatler, D., Mah, D.-K., & Yau, J. (Eds.) *Utilizing learning analytics to support study success.* Springer.

Jayaprakash, S. M., Moody, E. W., Lauría, E. J., Regan, J. R., & Baron, J. D. (2014). Early alert of academically at-risk students: An open source analytics initiative. *Journal of Learning Analytics*, *1*(1), 6-47. https://doi.org/10.18608/jla.2014.11.3

Long, P., & Siemens, G. (2011). Penetrating the fog: Analytics in learning and education. *Educause Review, 46*(5), 30–40.

Marbouti, F., Diefes-Dux, H. A., & Madhavan, K. (2016). Models for early prediction of at-risk students in a course using standards-based grading. *Computers & Education, 103*, 1–15. https://doi.org/10.1016/j.compedu.2016.09.005

Molnar, A., & McDonald, K. (2019, May 16–18). *Math diagnostics and relationship to course grades* [Poster presentation]. US Conference on Teaching Statistics 2019, State College, USA. https://www.causeweb.org/cause/uscots/uscots19/posters/2-13

R Core Team. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*, 461–464. https://doi.org/10.1214/aos/1176344136

Slade, S., & Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. *American Behavioral Scientist, 57*(10), 1510–1529. https://doi.org/10.1177/0002764213479366

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B, 58*(1), 267–288. https://doi.org/10.1111/j.2517- 6161.1996.tb02080.x

Utah Valley University Retention. (n. d.) Midterm grades. https://www.uvu.edu/retention/midtermgrades.html

Wang, A. Y., & Newlin, M. H. (2000). Characteristics of students who enroll and succeed in psychology Web-based classes. *Journal of Educational Psychology, 92*(1), 137–143. https://doi.org/10.1037/0022-0663.92.1.137

Watson, J. M. (1988). Student characteristics and prediction of success in a conventional university mathematics course. *Journal of Experimental Education, 56*(4), 203–212. https://doi.org/10.1080/00220973.1988.10806489

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research, 30*(1), 79–82. https://doi.org/10.3354/cr030079

INTRODUCING BAYESIAN INFERENCE WITH THE TAXICAB PROBLEM

BERG, Arthur and HAWILA, Nour
Pennsylvania State University, USA
berg@psu.edu

The taxi problem goes by many names in the literature including the Schrödinger problem, the German tank problem, the racing car problem, the horse-racing problem, and the taxicab problem. The basic problem goes like this: Suppose taxicabs in a certain city are numbered 1 to N, and one such taxicab is randomly selected, say number 1729. Based on this information, we wish to infer the total number of taxicabs, N, there are in the city. In this paper, we present a non-Bayesian and Bayesian approach to dealing with this problem that uncovers a wealth of statistical inference although we are dealing with a single data value. We provide reasonable assumptions on the potential number of taxicabs that leads to a Bayesian inference that combines the observed data with the additional assumptions into a coherent estimate of N. This paper offers an introduction to Bayesian inference for students as part of an introductory probability and statistics course.

INTRODUCTION

Here we provide a resource for introducing Bayesian statistics to tertiary students as part of an introductory probability and statistics course. The simply stated taxicab problem has a rich history with many well-known statisticians, including M. S. Bartlett, R. Fisher, R. C. Geary, H. Jeffreys, P.-S. Laplace, J. Neyman, C.S. Pierce, E. J. G. Pitman, and E. Schrödinger having been associated with the problem in various contexts. Before we delve into discussing a Bayesian approach to the taxicab problem, we first present some of the rich historical context of this problem including various applications in which the problem arises.

After introducing the different applications, we outline the key approaches to solving the problem. At each step, we explicitly state the assumptions that are being made and students should be challenged to critically evaluate such assumptions in different contexts of the problem. Our first step is to write down the likelihood corresponding to the stated problem. With the likelihood in place, we explore classical non-Bayesian solutions (also referred to as "frequentist solutions") to the problem. We start with the maximum-likelihood estimator and show that this standard estimation method is very conservative and underestimates the true value. We will also draw on a stick breaking model to intuitively calculate the expected values used to formulate an approximately unbiased estimator. Technical details are avoided; rather, we focus on key concepts. The article published in *Teaching Statistics* by Johnson (Johnson, 1994), is an excellent supplemental resource for this discussion.

What is lacking from the non-Bayesian solutions are intuitive or reasonable prior assumptions based on the context of the problem. For example, we can easily justify upper bounds for the number of taxicabs, and this additional information can be easily integrated into the analysis when a Bayesian approach is followed. Specifically, this information enters through the so-called prior in Bayesian inference. Different priors will be considered for different contexts of the taxicab problem.

Once the likelihood and prior are specified, we employ Bayes' theorem to update the prior based on observed data (the taxicab number that was randomly sampled) to produce the posterior distribution. The posterior distribution provides a probability for each possible value of the true parameter. We then discuss how we might reduce the distribution of values down to a single point estimate with a corresponding credible interval (the Bayesian version of the confidence interval).

A PROBLEM WITH A RICH HISTORY

In a letter written by the prolific American statistician Charles Pierce in 1911, Pierce attributes the problem to Pierre-Simon Laplace (Pierce, 1976):

"One of [Laplace's] problems professes to calculate from the fact that all balls in an urn are numbered 1, 2, 3, etc. and the fact that a ball has been drawn and found to bear a number N, what the probable number of balls in the urn is. But no deductive conclusion on the subject can be drawn from those premisses correctly."

Although Pierce clearly attributes Laplace, the authors carefully explored Laplace's extensive writings and various English translations with no success finding any description of this problem. Laplace indeed analysed numerous ball-and-urn problems, but we simply could not find a description of the problem at hand. A couple of decades after Pierce's letter, British statistician H. Jeffreys writes a letter in 1934 to another statistician R. A. Fisher attributing the problem to the Polish statistician J. Neyman (Fisher, 1990):

"[Neyman] once asked me the following: a man arrives at a railway junction in a town in a foreign country, which he has never heard of before. The first thing he sees is a tramcar numbered 100. Can he infer anything about the number of tramcars in the town? [Neyman] thought the question was significant and so did I, and we both had a feeling that there were probably about 200. I tried it on M.S. Bartlett, who thought it was meaningless but had the same feeling about 200."

Then in a 1944 paper, Irish statistician R. Geary attributes the problem to Nobel laureate E. Schrödinger (Geary, 1944):

"At a recent meeting of the Dublin University Mathematical Society, E. Schrödinger suggested the following ingenious problem as an illustration of Pitman's concept of closeness. In a town, cars are known to be numbered consecutively from 1. The numbers on r of the cars are noted: the problem is to find the closest estimate of the number of cars in the town."

Other variants of this problem have also appeared more recently in the literature:

(Tenenbein, 1971): "A spectator at a race track is observing a car race in which the cars are numbered consecutively from one to some unknown number N. He wishes to estimate the number of cars on the race track after observing that M cars numbered $X!$, $X"$, $\cdots$, $X\#$ have passed. Each car is equally likely to hold a given position in the race at any given time."

(Rosenberg & Deely, 1976): "Suppose we are at a horse race where we know there are no scratchings (i.e., the number of horses on the track is equal to the highest number on any horse). We take a moving picture of a particular section of the track and stop the film after M horses have passed by. Assuming that it is possible to read the numbers of the horses in the movie, we wish to estimate the number of horses taking part in the race."

Arguably the most significant application of this problem was during the Second World War, in which the serial numbers of captured German tanks were found to be marked sequentially from 1 to N (Ruggles & Brodie, 1947). Applying the same statistical inference that we discuss in this paper led to an estimate of 246 German tanks being produced each month during the war, which is substantially lower than the conventional Allied intelligence estimates indicating a monthly production of 1,400 tanks. After the war, German records validated the statistical analyses by confirming the actual monthly production number to be 245.

This shows that this simple problem can have many different and diverse applications. In the subsequent discussion, we will stick with the formulation introduced in the abstract: taxicab number 1729 is randomly selected among taxicabs numbered 1 to N, and we wish to estimate the value of N. The specific number 1729 is chosen due to its historical significance as the Ramanujan-Hardy taxicab number (Silverman, 1993). The first task in the inference – Bayesian and non-Bayesian alike – is to write down the likelihood corresponding to the data generating mechanism.

THE LIKELIHOOD

The likelihood simply encodes the probability of observing the data, which we will call M, given the true parameter N. Having observed just one taxicab, the likelihood is simply

$$\Pr(M|N) = \begin{cases} 1/N & if\ M \le N \\ 0 & otherwise. \end{cases}$$

If instead of just one taxicab, we observe k taxicab numbers $M_1, \ldots, M_k$ independently sampled from the N total taxicabs (with replacement), then independence of the data allows us to write the likelihood as follows

$$\Pr(M_1, \ldots, M_k|N) = \prod_{i=1}^{k} \Pr(M_i|N) = \begin{cases} 1/N^k & if\ max\{M_1, \ldots, M_k\} \le N \\ 0 & otherwise. \end{cases}$$

In particular, we see that this likelihood only depends on the maximum observed taxicab number, which we write as $\max\{M_1, \ldots, M_k\}$, thus making $\max\{M_1, \ldots, M_k\}$ a sufficient statistic.

Note that this likelihood assumes every value between 1 and N is possible and equally likely and that multiple samples are drawn independently with replacement. If the sampling was taken without replacement, say we observe three taxicabs at the same time, then we could modify the likelihood accordingly; see e.g. Berg (2021).

NON-BAYESIAN (FREQUENTIST) SOLUTIONS

The maximum-likelihood estimator is the value of M, or more generally $\max\{M_1, \ldots, M_k\}$, that maximizes the likelihood probability. Clearly, $1/N^k$ is a decreasing function in N, and, since the smallest possible value for N is $\max\{M_1, \ldots, M_k\}$, the maximum-likelihood estimator is $\max\{M_1, \ldots, M_k\}$. However, this estimator is rather unsatisfying as it only reports the lower bound of the possible values. An alternative approach is to estimate N with an unbiased estimator.

In order to construct an unbiased estimator, we first calculate the expected value of $\max\{M_1, \ldots, M_k\}$ and use that expected value to solve for N. Here, we only present a heuristic calculation of the expected value; a more rigorous calculation can be found in Johnson (1994). Let's suppose that instead of sampling k values from the discrete set $\{1, \ldots, N\}$, we randomly sample k values from the continuous interval $[0, N]$. In this case, the expected value of $\max\{M_1, \ldots, M_k\}$ can be intuitively calculated using a stick breaking analogy. If you randomly break a stick of length N at k randomly chosen positions, then the resulting k+1 pieces will have length $\frac{N}{k+1}$ on average (see Figure 1). Using this stick-breaking representation, we can intuitively see the expected value of $\max\{M_1, \ldots, M_k\}$ is

$$E[\max\{M_1, \ldots, M_k\}] = N - \frac{N}{k+1} = N\left(1 - \frac{1}{k+1}\right) = N\left(\frac{k}{k+1}\right) \quad \text{(continuous case)}.$$

Note that the above expression applies when the samples are taken on the continuous interval $[0, N]$. A small correction is applied when sampling from the discrete set $\{1, \ldots, N\}$ and depending on whether sampling is done with or without replacement.

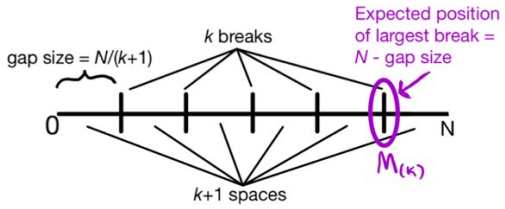


Figure 1. Using the stick breaking analogy to heuristically approximate the expected value of $M_{(k)}$.

Now we construct our approximately unbiased estimate of N by replacing the expectation in the above expression with $\max\{M_1, \ldots, M_k\}$ and solving for N. This leads to the following approximate unbiased estimator of N:

$$\widehat{N} = \max\{M_1, \ldots, M_k\}\left(\frac{k+1}{k}\right)$$

So, for having observed just one taxicab numbered M, this unbiased estimator simply doubles M, which is a much less conservative estimator than the maximum-likelihood estimator. However, there is more information encoded in the problem than just the value of $\max\{M_1, \ldots, M_k\}$. We explore this further in the next section.

THE PRIOR

The problem formulation we are focusing on is estimating the number of taxicabs in a certain city. Although we are not told the size of the city, we can describe some broad bounds that would cover most cities. For example, New York City has about 13,600 licenced taxis, London has about 70,600 licenced taxis, and Mexico City, the city with the most taxicabs, has approximately 140,000 taxicabs. Therefore, we could with reasonable certainty conclude the number of taxis in the unspecified city could range from 50 to 140,000. This is certainly a very wide range, but it is still information that could be utilized to augment the statistical inference. In this example, with such a wide range, the additional information would not provide much improvement, but if we knew more about the city, such as its population, more precise bounds could be constructed leading to more precise estimates.

This prior in Bayesian inference encodes the probabilities of each possible value before observing the data. So, if the feasible range for N is in the interval [50, 140000], then we need to specify a probability for each possibility. Lacking any insight as to the potential number of taxis, a natural prior would be the uniform distribution on [50, 140000] in which each value is equally likely. A somewhat more sophisticated approach would be to utilize published reports, such as Schaller (2005), that contain data on the number of taxis in a large sampling of cities to approximate a more realistic prior distribution.

Mathematically, we will denote the prior distribution as $\pi(n)$ as the probability distribution for the possible values n of N. So, if we take the prior to be a uniform distribution on [50, 140000], then $\pi(n) = 1/(140{,}000 - 50 + 1) = 1/139951$. In this case the prior mean is quite high – close to 70,000 – so instead we modify the prior probabilities to decrease proportionally with n. Specifically, we take $\pi(n) = c/n$, where the constant c is chosen so that the prior sums to one (in this case, $c \approx 0.126$). The prior mean for this "decaying prior" is 17,616, which is still quite high – close to the number of taxis in New York City – but far better than 70,000. It is often the case that different priors are considered to understand how the results vary with the priors.

We finally note that for different applications, such as estimating the number of German tanks or estimating the number horses at a horse race or estimating the number of cars at a racetrack, a different prior would be called for depending on the context of the given application. We would certainly use much smaller numbers when modelling the number of horses or the number of race cars, yet the non-Bayesian estimators presented above do not change according to these substantial differences across the applications.

THE POSTERIOR AND BAYESIAN INFERENCE

Once the prior distribution has been pinned down, the posterior distribution is calculated using Bayes' theorem:

$$\pi(N \mid \max\{M_1, \ldots, M_k\}) = \frac{\pi(N)\, Pr(\max\{M_1, \ldots, M_k\} \mid N)}{\sum_N \pi(N)\, Pr(\max\{M_1, \ldots, M_k\} \mid N)}$$

This seems like a monstrous formula, but it's not so bad; it's basically just the product of the prior $\pi(N)$ with the likelihood $\Pr(\max\{M_1, \ldots, M_k\}|N)$ but then normalized (dividing by its sum) so that it adds up to one. The approach to calculate $\Pr(\max\{M_1, \ldots, M_k\}|N)$ can be found in Berg (2021). Here, we will assume just one taxicab was observed ($k = 1$), so we will use the likelihood $Pr(M|N)$ above. After multiplying $\pi(N)$ by $Pr(M|N)$ and then normalizing the vector so that it sums to one, we obtain the

posterior probability of N given M. This is an entire distribution of values for N, but if we wanted to report a single estimate, we would report a central tendency like the mean or the median of the posterior distribution. Additionally, we can summarize the posterior distribution with a $(1-\alpha)\%$ credible interval by identifying values of N with posterior probabilities that sum to $1-\alpha$ for a user-specified value of α.

SHINY APPLICATION

Implementing the computations required for the Bayesian analysis may be a barrier to some, so we developed an interactive Shiny application (Chang et al., 2021) to assist with these computations. The Shiny app, accessible at https://glow.shinyapps.io/taxicab/, implements the Bayesian and non-Bayesian estimators of the taxicab problem. The source code is accessible at https://github.com/NourHawila/taxicab. After the user provides the data (e.g., observed taxicab numbers), smallest and largest feasible values of N (parameters $N_{\min}$ and $N_{\max}$), and level α, the application calculates and graphs the posterior distribution for N and displays the maximum-likelihood estimate, approximate unbiased estimate, prior mean, posterior mean, posterior median, and the lower and upper bounds of a $(1-\alpha)\%$ credible interval.

We now return to the originally stated taxicab problem having observed the taxicab number 1729. The estimates of the total number of taxicabs based on the maximum-likelihood estimator and the approximate unbiased estimator are 1729 and 3458, respectively. In Table 1 we present the Bayesian for different prior parameters. We see that the Bayesian inference is indeed sensitive to the prior parameters. The more accurate the prior distribution can be specified, the more accurate the Bayesian inference becomes. We also see that the posterior median is consistently smaller than the posterior mean as the posterior distribution is right-skewed.

Table 1. Bayesian analysis of the taxicab problem having observed taxicab number 1729 with four different priors

Parameters			Bayesian results			
Nmin	Nmax	Prior	Prior Mean	Posterior Mean	Posterior Median	95% Credible Interval Upper Bound
50	140,000	Uniform	70,025	31,518	15,561	112,388
50	140,000	Decaying	17,610	15,610	3,417	28,015
50	10,000	Uniform	5,025	4,762	4,159	9,160
50	10,000	Decaying	1,875	4,208	2,949	8,071

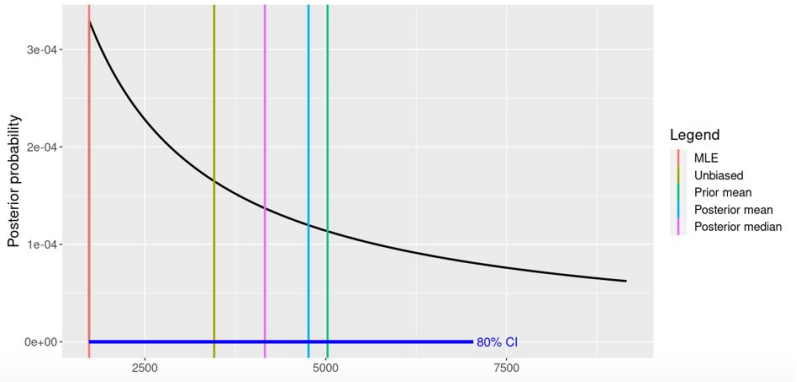


Figure 2. Posterior probabilities of the taxicab problem having observed taxicab number 1729 are illustrated along with five different estimates for N

DISCUSSION

We recognize that Bayesian methods are not as utilized for inference in practice as often as other Frequentist methods and that Bayesian approaches require a proper understanding of conditional probability which could be a difficult task to some students (Moore, 1997). However, Bayesian methods are becoming more popular due to the advances in software power and emphasis on computational thinking in Education. In this paper, we make us of the simplistic structure of the taxicab problem and

its historical roots to provide an excellent gateway problem for students to be introduced to Bayesian methods.

We highlighted several different applications related to this problem, presented non-Bayesian solutions, and detailed a Bayesian approach to solving the problem. The Bayesian solution allows for more information of the problem to be utilized but is also computationally more complex. To facilitate the computation of the Bayesian solution, an accompanying interactive application is described and provided online.

When used in the classroom, additional Bayesian examples and applications that could follow the taxicab problem include Eadie, Huppenkothen, Springford and McCormick (2019), which applies Bayesian statistics to modelling the colours of M&M's candies, Bárcena, Garín, Martín, Tusell et al. (2019), which applies Bayesian statistics in finding the sunken nuclear submarine USS Scorpion that sank in 1968, and Kuindersma & Blais (2007), which uses Bayesian statistics in analysing the probability a flipped cylinder (representing a thick coin) comes to rest on its edge.

BIBLIOGRAPHY

Bárcena, M. J., Garín, M. A., Martín, A., Tusell, F., & Unzueta, A. (2019). A Web Simulator to Assist in the Teaching of Bayes' Theorem. *Journal of Statistics Education*, *27*(2), 68–78. https://doi.org/10.1080/10691898.2019.1608875

Berg, A. (2021). Bayesian Modeling Competitions for the Classroom. *Colombian Journal of Statistics*, *in press*, 1–11.

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2021). *shiny: Web Application Framework for R* (1.6.0). https://cran.r-project.org/web/packages/shiny/

Eadie, G., Huppenkothen, D., Springford, A., & McCormick, T. (2019). Introducing Bayesian Analysis With m&m's®: An Active-Learning Exercise for Undergraduates. *Journal of Statistics Education*, *27*(2), 60–67. https://doi.org/10.1080/10691898.2019.1604106

Fisher, R. A. (1990). *Statistical Inference and Analysis* (J. Bennett (ed.); xviii). Clarendon Press.

Geary, R. C. (1944). Comparison of the Concepts of Efficiency and Closeness for Consistent Estimates of a Parameter. *Biometrika*, *33*(2), 123. https://doi.org/10.2307/2334111

Johnson, R. W. (1994). Estimating the Size of a Population. *Teaching Statistics*, *16*(2), 50–52. https://doi.org/10.1111/j.1467-9639.1994.tb00688.x

Kuindersma, S. R., & Blais, B. S. (2007). Teaching Bayesian Model Comparison with the Three- Sided Coin. *The American Statistician*, *61*(3), 239–244.

Moore, D. S. (1997). Bayes for Beginners? Some Reasons to Hesitate. *American Statistician*, *51*(3), 254–261. https://doi.org/10.1080/00031305.1997.10473972

Pierce, C. S. (1976). The New Elements of Mathematics. In C. Eisele (Ed.), *Mathematical Miscellanea* (Vol. 3, Issue 3). https://doi.org/10.1016/0315-0860(77)90070-2

Rosenberg, W. J., & Deely, J. J. (1976). The Horse-Racing Problem-A Bayesian Approach. *The American Statistician*, *30*(1), 26–29. https://doi.org/10.2307/2682883

Ruggles, R., & Brodie, H. (1947). An Empirical Approach to Economic Intelligence in World War II. *Journal of the American Statistical Association*, *42*(237), 72–91.

Schaller, B. (2005). A Regression Model of the Number of Taxicabs in U.S. Cities. *Journal of Public Transportation*, *8*(5), 63–78. https://doi.org/10.5038/2375-0901.8.5.4

Silverman, J. H. (1993). Taxicabs and Sums of Two Cubes. *The American Mathematical Monthly*, *100*(4), 331–340.

Tenenbein, A. (1971). The Racing Car Problem. *The American Statistician*, *25*(1), 38–40.

A STUDY OF WIL IN STATISTICS AND ANALYTICS: WHAT HAS BEEN ACHIEVED AND WHAT CAN BE IMPROVED?

Yan Wang, Denwick Munjeri and Mali Abdollahian
RMIT University
yan.wang@rmit.edu.au

Within the traditional science area, especially for the Mathematics discipline, the number of project-based and placement-based activities are less compared to other disciplines. The work integrated learning (WIL) course has exposed and placed our students directly to industry through internships and projects offered by industry organisations. This research will study the WIL course of two programs within the mathematics discipline, Master of Analytics and Masters of Statistics and Operations Research at RMIT University. A study was carried out to demonstrate the impact of our good practice of WIL in Statistics and Analytics, which can be shared with mathematics programs at other universities. Meanwhile the study also shows the barriers to WIL in statistics that may be further improved to value the WIL activities in the mathematics discipline.

A JOB-READY ASSESSMENT FOR POST-GRADUATE LEVEL INTRODUCTORY BIOSTATISTICS COURSES: DESIGN AND IMPLEMENTATION

Darsy Darssan, Pakhi Sharma, Alexandra Robbins-Hill and Gail Williams
The University of Queensland, Australia
d.darssan@uq.edu.au

Learners enrolled in master level introductory Biostatistics courses have backgrounds in a variety of life sciences. They include medical doctors, health service researchers, Doctor of Philosophy candidates, and students in Masters of Public Health or epidemiology programs within medical faculties. Their objectives are usually to learn the fundamentals of Biostatistics so they can apply basic quantitative methods to their research projects and also to communicate with biostatisticians in their team. Traditional assessments in biostatistics courses often fail to provide an opportunity for them to evaluate these professional skills. At the University of Queensland, using a flipped-classroom structure suitable for distance mode learning we implemented a job-ready assessment with four components. These were: (i) exposure to the real world by sourcing a unique dataset from workplaces, projects, or open databases, (ii) simultaneous learning and application throughout the course, with loosely defined instructions that mimic real-life challenges in requiring careful selection of appropriate theoretical concepts to use for various research questions, (iii) exchanging feedback with classmates in order to train their cognitive ability to assess the work of others and constructively accept or reject feedback, and (iv) production of a four-minute video presentations for a lay audience to simulate the professional workplace.

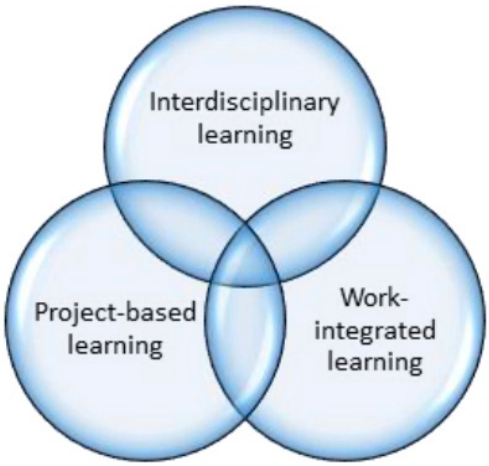
BUILDING EMPLOYABILITY CAPABILITIES IN DATA SCIENCE STUDENTS: AN INTERDISCIPLINARY, INDUSTRY-FOCUSSED APPROACH

Sonia Ferns[1], Aloke Phatek[1], Susan Benson[2], Andrew St John[3] and Nina Kumagai[1]
[1]Curtin University
[2]Lab Tests Online Australasia and Pathwest
[3]Lab Tests Online Australasia
s.ferns@curtin.edu.au

In the contemporary workplace, competent data scientists capable of interdisciplinary collaboration are in high demand. In addition, employers are seeking innovative, agile, and motivated employees who are analytical, communicative, collaborative, resilient and creative. To nurture these attributes, a student data scientist needs to experience a plethora of learning opportunities that involve collaboration in interdisciplinary contexts and engagement with industry partners.

Curtin University and Lab Tests Online Australasia (LTOAU) collaborated to provide an interdisciplinary, industry-focussed learning experience for data science students. Students analysed text complexity of online consumer education content, website Google analytics and developed an online survey. LTOAU has used the insights from this work to inform redesign of processes and content. Students reported improved self-awareness, and teamwork, decision-making and leadership skills from this interdisciplinary learning experience. Tackling a real-world problem as part of an interdisciplinary team empowered them to understand and respond to client requirements, embrace diversity, learn from others, and establish trust and equity within a focussed team.

To navigate the dynamic and unpredictable landscape of the future workplace, graduates require transferrable skills across global contexts. This presentation will highlight how interdisciplinary, industry focussed learning experiences can provide these skills to data science students, thereby enhancing their employability.



STEERING STUDENTS PAST THE 'TRUE MODEL MYTH'

Damjan Vukcevic[1,2], Margarita Moreno-Betancur[3,4], John Carlin[4,5], Sue Finch[1], Ian Gordon[1] and Lyle Gurrin[5]
[1]School of Mathematics and Statistics, University of Melbourne
[2]Melbourne Integrative Genomics, University of Melbourne
[3]Department of Paediatrics, University of Melbourne
[4]Clinical Epidemiology & Biostatistics Unit, Murdoch Children's Research Institute
[5]Melbourne School of Population and Global Health, University of Melbourne
damjan.vukcevic@unimelb.edu.au

Statistical teaching often focuses on models and techniques, with much less time devoted to emphasising the primacy of the real-world questions that these are meant to answer. One consequence is what we call the 'true model myth': the belief that statistics is about finding the 'best' model or technique to apply to the data.

To help students avoid this, we propose explicitly teaching the idea of a 'statistical investigation', which would be rather like a scientific investigation. This starts with well-posed questions, to be investigated via potentially several different analyses (analogous to scientific experiments), with the results drawn together to form the final conclusions.

The examples we provide should reinforce this. When we teach new techniques, we often keep things simple and show only a single analysis. We can remind students of the larger story by replacing or supplementing these with examples where multiple techniques are used. We present several such examples and exercises that would fit naturally within existing curricula.
Although this change might seem to complicate teaching, we expect that it will actually lead to greater student satisfaction since students will develop skills for "scaffolding" problems in the context of real-world questions.

In: A. A. Bilgin and S. Budgett (Eds.). Proceedings of the Tenth Australian Conference on Teaching Statistics, July 2021 [Virtual]

ONLINE Q&A IN STATISTICS – USING THE STACKEXCHANGE NETWORK

O'NEILL, Ben
Research School of Population Health, Australian National University
ben.oneill@anu.edu.au

We discuss the facilities on the StackExchange network, with particular attention to the CrossValidated website for statistical questions and answers. We set out information on the characteristics of this network and the available expertise among top-ranked users. We identify potential benefits for students and associated benefits for their teachers. We also examine some potential pitfalls that teachers should guard against when dealing with students using the StackExchange network in their coursework.

Students undertaking statistics courses in secondary and tertiary education are generally reliant on their course lecturer. They may have access to further assistance from course tutors, and other staff in their department, but this is still a substantial limitation. Among the "social media generation" it is perhaps natural that students have sought out further assistance from online social question and answer (SQA) sites devoted to technical and scholarly topics that they study in their courses. In particular, many university students can be seen using the popular StackExchange network of SQA sites to assist them with their courses. This online platform opens up a world of technical expertise at no cost to the student, and it provides a handy resource for students beyond the expertise of their course lecturers.

CROSSVALIDATED AND THE STACKEXCHANGE NETWORK

StackExchange is a network of social question and answer (SQA) websites allowing users to post questions, answers, and comments, on a range of topics. Pertinent to statistics are CrossValidated (https://stats.stackexchange.com) for probability, statistics, data analysis and visualisation, and machine learning; StackOverflow (https://stackoverflow.com/) for statistical programming; and Mathematics (https://math.stackexchange.com/) for mathematics. At the time of writing, CrossValidated has 7,937 tracked users, and Mathematics has 32,780 tracked users.[i] The network is presently the most popular reference site on the internet, attracting a large number of daily visits.[ii]

Sites on the StackExchange network use individual questions to create a repository of answers that are of enduring value to a broad audience (Anderson *et al* 2012). Sites use a "gamification" method that adds game-design elements to the platform (Deterding *et al* 2011; Robson *et al* 2015). Gamification acts to incentivise contributions of questions and answers, and allows self-moderation of the site by users. Users can "upvote" or "downvote" questions and answers, and they receive "reputation" points when one of their questions or answers is upvoted, and lose a small amount of points when one of their questions or answers is downvoted. The questioner can "accept" a single answer, sending this answer to the top of the page. Users can also gain "badges" for achieving certain requirements on the site. As users gain "reputation" on a site, they are granted privileges, and can undertake moderating tasks. There are also elected moderators, who hold the power to close or delete questions or answers. In Figure 1 we show part of the user profile of a highly-ranked user on CrossValidated, who has previously served as one of its moderators.

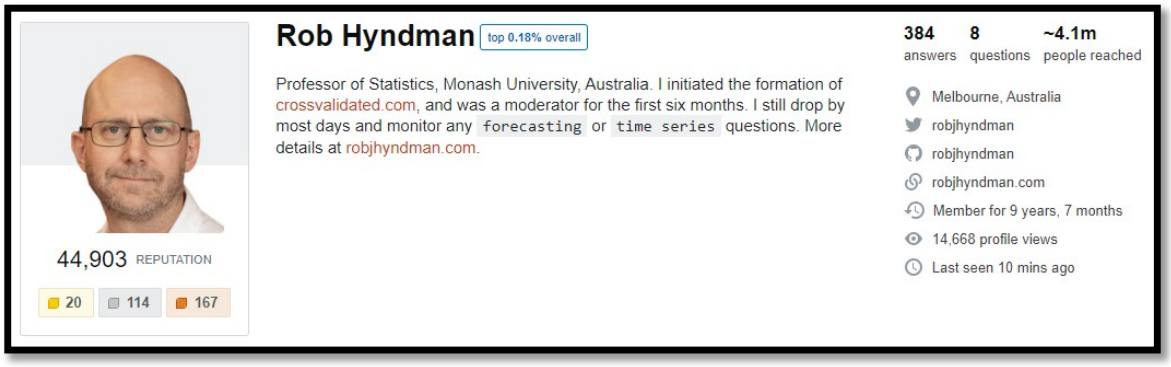


Figure 1. User profile for Professor Rob Hyndman (presently ranked #14 of 7,937 tracked users)

The voting and "gamification" on the StackExchange network have some salutary effects that assist self-moderation. Voting on questions and answers allows popular questions to gain visibility, and it usually ensures that good answers "rise to the top". Although most users on CrossValidated are statistical novices, virtually all high-reputation users are professional statisticians working in either academia or industry. Thus, while there is no necessary connection between popularity of an answer — as expressed by net upvotes— and correctness, in practice the most popular answers to a question are generally of higher quality, and wrong answers are usually downvoted to net-negative scores, or deleted. (The net score of answers is less useful when compared across questions, since it is affected heavily by the level of interest in the topic.) In Figure 2 we show an example of what a question and accepted answer look like; the numbers on the left side of the question and answer (between the up/down vote buttons) show net upvotes, and the tick on the answer shows that it was accepted by the questioner. Both the CrossValidated and Mathematics sites allow users to use LaTeX formatting to insert equations, and R Markdown syntax to add references to computer code.

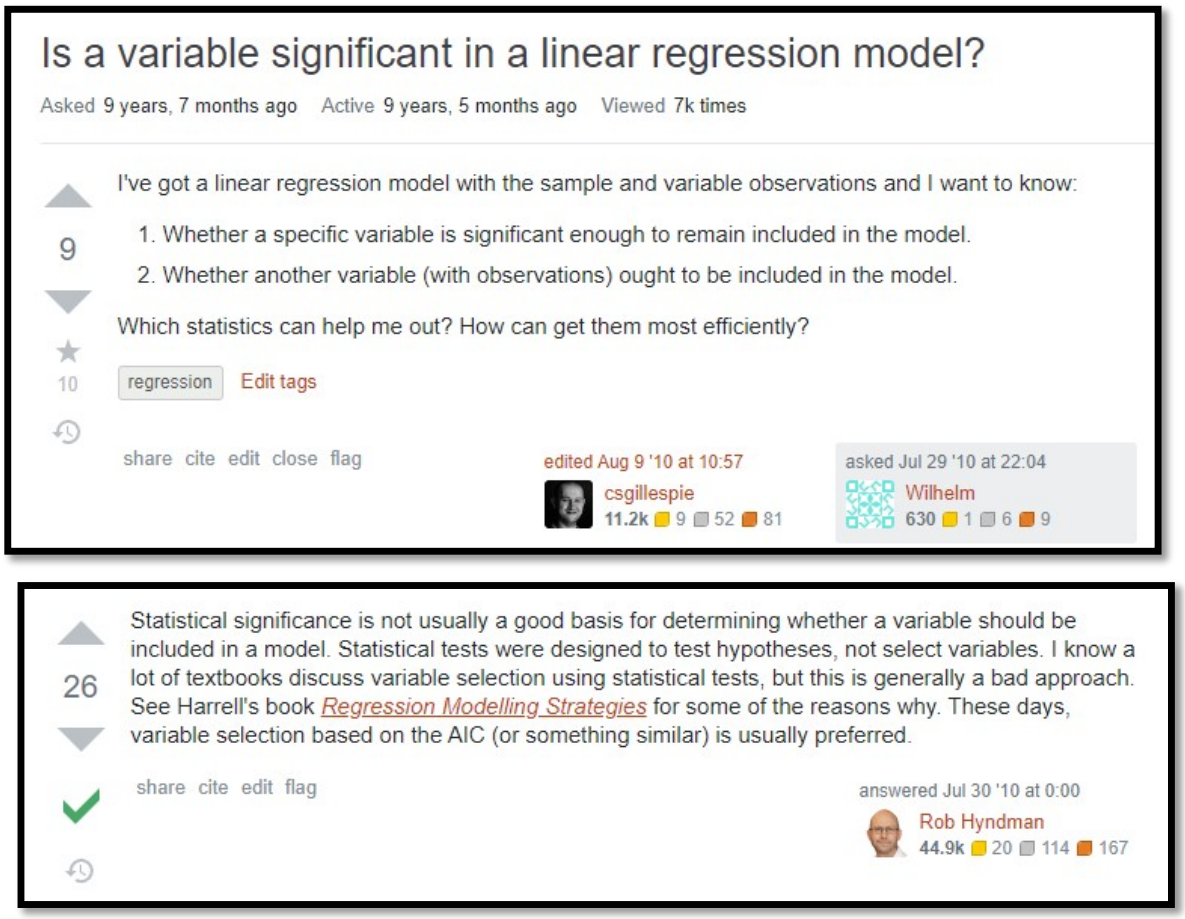


Figure 2. A question with an accepted answer; other answers and comments are omitted

An important aspect of the StackExchange network is that it constrains user communication activities solely to posting questions, answers, and comments. The site's motto is "Ask questions, get answers, no distractions", and the descriptive information further elaborates that "[t]his site is all about getting answers. It's not a discussion forum. There's no chit-chat."[iii] Users are alerted to answers to their questions, comments on their answers, and "directed" responses to their comments (if a respondent includes the username of the user they are responding to), but there are no broader contact mechanisms, and extended discussion in comment threads is discouraged. Accounts on the website are not bogged down with an "inbox" full of messages requiring attention. (Notwithstanding this limitation, most high-ranking users give enough information on their profile that you can locate and message them outside the site.) This gives the network a deliberate "less is more" quality that focuses on its core task, and ensures that administration of an account is relatively undemanding — a user can be absent from the

site without accruing messages requiring their attention when they return. The simple registration process means that many registered users on the network create accounts in order to post only a single question, and once they get an answer they do not engage in further activity. Consequently, users are classified as "tracked users" and are tracked in the reputation rankings only if they have at least 200 reputation points (which is equal to twenty total upvotes on their questions and answers).

Q&A VOLUME AND PARTICIPATION OF EXPERTS ON CROSSVALIDATED

In Figure 3 we show the weekly frequency of questions and answers on CrossValidated since 2010. In the early part of this period there were more answers than questions (i.e., an average of more than one answer per question) but there has been rapid acceleration of questions up to about 2017-2018. Since that time the moderators have made efforts to be more diligent in closing duplicate questions and low-quality questions, and this has led to a slight drop-off in recent years. Even with this moderation effort, there remains a persistent gap between questions and answers.

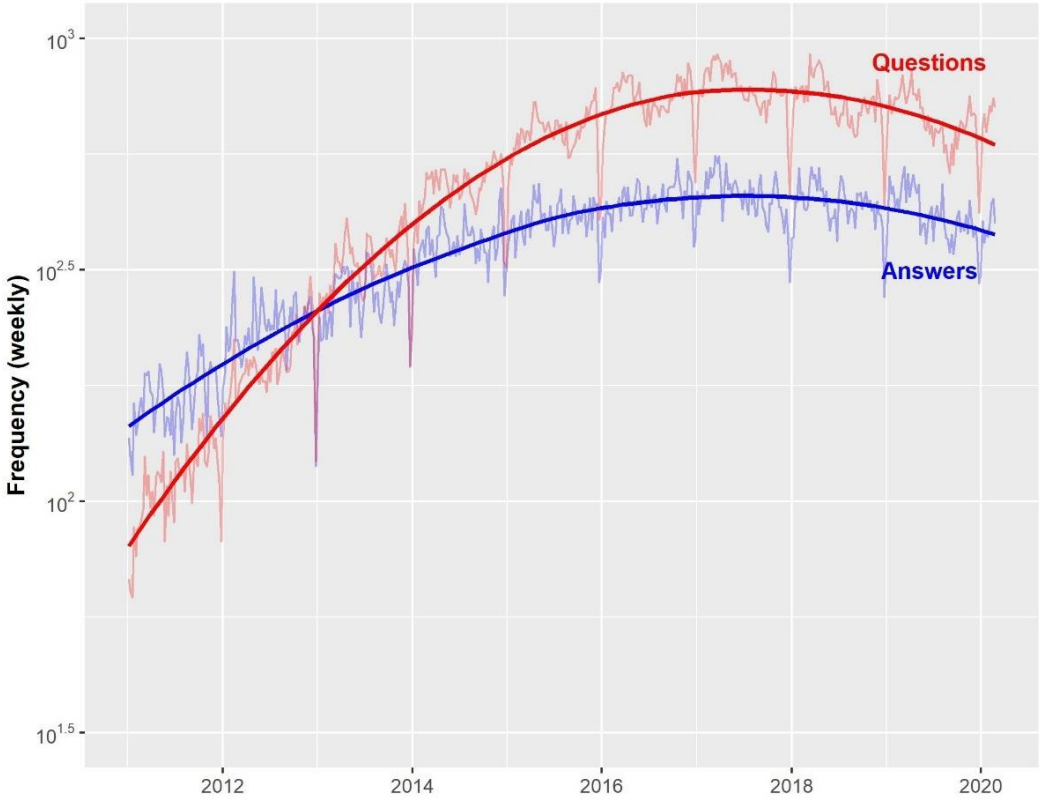


Figure 3. Questions and answers on CrossValidated (with smoothing)
Data taken from https://stats.stackexchange.com/site-analytics [Queried 3 March 2020]

Various studies have examined expert users on SQA sites on the StackExchange network (see e.g., Zhang et al 2007, Pal et al 2012, McLeod 2014). These users are generally identified by a range of metrics, including the user reputation, ratio of upvotes to downvotes received, and ratio of answers to questions. Expert users are those who answer substantially more questions than they ask, receive substantially more upvotes than downvotes, and have a high reputation score. Studies on expert users find that a small number of expert users generate a large amount of answer content on the sites. User reputation is highly positively skewed, with a large number of accounts having low reputation, and a small number of expert users with high reputation. While these characteristics are often taken as definitional aspects of expertise in literature on SQA sites, it is possible to view the profiles of top users to try to determine their professional background, to obtain an outside measure of their expertise. User details on the network are limited to what users choose to disclose in their profiles. Some users are identifiable by name, and list their profession, while others use aliases and do not disclose professional information. At the time of writing, out of the top 30 ranked contributors on CrossValidated, there were 25 with professional positions that were identifiable. Most of these users work in professional fields

involving research in statistics, data science, or machine learning, and a few work in applied scientific or economic roles that involve heavy use of statistical analysis. There were 14 users with professional academic positions, 2 statistical consultants, 2 statisticians in industry, 6 data scientists/machine learning specialists in industry, and one professional in finance and administration. The vast majority of identifiable top users hold doctoral qualifications in statistics or other STEM fields involving heavy application of statistics.[iv]

Although the top echelon of contributors on CrossValidated consists mostly of academic faculty and professional statisticians, data scientists, and machine-learning specialists, the converse is not true the vast majority of academic and professional statisticians, data scientists and machine-learning specialists are not present on CrossValidated. In the time that CrossValidated has been operational there has been a substantial increase in the rate of incoming questions on the platform, and its existing expert users have difficulty keeping up with the volume.[vi]

There have been calls within the platform to try to recruit additional experts to the site through outreach to universities and conferences. (Please consider this my outreach to this audience!) Though there may be some minor professional benefits to participation, the existing expert users on the site are "benevolent users" who contribute time without expectation of professional benefit. For users employed as academic faculty in universities, contributions on CrossValidated could potentially be counted as "service to the profession", but the answers posted do not count as academic publications. For professional statistical consultants, visibility as a high-reputation user on the network can potentially assist with professional networking. Some academics may see participation in the StackExchange network as a natural extension of their teaching activities, allowing them to reach a much larger audience than in lectures at their university. The site can also showcase a user's ability to explain statistical concepts in a clear and simple manner.

STUDENT QUESTIONS ON CROSSVALIDATED

Online learning environments offer teachers and students a platform that fosters collaboration and interaction with a body of peers (Alagic and Alagic 2013). The SQA facilities on CrossValidated and other sites on StackExchange offer some potential benefits for students and teachers in statistical courses. Use of the site has potential benefits for students: (1) a broad source of statistical expertise from experts other than their course lecturer; (2) practice framing statistical questions and problems clearly for an outside audience; (3) answers to their specific statistical questions; (4) a broader repository of statistical questions and answers on related problems; (5) practice using LaTeX and R Markdown formatting (respectively) for equations and computer code; and (6) broader technological competence using online SQA facilities. Use of the facility lets students supplement the expertise of their course lecturer with answers from other experts, allowing greater resources while outside the class. Powell et al (2017) present a model for teachers to support student use of online collaborative environments, and decentralise their own role. In addition, CrossValidated allows the lecturer to gain supplementary assistance from other experts. Notably, questions about statistical education fall within the ambit of the CrossValidated site, allowing lecturers to pose questions on how best to explain statistical concepts to students. Some of the most popular questions and answers on the site relate to explanations of the "intuition" behind statistical rules or methods.

Questions by novice users are often unclear or poorly framed, and it is common for this to elicit comments on the question seeking clarification. Novice users sometimes proceed from false premises, or ask narrow technical questions that do not adequately address statistical goals. (An example of this is questions about how to transform data to achieve normality of variables; this is rarely useful in practice since it is almost always preferable to model the original data with a model that does not assume normality.) In such cases, expert users frequently encourage these novice users to set out the overall goal of their analysis, to allow a holistic assessment of appropriate methods. This is consistent with the general experience of statistical inquiry, which involves solving complex and ambiguous inferential problems (Wild et al 2018; Makar 2018). There are many different types of statistical questions and these have different criteria for what makes a good question (Arnold and Franklin 2021). In many cases the goal of the problem requires refinement before a question is answerable. Research on the statistical reasoning of young children has exposed similar challenges concretising ambiguous inference problems (Makar 2014; Makar 2016).

On CrossValidated, users are encouraged to edit their questions with new information until the

question is clear enough to attract an answer; in cases where this does not occur, the question may be closed for lack of clarity. This process encourages students to refine their questions, and may bring to light false premises in the question, or aspects of the question where they are unable to explain clearly what they want to know. This is particularly challenging in the case of questions relating to syntax or error messages occurring in statistical programming. (Questions on statistical programming are split between CrossValidated and StackOverflow, depending on their statistical content; technical questions about syntax/programming go to StackOverflow.) In cases where users seek assistance to deal with a coding deficiency or error, they are asked to provide a "minimal reproducible example" of the problem (i.e., a set of commands that can be run by other users on the same program to replicate the error, stripped of extraneous aspects). This can be quite difficult for some students, but it encourages them to learn to describe their statistical programming problems clearly, without extraneous information.

One additional aspect of the free-flowing nature of an SQA platform that is challenging to some students is that one sometimes encounters cases where answers suggest different methods, or disagree on some statistical issue, or show a range of alternative methods to arrive at a solution to the same problem. Upvoting and downvoting gives a sense of which answers are popular among users, but the student may still need to make decisions on which answers seem most plausible. Burghardt et al (2017) suggests that users often use simple cognitive heuristics to decide which answers to upvote or accept, and they may also be "biased" towards upvoting answers that are already popular. (Though arguably, they are just being good Bayesians, making use of implicit information on the likelihood that other users are also good judges of answers.) In any case, this aspect of the student experience may be seen as a "difficulty", or it may be viewed as a constructive aspect of the use of an SQA site, since it ensures that the student is exposed to differences of opinion on technical matters, and there is no single "authority figure" to definitively settle the issue. (Though arguably there are some high-reputation users who have such tremendous levels of expertise in their topic that one hesitates to disagree with them!)

POTENTIAL DANGERS AND PITFALLS FOR PEDAGOGICAL INSTRUCTION

One potential pitfall of students using StackExchange is the possibility that seeking expert help on questions may become a substitute for personal effort and engagement with the statistical problem under consideration, such that students do not learn the material. In particular, the use of StackExchange to get answers to "textbook problems" that are assigned as homework or assessment is a possible pitfall that could diminish pedagogical success, or cause problems in formal assessment of student knowledge. Expert contributors on the network (many of whom are academics) are attuned to this issue, and there have been questions and answers on the "meta" site setting out ideas for how to deal with questions that "smell like homework".[vi] The consensus is that these "textbook problems" are useful aids for statistical learning, so these questions should not be closed or ignored. Nevertheless, contributors are reluctant to give contemporaneous answers to problems that look like they might be assessment or homework, so these problems are tagged with the "self-study" tag, and there are special rules for dealing with them.[vii] Users posting "self-study" questions must show what they have done so far to solve the problem, and which part they are stuck on. Contemporaneous answers to these questions are "hints" rather than full solutions. Since many contributors are academics themselves, these hints are generally calibrated fairly well to assist the student, without answering the question for them.

Notwithstanding these precautions, there are cases where expert contributors cannot identify whether a problem is an assigned assessment item for a student, or just a practice problem of interest for learning. The kinds of toy probability problems that regularly appear as homework items are also useful questions for general statistical learning, so it is not unusual for these questions to accrue full solutions over time. In some cases, contributors will give "hints" for recent self-study questions, but they may give worked solutions for old self-study questions (a delay of six months answering such a question ensures that the questioner cannot use the answer in an assessment item in a course occurring in that semester). This means that, over time, the site accrues a repository of solutions to textbook problems, particularly in probability theory, and an enterprising student may be able to find the solution to their homework problem with a rigorous search.

Opinions differ on whether this is problematic, and indeed, questions on this issue have arisen on the associated "meta" sites connected to CrossValidated and Mathematics. The present author is of the view that having a repository of worked solutions to these kinds of toy problems is not harmful to student learning, since students still have an incentive to learn the relevant material in order to pass in-

person examinations. (During in-person examinations the student must obviously perform without the aid of the CrossValidated website, and even for remote online examinations, the asynchronous nature of the network and the elapsed time required to get an answer to a question will usually be too long to assist the student in this context.) Answers to "self-study" problems on CrossValidated can substitute work on an assignment problem, but they cannot substitute the knowledge required to replicate good performance on an examination. Thus, the present author recommends that the StackExchange network be used as a tool to assist learning, but (unless they want to rely on an honour system) course lecturers should retain one or more in-person examination items in the assessment for their courses. This ensures that students are unable to "outsource" their assessment to experts on the StackExchange network.

IMPLEMENTING USE OF THE NETWORK IN STATISTICS CLASSES

The StackExchange network is over a decade old, and other online SQA facilities capable of assisting with statistical questions are of similar age (though message-boards and forums are older). The present generation of university students are the "canaries down the coalmine" with respect to pedagogy augmented by SQA platforms. For good or ill, their education will be affected by access to online platforms to ask questions to a broad body of experts.

Rather than leaving students to their own devices, it is possible for educators to incorporate use of the CrossValidated website or the broader StackExchange network as a formal aspect of their classes. Depending on the goals of the course, this could entail lessons designed to aid students in use of the platform, and even assessment requirements involving use of the site. The expectations of students should have regard to their own level of understanding of probability and statistics; novice users may be expected to post a reasonable question, but it will be difficult for them to provide answers that receive upvotes. In some contexts, students will already have done some statistical programming and can be expected to augment their questions/answers with coding, but in other cases this preliminary knowledge will be lacking and so this aspect of the site may be too difficult.

Teachers who wish to incorporate use of the CrossValidated website into their teaching should first spend some time creating their own account, practicing some questions and answers, and becoming generally comfortable in their own use of the site. Once this is accomplished, teachers should be able to give formal instruction covering the following material: (1) assisting students to create their own user account and understand the basic mechanics of the site; (2) assisting students with posting questions and answers about probability or statistics problems; (3) providing instruction and practice sessions on the use of LaTeX syntax to write mathematics; and (4) providing instruction and practice constructing a "minimum reproducible example" of a coding problem for data analysis or statistical programming. Some ideas for simple (non-onerous) activities for students include:

(1) **Become a user:** Create a user profile and fill it in with user details and an avatar. You may use a pseudonym for your user-name if you wish, and you do not need to give information that would identify you if you do not wish to do so.

(2) **Post a well-received question:** Post a question about a topic within the scope of the site (e.g., probability, statistics, machine learning, etc.) and receive at least one upvote on the question. If you receive comments on your question seeking clarification, edit your question until it is clear.

(3) **Accept an answer:** Upvote and accept an answer to a question you have asked.

(4) **Post an answer:** Post an answer to a question; edit your answer to improve it if it receives critical feedback or downvotes from other users. (Do not be too upset by this; it is not unusual for new users to have their answers downvoted or critiqued.)

(5) **Use LaTeX for mathematics:** Post a question or answer that uses LaTeX syntax to set out the required mathematical details. (There are a number of instruction pages for LaTeX syntax that are available online or in textbooks on the topic.)

(6) **Use computer code:** Post a question or answer using `coding font` to set out coding details for a question or answer. (This requirement is only appropriate if the students already have some existing experience in statistical programming in a scripted statistical language.)

(7) **Cite an outside source:** Post at least one question or answer where you cite an academic paper,

lecture notes, or online material, with an appropriate citation and/or hyperlink to the source.

(8) **Become a fan of another user:** Browse one or more of the existing user profiles and review and upvote at least ten questions/answers from a favourite user.

(9) **Become a generous CV.SE citizen:** Cast at least fifty total upvotes, and more upvotes than you have received. Cast at least one upvote for a question/answer from another student in the course.

(10) **Become a ranked user (more challenging):** Gain at least 200 reputation on the site.

The above activities are requirements that students in university should not find too onerous, and even at upper high-school level, some of these activities would be within the capabilities of students. Teachers could reasonably set some or all of these activities as assessable work during a course, and might also ask students to reflect on their use of the site as part of their assessment. Teachers should note that some activities require "site privileges" that are not available to new users until they earn some "reputation". For example, in order to upvote a question or answer a user must first earn 10 reputation (one upvote on their own questions/answers). Similarly, users cannot downvote a question or answer until they earn 125 reputation (a bit over twelve upvotes on their own questions/answers). Other actions, —such as editing the questions/answers of other users or conducting moderation tasks— require a high level of reputation that would be difficult for students to acquire during a single course.

One advantage of explicit instruction in using CrossValidated is that it gives educators a chance to guide students on the ethical and pedagogical aspects of using the site. Students can be alerted to the possible pitfalls to their education if they misuse the site (e.g., using it as a substitute for learning) and teachers can have class discussions on what the students hope to get out of the site. Asking students to give an end-of-course "reflection" on their use of the site could also help to inculcate good practice. In particular, the students could be asked to discuss how helpful/unhelpful it was to have access to experts in the field as an additional resource. Did they find the answers to their questions helpful? Were they too technical? Were students tempted to "cheat" by outsourcing their homework to others? Did they learn anything about probability or statistics from reading the questions and answers on the site?

If the site proves useful to students, teachers may also consider contacting a high-ranked user on the site to ask them to give a talk to the students about their work in the profession and what they had to learn to give good answers on the site. Several high-ranked users are identifiable professionals in the field who would be amenable to contact from other statistics teachers. (The present author is one of them.) In any case, the site opens up opportunities for assistance from professional statisticians and other expert users, and it provides a good source of expertise and potential networking.

CONCLUDING REMARKS

The StackExchange network is a resource available to students that allows them to seek expert assistance on questions in a range of topic areas, including specialist assistance in statistics, data science, and mathematics. In particular, the CrossValidated site provides a platform to ask and answer questions on probability, statistics, data science, machine learning, and other topics related to statistics and data analysis. This SQA site has potential teaching benefits, insofar as it provides a broad source of expert help, and it encourages students to develop capability in framing questions. Teachers should consider the availability of this platform in designing their teaching practices and assessments.

REFERENCES

Alagic, G. and Alagic, M. (2013) Collaborative mathematics learning in online environments. In Martinovic, D., Frdiman, V. and Karadag, Z. (eds) Visual Mathematics and Cyberlearning: Mathematics Education in the Digital Era (Volume 1). Springer: Dordrecht, pp. 23-48.

Anderson, A., Huttenlocher, D., Kleinberg, J., and Leskovec, J. (2012) Discovering value from community activity on focused question answering sites: a case study of stack overflow. *KDD '12: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 850-858.

Arnold, P. and Franklin, C. (2021) What makes a good statistical question? *Journal of Statistics and*

Data Science Education **29(1)**, pp. 122-130.
Burghardt, J., Alsina, E., Girvan, M., Rand, W. and Lerman, K. (2017) The myopia of crowds: a study of collective evaluation on Stack Exchange. *PLoS One* **12(3)**, pp 1-19.
Deterding, S., Sicart, M., Nacke, L.,O'Hara, K. and Dixon, D. (2011) Gamification: using game-design elements in non-gaming contexts. *CHI Proceedings*, pp. 2425–2428
MacLeod, L. (2014) Reputation on Stack Exchange: tag, you're it! *Proceedings of the 28th International Conference on Advanced Information Networking and Applications Workshops*, pp. 670-674.
Makar, K. (2014) Young children's explorations of average through informal inferential reasoning. *Educational Studies in Statistics* **86(1)**, pp. 61-78.
Makar, K. (2016) Developing young children's emergent inferential practices in statistics. *Mathematical Thinking and Learning* **18(1)**, pp. 1-24.
Makar, K. (2018) Rethinking the statistics curriculum: holistic, purposeful and layered. *International Conference on Teaching Statistics* **10**.
Meng, X., Webster, S.A. and Butler, B.S. (2013) Motivational effects of badge systems on participation in Stack Exchange social Q&A online community. *Proceedings of the Nineteenth Americas Conference on Information Systems* (Chicago), pp. 1-7.
Pal, A., Chang, S. and Konstan, J.A. (2012) Evolution of experts in question answering communities. i*International AAAI Conference on Web and Social Media*.
Posnett, D., Warburg, E., Devanbu, P. and Filkov, V. (2012) Mining Stack Exchange: expertise is evident from initial contributions. *2012 International Conference on Social Informatics*, pp. 199-204.
Powell, A.B., Alqahtani, M.M. and Singh, B. (2017) Supporting students' productive collaboration and mathematics learning in online environments. In Jorgensen, R. and Larkin, K. (eds) *Stem Education in the Junior Secondary*. Springer: Singapore, pp. 37-56.
Robson, K., Plangger, J., Kietzmann, H., McCarthy, I. and Pitt, L. (2015) Is it all a game? Understanding the principles of gamification. *Business Horizons* **58(4)**, pp. 411-420.
Wild, C., Utts, J. and Horton, N. (2018) What is statistics? In Ben-Zvi, D., Makar, K. and Garfield, J. (eds) *International Handbook of Research in Statistics Education*. Springer: Cham, pp. 5-36.
Zhang, J., Ackerman, M.S. and Adamic, L. Expertise networks in online communities: structure and algorithms. *Proceedings of the 16th International Conference on the World Wide Web*, pp. 221-230.

[i] Query of tracked users (minimum 200+ reputation) on these sites was conducted on 3 March 2020.
[ii] Highest ranked website among "reference" websites (query conducted on alexa.com on 3 March 2020).
[iii] See e.g., the description at https://stats.stackexchange.com/tour (retrieved 3 March 2020)
[iv] Query performed on 3 March 2020; see supplementary materials for details.
[v] See e.g., https://stats.meta.stackexchange.com/questions/5325/
[vi] See e.g., stats.meta.stackexchange.com/questions/12/; stats.meta.stackexchange.com/questions/2412/.
[vii] See description of "self-study" tag at https://stats.stackexchange.com/tags/self-study/info

ASSESSMENT RANDOMISATION IN STATISTICS AND RELATED DISCIPLINES

ZAMMIT-MANGION, Andrew
School of Mathematics and Applied Statistics,
University of Wollongong, Wollongong, NSW
azm@uow.edu.au

Assessment randomisation is a strategy in assessment design wherein every student in a cohort is assigned tasks that are different from all other students in that cohort. This paper describes how assessment randomisation has benefits that go beyond that of remote assessment facilitation. In particular, randomisation can be used to effectively increase student engagement and collaboration, to encourage learning, to make assessment equitable, and to reduce academic workload. Here, I discuss these benefits, along with implementation aspects in the context of subjects in Statistics and related disciplines, using the R Software package ***exams****.*

Keywords: *Equity, Online Learning, Randomised Questions, Remote Assessment, Workload*

1 INTRODUCTION

Covid-19 has been cataclysmic for the education industry. In most colleges and universities around the world, academic staff had to put their content online, and begin to virtually engage and teach their students, in the matter of a few weeks in early 2020. This has led many to change their lecture style and formats, and to develop creative ways in which to encourage student engagement. A particularly challenging aspect of shifting to online learning was remote assessment. Notably, many summative assessments that were usually carried out under strict invigilation were now being completed by students in the comfort of their own homes, with access to the internet and to channels of communication with fellow students. This shift naturally leads to concerns related to academic integrity and to the viability of the prevailing assessment model.

Assessment randomisation addresses these concerns. With assessment randomisation, every student receives assessment tasks that are different from those given to every other student in a cohort. This is an effective way to help ensure that submitted work is a student's own when being remotely assessed. The uptake of randomisation worldwide is reflected in the increased adoption of software that can generate randomised tasks, such as the *R Software* (R Core Team 2020) package *exams*. Beyond ensuring academic integrity, randomisation can lead to benefits other than academic integrity facilitation in the ecosystem of assessment strategies, even when the primary aim of the assessment is formative.

In this paper, I focus on the role randomisation plays in the assessment of Statistics subjects. After briefly describing principles of assessment that may be put into practice through randomisation, I discuss the relevance to academic workload. I then proceed to describe practical aspects of assessment randomisation and outline some drawbacks of randomisation. This paper is also accompanied by a series of 14 short video clips that contain detailed instruction on how one can implement randomisation using *R* and the package *exams* (https://andrewzm.thinkific.com/courses/assessment-randomisation).

2 PRINCIPLES OF ASSESSMENT AND THE ROLE OF RANDOMISATION

Assessment generally serves one, or both, of two purposes: to foster learning, and/or to measure outcomes of students learning to certify or accredit expertise in a subject area (Boud and Falchikov 2007). Assessment can be used for other purposes, such as to gather feedback on teaching efficacy and to adjust the teaching process accordingly (Trumbull and Lash 2013), but we do not consider these here. In this section, I describe the following assessment principles that can be put into practice through randomisation:

Principle 1: Reduce opportunities for academic misconduct.
Principle 2: Design for learning.
Principle 3: Ensure equity and inclusivity.

2.1 PRINCIPLE 1: REDUCE OPPORTUNITIES FOR ACADEMIC MISCONDUCT

In order to certify or accredit, assessments that are even partly summative in nature need to be such that they "reduce opportunities to engage in academic misconduct" (Hughes and McCabe 2006). This is particularly challenging in today's world, where answers to assignment questions are often freely shared on social media sites by students. Several "educational websites," such as *Chegg.com*, also offer services that provide answers to student questions, which are then publicly searchable online. For time-restricted high-stakes summative assessments, such as final examinations, invigilation is the *de facto* standard by which opportunities for academic misconduct are reduced. Remote invigilation is not as straightforward as in-person invigilation. For example, while online proctoring software solutions, such as *Examity* and *ProctorU*, have been used effectively to mitigate the risk to integrity in examinations, they are generally seen as intrusive (e.g., Stewart 2020), and come at a considerable cost to the accrediting institution.

Randomisation plays a straightforward role here: It provides a way to generate assessment tasks that are specific to each individual. If every student has a different assessment task, then educational websites and cheating will have a diminishing effect. Sharing of answers through social media networks or by other means will also be less effective; even if two task descriptions are largely similar, students will need to work through their own tasks and, at a bare minimum, adjust components of their responses accordingly. Behaviour among students is also likely to change: While many well-meaning students will gladly share their answers with fellow students, it is less likely that they will complete a specific assessment task for their friends. This last point is particularly pertinent in time-constrained, un-invigilated, high-stake assessments, where one generally will have no time to spare to carry out other people's work.

2.2 PRINCIPLE 2: DESIGN FOR LEARNING

There is widespread consensus that assessment and instruction are not separate entities, and that assessment is a "tool for learning" (Dochy et al. 2007). Assessments, whether primarily formative or otherwise, need to be intellectually stimulating and promote a deep approach to learning (Marton and Saljo 1997). Design elements such as scaffolding (the provision of hints and pointers to aid problem solving; see Shepard 2005), including real-world appeal (Dochy et al. 2007, Kang et al. 2014), are often used to put this principle into practice.

One may aid learning via assessment by designing tasks that promote interaction and collaboration among students. Such a strategy is likely to facilitate learning, since interaction between students is often associated with increased motivation and positive attitudes to learning (McKeachie 2007). However, when assessment tasks need to be completed by each student individually, this design strategy can be at odds with Principle 1: many subject coordinators in Statistics subjects will be all too familiar with the phrase "I was stuck, and just asked for some help" when investigating cases related to academic misconduct, where submitted assignments from two or more individuals are identical (mistakes and all!), and leave no doubt that one student blatantly copied from the other.

Randomisation has an interesting role to play here: It provides a way for students to interact, and engage with each other, while reducing the opportunity for academic misconduct. Specifically, if all assessments are different, but relatively similar, then a student who is confused by an assessment component will need to comprehend, and seek to understand, how the other student tackled that assessment component; copying verbatim would not be an option. Thus, randomisation can aid assessors accomplish what is typically deemed very difficult: designing assessments that allow students to collaborate and help each other without providing an environment conducive for gross academic misconduct. This, in turn, leads to increased opportunity for learning.

2.3 PRINCIPLE 3: ENSURE EQUITY AND INCLUSIVITY

Equity and fairness are fundamental to any assessment task, at the very least for quality assurance in measuring learning outcomes. Specific design strategies employed to ensure this principle is put into practice include the use of a language and math notation that are closely aligned to the course content, the use of contextual settings in questions that do not penalise any minority group, and anonymous marking (Boud 2007). Randomisation at first might seem to derail this principle since, 'by chance,' questions given to one student might be more challenging than those given to other students. However, it is relatively straightforward to generate randomised tasks with similar difficulty, and that require a

similar amount of intellectual engagement in order to complete satisfactorily. For example, in an assessment, one could randomise the problem context and the numbers used (see Section 4 for specific examples), but not the specific topics and learning outcomes that are being assessed through the task.

Notably, equity and inclusivity can not only be ensured when using randomisation, but also promoted. For example, students in some groups might be less socially connected to other students in the class cohort than those in other groups, and have less opportunity for direct help on solving a specific assessment task. If all students have different assessment tasks then, as discussed in Principle 1, the opportunity for students to benefit from having access to solutions is greatly reduced. This helps put all assessed students on equal footing.

In this section I have argued that randomised assessment can help enforce some important principles of assessment. However, its feasibility in terms of staff workload needs to also be considered; this is the subject of the next section.

3 ACADEMIC STAFF WORKLOAD

There is no doubt that designing and providing feedback on assessments is time consuming and resource intensive. Indeed, it would ultimately be counterproductive to the academic and their employer if the effort required to implement assessment randomisation is so high that it results in a loss of motivation and hence a reduction in job satisfaction (Houston et al. 2006). Tertiary institutions are well aware that assessment may unduly affect staff workload; for example the University of Wollongong Assessment and Feedback Policy (University of Wollongong 2020) states that "tasks need to be intellectually challenging and enable students' learning without placing undue burdens on either staff or student workloads," while the University of Technology Sydney Coursework Assessment Policy (University of Technology Sydney 2020) states that "[s]ubject assessment patterns must involve reasonable workloads for both students and staff."

While designing randomised assessment tasks is more time consuming, it is not necessarily more so than the 'traditional' assessment model, where tasks are re-designed on a regular basis in order to adhere to Principle 1. Indeed, randomisation offers a way to *reduce* academic staff workload in the long run through the shift in focus from individual assessment design to group-based assessment design. Specifically, when designing a randomised assessment task, the aim is not to design just one task that assesses the learning outcomes of the student, but to design a group of *quasi* equivalent tasks that do so. For example, if the aim of the assessment is to assess, or train, the capability of the student in performing hypothesis testing relating to the mean of the population, one would now design a (potentially infinitely-large) class of problems that targets this learning outcome, and not just one problem. More effort may be needed into finding a class of problems which target the outcome; however, once this is established, it can be used year-on-year with little or no adjustment, and with little risk to academic integrity.

Another way in which assessment randomisation offers a reduction in workload is through immediate, and automatic, student feedback. Specifically, several online learning environments such as *Moodle*, *Blackboard*, and *OpenOLAT,* as well as some platforms from textbook publishers, offer the functionality to automatically check students' answers, and automatically grade submissions. This, however, is only likely to be useful when the math or statistical problems posed are rather 'mechanical' in nature, and are given to the student for training and for the student to gain confidence, before being presented with tasks that are more intellectually challenging. In the latter case, dedicated and time-consuming student feedback is often still required.

4 RANDOMISATION IN PRACTICE

Assessment randomisation can help enforce some important principles of assessment and also make the assessment process more efficient. But how does one randomise in practice? Focusing on the conventional question-and-answer format of assessment, there are typically two ways in which randomisation can be used:

1. *Random numeric or textual entries*
2. *Random task selection*

First, random numeric or textual entries is the most straightforward way in which to randomise questions. Here, the question is the same for every student in the cohort, but selected numbers or words within the question are different for every student. For example, if the question is on hypothesis testing,

the null hypothesis, or the data on which to base the test on, could be different for each student. Each student could also be asked to prove a result (e.g., prove that the 95% confidence interval for some parameter is [0.1, 0.2]) that is different from that of other students. Text could also be randomised; for example, a student may be asked to write down the definition of an x-process, where x takes values in {"Gaussian", "Poisson", "Markov", "auto-regressive", …}.

Second, in random task selection, tasks that assess similar learning outcomes with similar difficulty are put into groups. Then, each student is allocated a task from within this group. This task could be a simple multiple-choice question, a project, an essay, or one that requires a high degree of specialised scaffolding. For example, in queuing theory, one group of questions could be assessing the student's capacity to derive the expected properties of queues, such as the expected queue length. Each question in this group would be placed in a different real-world setting, but assess the same learning outcome.

When only a few numeric or textual entries in a question are randomised, the risk for academic misconduct (Principle 1) is higher, but student interaction is promoted (Principle 2), and equity (Principle 3) is ensured. The workload on the staff member implementing the randomised assessment task is also relatively low. On the other hand, an assessment where entire questions, or tasks, are different, is more immune to misconduct (Principle 1), but discourages student interaction (Principle 2), and more effort is needed to ensure that the assessment is equitable across the entire student cohort (Principle 3). The workload associated with generating task groups is also higher.

A compromise between the two ways of randomising can reap the benefit of both worlds. For example, when assessing the student on analysing systems that can be modelled using Markov chains, when generating a random question, a contextual setting may be selected from a few and, within that setting, state transition probabilities may be randomly generated. All students would then be assessed on similar learning outcomes, for example, on the ability of finding the stationary probabilities associated with each state.

5 THE R PACKAGE EXAMS

Several online learning management systems, such as *Moodle*, provide the option to randomly generate numbers or text, and to randomly select questions from question groups. However, these systems tend to be relatively limited in the functionality they provide. For example, when scaffolding assessments, one may want to ask the student to prove an intermediate result, in which case the result needs to be computed, numerically or otherwise, for each randomised task. This can be difficult, or impossible, to do in an online learning management system.

The desired flexibility can be achieved if randomisation is carried out within a fully- fledged programming environment. For statisticians and mathematicians, the programming software *R* is a natural choice. *R* natively supports many operations carried out in Statistics, for example, operations related to hypothesis testing. *R* could be used to generate random data from randomised models, generate sophisticated, presentable plots, and much more. The *R* package *exams* provides the link between *R* and randomised assessment tasks, by allowing the user to specify the random components of assessment tasks, scaffolding through the provision of intermediate results and task-specific guides and pointers, and the corresponding solutions to those tasks. The package allows the user to generate randomised questions in a variety of ways, for example by generating multiple Portable Document Format (PDF) files (one per student), or by generating an eXtensible Markup Language (XML) file for importing into *Moodle*. A detailed discussion on the usage of the *exams* package is beyond the scope of this work; instead, I provide a series of short videos, available at https://andrewzm.thinkific.com/courses/assessment-randomisation, that give a gentle introduction to using this package. More resources and exercise templates are available on the package website http://www.r-exams.org/.

6 DRAWBACKS OF RANDOMISATION

There are a few drawbacks to randomisation that are worth noting. First, designing a random assessment task is a considerable time investment, and will only pay dividends if given to more than one student cohort. Hence, randomisation is infeasible for one-off courses, or for courses where the learning outcomes are prone to regular change.

Second, assessment task validation is more difficult. While, traditionally, the academic or tutor would work through every assessment task to gauge its difficulty and validity, this approach is no longer

feasible when assessment tasks are random. This is where the use of a software package like *R* can be particularly useful as one can test for conditions, and adjust task questions accordingly. For example, *R* could be used to flag a situation when randomly generated numbers would lead to a division by zero when problem solving, and not use those numbers in a randomly generated assessment task. Still, identifying and implementing test conditions can be time consuming.

Finally, although solutions can usually be generated for each assessment task, locating problems in a student's logic, or math working, and providing corresponding feedback will be more time consuming since questions, and numbers, will be different for every student. Appropriate scaffolding, for example by breaking down a large assessment task into small components, can be used to help the feedback process.

7 CONCLUSION

Learning and teaching in the higher education sector worldwide was thrown into disarray with the sudden onset of the Covid-19 global pandemic. It is only because technology is deeply entrenched in our lives and educational systems that most universities and institutions managed to still deliver effective learning experiences to their students. Technology-rich learning will only become more central to teaching in time, and assessment randomisation will likely play an increasingly important role. In this paper I have argued that the advantages of randomised assessment go beyond the facilitation of remote assessment, and that it may have a positive impact on student learning experience and on staff workload.

The focus of this paper was on randomisation in the context of subject learning outcome assessment. However, randomisation can be used elsewhere in the course. For example, another positive aspect to randomisation relates to the way in which it facilitates learning by practice. This approach to learning is important in several technical subjects, such as Mathematics and Statistics, where the instructor provides the students with several different contextual scenarios to which to apply knowledge learnt on a new topic. Randomisation has obvious benefits here, since the instructor may design a system that can provide the student with virtually endless opportunities to learn by practice in a given subject.

REFERENCES

Boud, D. (2007). Reframing Assessment as if learning were important. In *Rethinking Assessment in Higher Education*, edited by Boud, D., & Falchikov, N. Routledge: New York, NY.

Boud, D., & Falchikov, N. (2007). Introduction. In *Rethinking Assessment in Higher Education*, edited by Boud, D., & Falchikov, N. Routledge: New York, NY.

Dochy, F., Segers, M., Gijbels, D., & Struyven, K. (2007). Assessment engineering: breaking down barriers between teaching and learning, and assessment. In *Rethinking Assessment in Higher Education*, edited by Boud, D., & Falchikov, N. Routledge: New York, NY.

Houston, D., Meyer, L. H., & Paewai, S. (2006). Academic staff workloads and job satisfaction: Expectations and values in academe. *Journal of Higher Education Policy and Management*, *28*(1), 17-30.

Hughes, J. M. C., & McCabe, D. L. (2006). Understanding academic misconduct. *Canadian Journal of Higher Education*, *36*(1), 49-63.

Kang, H., Thompson, J., & Windschitl, M. (2014). Creating opportunities for students to show what they know: The role of scaffolding in assessment tasks. *Science Education*, *98*(4), 674-704.

Marton, F., & Saljo, R. (1997). Approaches to learning. In *The Experience of Learning*, edited by Marton, F., Hounsell, D., & Entwistle, N. Scottish Academic Press: Edinburgh, Scotland.

McKeachie, W. J. (2007). Good teaching makes a difference – and we know what it is. In *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective*, edited by Perry, R. P., & Smart, J. C. Springer: Dordrecht, The Netherlands.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing: Vienna, Austria. URL https://www.R-project.org/.

Shepard, L. A. (2005). Assessment to promote learning. *Educational Leadership*, *6*(3), 66-70. Stewart, B. (2020). Online exam monitoring can invade privacy and erode trust at universities.

The Conversation. URL https://theconversation.com/online-exam-monitoring-can- invade-privacy-and-erode-trust-at-universities-149335.

Trumbull, E., & Lash, A. (2013). Understanding formative assessment: Insights from learning theory

and measurement theory. WestEd: San Francisco, CA. URL https://www2.wested.org/www-static/online_pubs/resource1307.pdf

University of Technology Sydney (2020). Coursework assessments policy. URL https://gsu.uts.edu.au/policies/documents/coursework-assessments-policy.pdf

University of Wollongong (2020). Teaching and assessment: Assessment and feedback policy. URL https://documents.uow.edu.au/content/groups/public/@web/@gov/documents/doc/uow222905.pdf

DIFFERENT APPROACHES TO GROUP PROJECT WORK IN STATISTICS CLASSES

BILGIN, Ayse Aysin PRVAN, Tania
Department of Mathematics and Statistics,
Faculty of Science and Engineering,
Macquarie University, NSW 2109
tania.prvan@mq.edu.au

Authentic problem solving, experiential learning and work integrated learning (WIL) in statistics education are effective ways to prepare students for life after their studies at university. Project assessments provide opportunities for these experiences. Although project-based assessments usually have a heavy workload, both for students and the academics, they provide opportunities for students to obtain and/or improve soft skills, including their employability skills. In this paper, we discuss different ways of setting up project work for students in an undergraduate WIL unit and a postgraduate project unit. The similarities and differences between them and the benefits to students are discussed along with our own reflections, which include how the assessment workload can be reduced by using carefully designed rubrics. Students' reflections on their learning and experiences of project work support the value of project work.

INTRODUCTION

Group projects are identified as a pedagogical tool which contributes to developing employability skills because they provide students with opportunities to engage in active and collaborative learning experiences (Alexander, Cutrupi, Smout 2019). Authentic problem solving (American Statistical Association, 2016; Bilgin, Newbery, Petocz, 2015), experiential learning (Taback, 2018; Kolb, 1984) and work integrated learning (WIL) (Bilgin, Bulger, Petocz, 2018; Smucker, Bailer, 2015) in statistics education, which can be achieved by using project assessments, are effective ways to prepare students for life after their studies at university. In statistics, students nearly always encounter textbook data which is often unrealistic compared to data from real life problems. Group projects with authentic data which is inherently messy, possibly full of missing observations, ill-defined variables, confusion due to meta data (such as misunderstanding the variables due to cultural differences between different countries or workplaces or disciplines) provide opportunities for students to practice what they learnt on real data sets. They also give students an avenue to practice their communication skills both oral through discussions within their groups or with the project owners (if WIL) or presentations, and written through writing project reports. An additional benefit of group projects is to improve teamwork skills. Undoubtedly, feedback throughout the group project work and after it is completed is critical for student development, whether it is provided by their peers or by the lecturer(s).

Higher education has moved from being about "the pursuit of impartial truth through research and teaching" towards, making graduates job-ready through incorporating employability skills into the curriculum (Sin, Tavares, Amaral, 2019). The adoption of the Bologna process by European countries has enabled them to ensure comparability in the standards and quality of higher- education qualifications across Europe and has also highlighted the importance of embedding employability skills into discipline specific teaching as one of the priorities (Sin, Tavares, Amaral, 2019). The AdvanceHE (formerly the Higher Education Academy UK) included embedding employability as one of its strategic areas of priority for change (2015). Researchers in other countries contributing to the discussion state that "The Employability Agenda is a core driving force for tertiary education and will remain so for as long as higher and vocational education are seen to be avenues for shaping the transition of post-secondary and mature learners to work and further learning." (Higgs, Letts, Crisp, 2019).

This paper presents different ways of setting up group project work in two units to improve employability skills. The similarities and differences between them and the benefits to students are discussed along with our own reflections. Students' reflections on their learning and experiences of group project work support the value of project work.

THE AUTHENTIC GROUP PROJECTS: DATA SELECTION AND GROUP FORMATION

There are many aspects for designing a project, such as whether it will be individual or groupwork, whether data will be provided by an academic or to be sourced by the student. For senior (i.e. third year or master's) units, it is not uncommon to have an individual project supervised by an academic on an one-on-one basis, similar to a mini thesis, with a defined research question and already collected data or ready to go data collection instruments. Although it is not new to have group projects for statistics units (MacGillivray 2005), with the increased emphasis on employability skills, teamwork has become a necessity in the statistics curriculum.

Two units' groupwork designs will be discussed in this paper. With the increased emphasis on teamwork, the design of the master's unit changed from having individual projects to group projects. The third-year unit was designed as WIL experience, where it is important to have group projects instead of individual projects, since it is harder to find external projects for each individual student and it is much harder to find an external supervisor due to shortage of statistically trained people in industry (American Statistical Association, 2015). Working in groups, enables students to learn from each other, discuss the problems at hand and seek help from their academic supervisor when needed. There were 3 groups for thirteen master's students and 10 groups for 34 third-year students. Given the COVID-19 restrictions, in both units, groups met with the lecturer(s) weekly on zoom to present their progress and discuss any issues related to their projects such as identifying the suitable statistical technique(s) for their analysis, how to document their analysis, how to write the project report and technical questions related to statistical analysis. Before the final submission of the project reports, students were able to seek feedback to their written project (formative assessment) which gave them an opportunity to improve the final project report.

The master's projects were based on one big data set (WVS-W7, 2017-2021) which students could choose different countries for their analysis and come up with their own research questions. The third-year projects had external partner (i.e. industry or non-profit organisation) problems (some had data which required analysis, some had problems required students to design a study or survey instruments). Either way, the projects were based on authentic, messy and complex problems which were quite different from their previous learning experiences based on text-book data sets. The importance of exploratory data analysis (Tukey, 1977) became evident to students as well as the importance of data visualisation prior to formal statistical inference. They became aware that there is no one correct approach to solving complex problems. Students gained valuable experience working with real, authentic problems.

Finding publicly available real data or real problems from industry required a substantial amount of time prior to the beginning of semester by the lecturer(s). However, the effort and time invested in finding real data or problems led to increased student engagement. The World Values Survey Database -Wave 7 (WVS-W7, 2017-2021) was chosen for the master's projects because the fieldwork was recently completed which included responses from 77 countries with around 300 questions on a wide range of topics. The students selected the country for their group and each student posed two research questions which meant that each student could choose topics that most interested him or her and still work in a group. Choosing the country of interest and posing research questions themselves kept students motivated and resulted in each student taking ownership of his or her learning. The projects for WIL students included topics such as "Driving through floodwater: State Emergency Services experiences", "Evaluation of a mobile App for healthy lifestyle behaviour change" and "Statistical approaches for evaluating quality of care in aged care facilities". The projects for WIL unit were sourced from interested industry partners and/or non-profit organisations prior to semester starts. The suitability of the projects was carefully assessed by the academic (i.e. it can be completed in one semester, students are expected to have necessary technical skills for the given project and organisations comply with ethical practice).

Students chose their group members (master's) or were allocated to groups based on their preferences for the projects (third-year) at the beginning of the semester. There are advantages and disadvantages for both ways (Mellor, 2012). For example, when students decide who they want to work with, they usually choose their friends which might prevent them practicing some of the teamwork skills (i.e. forming, storming, norming, performing) (Tuckman, 1965), on the other hand, they quickly move to dealing with the task instead of initial team dynamics. When students choose which project to work on, they have intrinsic motivation (Bilgin et al. 2015) to complete the task.

COMPONENTS OF RUBRICS FOR GROUP PROJECTS AND FAIR MARKING

The rubrics are helpful for informing students about the expectations of the project, standardising the marking and allowing lecturers to mark different projects consistently and holistically. They are widely used in social sciences but not that prevalent in statistics. Researchers have shown that the use of rubrics could improve student learning (Reddy, & Andrade (2010). Usually assessments in statistics require students to solve a given problem where the possible answers can be used as marking guide and therefore there is no need for a rubric. When group projects on authentic problems are assessed and where each group works on different research questions, it is difficult or impossible to write a marking guide. The creation and use of a rubric becomes a necessity to ensure fair marking.

For units discussed in this paper, we developed rubrics which were shared with the students along with the group work project requirements. Due to the assessment policy in our institution, group work assessments are required to have individual parts to identify individual student contributions to the group work. The rubrics included project format (i.e. suggested sections), length of the project (i.e. either word count or page length) and various other additional information such as language requirements and individual part requirements. The suggested sections of the reports were different but similar in both units. The sections included an abstract or executive summary; an introduction where the need for the project was described and aims of the project clarified; the description of the data set; statistical methods and justification for choosing the methods; results; discussion and conclusion where most important findings are summarised, implications of the findings and the limitations of the current results are discussed along with identifying any future research questions. An abstract for one of the master's projects is given in Table 1 for the project titled "*An exploration of the 2017 -2021 World Values Survey Wave 7: New Zealand".* As can be seen from abstract (Table 1), the project was not trivial and led to new and original work. The cohesiveness of the abstract demonstrates that the students did collaborate effectively.

Table 1: A master's project abstract[1]

This report interrogates data gathered by the 2019 New Zealand World Values Survey. It explores the determinants of individual happiness to discover the factors that exert the greatest effect on happiness. It then examines the roles that individual beliefs, values, and demographic factors play in determining left-right political orientation. Ordinal logistic regression with stepwise variable selection is used to develop a parsimonious explanatory model for self-rated happiness. This report finds that the most important determinants of happiness are degree of individual freedom of choice; financial satisfaction; security; degree of belief in God; sex; and number of children. To model political orientation, a conditional inference tree is used as a variable-selection procedure over the entire dataset. This modeling determines that the factors exerting the largest effects on political orientation are economic beliefs about wealth redistribution; satisfaction with the political system; willingness to protest; trust in labour unions; and beliefs about whether homosexuals are as good parents as other couples. These beliefs and values are observed to be more important predictors than any demographic traits.

The individual parts for each student were included as appendices in the group projects. They were limited to two pages for each student where they provided their individual research questions and/or the methodology used including why it was chosen, their results and conclusion. This allowed the independent component of group work to be assessed as well as identifying free loaders, if any.

In addition to the group project rubric, in the third-year unit students were required to fill in a form for themselves (Figure 1) and a similar form for their group members to enable identification of each individual student's contribution to the group process and dynamics. Self and peer evaluations of the contributions were used to make adjustments to the group mark if there were any inconsistencies between them. This meant that as well as having an individual mark for the individual part, students might have different marks for the group part of the project.

[1] Due to confidentiality, we are not able to show an example project abstract for the third-year industry-based projects. However, a summary report shared by partner organisation can be found in a post made on 23 October 2020 at https://www.facebook.com/WolliCreekBirdos

Your Name:	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Attended all meetings with the client	❍	❍	❍	❍	❍
Attended all the group meetings without the client	❍	❍	❍	❍	❍
Was on time for the meetings (with or without the client)	❍	❍	❍	❍	❍
Asked questions to client to clarify grey issues (e.g. the aims of the project)	❍	❍	❍	❍	❍
Used active listening skills during the meetings	❍	❍	❍	❍	❍
Assisted in keeping the group on focus during the client meetings	❍	❍	❍	❍	❍
Went to the meetings prepared with available resources (e.g. project description, sample data)	❍	❍	❍	❍	❍
Learned new statistical techniques or build up my existing knowledge by interacting with my peers in meetings	❍	❍	❍	❍	❍
Improved my communication skills by interacting with my peers in meetings	❍	❍	❍	❍	❍
I would not be able to complete this project alone within the time given to us	❍	❍	❍	❍	❍

Figure 1. Self-Evaluation of Contributions to the Group Process and Dynamics

Carefully designed rubrics help to inform students about the expectations and enables them to have better learning outcomes. They use it to assess their group's project report before submission (self-assessment) and improve their report if they identify gaps or mistakes. Limiting the report either by words or page limit, forces students to synthesise their analysis and choose their words carefully for the report which helps them to improve their written communication skills. To be able to achieve these, they need to have discussions within their group, which contributes towards improvements in their oral communication skills. The benefits to academics are that they assess carefully written projects and since each project is dealing with different research questions, it is not boring to read and assess them. The rubrics help academics to be consistent and save marking time.

DISCUSSION AND PEDAGOGICAL IMPLICATIONS

There is no doubt that groupwork and real problems in industry are encountered daily therefore it is important to give students a taste of working in groups while solving real problems. Working with real data sets to answer real research questions is interesting, raises curiosity of the students and enjoyable but also challenging, especially for undergraduate students (Bilgin et al 2015). Working with a group of peers alleviate the burden on students and empower them to deal with the challenges collectively. Therefore, groupwork is potentially beneficial to increasing intrinsic motivation, defined as the doing of an activity for no reason other than the rewards in the activity itself (Ryan & Deci 2000).

In this paper we have described the design of two group work projects, one for a third-year work integrated statistical consulting unit and another one for a Master of Applied Statistics project unit. Our reflections and student feedback for the evaluation of the units indicates that students benefit from working in a team on authentic problems either provided by a client or posed by students based on real (i.e. messy with a lot of variables and observations) data. A third-year student commented on the benefits of group work "I personally learnt that I cannot get it my way all the time and realised that I must appreciate other points of view. The group projects also helped me to improve my communication skills both with group members and the client." One master's student wrote that "I was looking forward to doing an individual project before the semester started but found out at the beginning of the semester that it will be a group project. Though I didn't like the idea, I now appreciate the fact that I've learnt something from others while working with them."

The learning potential of working in groups is often underutilised (Johnson et al 2007). While working in groups, collaborative learning could provide opportunities for students deep learning through high quality social interactions (Visschers-Pleijers et al 2006). Effective collaboration can be achieved by student autonomy and self-regulatory behaviour while working on a complex task with other students which leads to something new and original (Scager et al 2016). Students learn more by discussing and sharing their knowledge with their peers.

Within the group, students improve/practice their oral communication skills with discussions amongst themselves to clarify research questions and with further discussions to identify appropriate analyses of the data. Third year students also had to do a group oral presentation of their results after completion of the project. Throughout the semester, the team dynamics could add extra workload for students as noted by a student "Main challenge was that working as a team effectively took more work" but teamwork also enables them to improve their negotiation, problem solving and time-management skills. Evaluation of the third-year unit showed that 100% of the students who responded to the survey either agreed (50%) or strongly agreed (50%) that they *developed ability to work as a team member*. Students also commented that "Being in a group kept it engaging as it put more pressure on me to work hard as my efforts not only effected myself but my group and clients as well." On a 6-point Likert scale, the average for *I can apply my knowledge in a way that helps to solve 'real-life' problems* increased from 4.3 (std = 1.2) before the group project work to 5.5 (std = 0.5) after the group project work.

Pedagogically, well-designed group project work cannot be achieved by bringing together individual work such as solving a number of questions in an adhoc manner. It requires continuous engagement, discussions and collaborations among students to complete the work. Such assessments usually have a heavy workload, both for students and the academics. However, they provide opportunities for students to develop most needed skills in their professions and for academics to mentor their students throughout the learning journey.

While group projects remove the angst of students finding a supervisor and an individual project, some students felt deprived of missing out on the one on one experience of working on an individual project and being mentored by an academic. Frustration at some group members not pulling their weight was communicated privately to the lecturer(s) by some students. In the third- year unit, use of peer and self-evaluation of contributions gave students the confidence that free- loaders will be identified and the group project mark will reflect their (less than desired) contributions. Although, co-ordination of peers in the group to prepare a coherent report was challenging for the students, it helped (at least some of) them to develop their leadership skills.

Our experience is that group projects have a place in the final year of a degree, be it undergraduate or postgraduate coursework, by providing lecturers with an assessment tool to assess whether students have consolidated their learning and opportunities for students to practice their statistical skills in a safe environment before moving to the real world of industry. Our students were fully engaged in weekly (zoom) discussions, this was noticed and commented on favourably by a peer reviewer of one of the lecturers. Students learnt how to collaborate and produce a substantial piece of written work that was coherent just like many will have to do in a workplace. Additionally, the experiences gained through the group project work could be used as a case study by the students in their job application(s) to show how they were able to work both collaboratively and independently in a group as an effective team member.

Carefully designed rubrics enabled the lecturers to communicate expectations from the beginning which was noticed by the students "… the assessment criteria and grading standards very clear…" The rubrics were instrumental marking the projects consistently and saved marking time.

REFERENCES

AdvanceHE (2015). Essential Frameworks for Enhancing Student Success - Embedding Employability in Higher Education. Retrieved from https://www.advance- he.ac.uk/guidance/teaching-and-learning/embedding-employability

Alexander, S., Cutrupi, J., Smout, B. (2019). Taking a whole university approach to employability. In J. Higgs, W. Letts, & G. Crisp (Eds.), *Education for Employability (Volume 2): Learning for Future Possibilities (pp. 117-132). (Practice Futures; Vol. 4).* Brill. https://doi.org/10.1163/9789004418707_010

American Statistical Association (2016). "Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report".

American Statistical Association. (2015, October 1). More students earning statistics degrees; Not enough to meet surging demand. *ScienceDaily*.

Bilgin, A.A., Bulger, D., Petocz, P. (2018). Industry collaboration through work-integrated learning in a capstone unit. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, 8-13 July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute.

Bilgin, A. A., Newbery, G., Petocz, P. (2015). Engaging and motivating students with authentic statistical projects in a capstone unit. In: M.A. Sorto (Ed.), *Advances in statistics education: developments, experiences and assessments. Proceedings of the Satellite conference of the International Association for Statistical Education (IASE), 22 – 24 July, Rio de Janeiro, Brazil.* Higgs, I., Letts, W, Crisp, G. (2019). *Education for Employability (Volume 2): Learning for Future Possibilities. (Practice Futures; Vol. 4).* Brill.

Johnson, D. W., Johnson R. T. (2007). An educational success story: social interdependence theory and cooperative learning. *Educational Researcher*, *38*(5), 365-379.

Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development* (Vol. 1). Englewood Cliffs, NJ: Prentice-Hall.

MacGillivray, H. (2005). Helping Students Find Their Statistical Voices. In B. Phillips & L. Weldon (Eds.), Statistics Education and the Communication of Statistics. *Proceedings of the Satellite conference of the International Association for Statistical Education (IASE), April 2005, Sydney, Australia.*

Mellor, T. (2012). Group work assessment: some key considerations in developing good practice, *Planet*, 25:1, 16-20, DOI: 10.11120/plan.2012.00250016

Reddy, Y.M., Andrade, H. (2010). A review of rubric use in higher education, *Assessment & Evaluation in Higher Education*, 35:4, 435-448, DOI:10.1080/02602930902862859

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, *25*(1), 54–67.

Scager, K., Boonstra, J., Peeters, T., Vulperhost, J., Wiegant, F. (2016). Collaborative Learning in Higher Education: Evoking Positive Interdependence. *CBE – Life Sciences Education*, 15(4), ar69, 1-9.

Sin, C., Tavares, O., Amaral, A. (2017). Accepting employability as a purpose of higher education? Academics' perceptions and practices. *Studies in Higher Education*, 44(6), 920-931. DOI: 10.1080/03075079.2017.1402174

Smucker B.J. and Bailer A.J. (2015). Beyond Normal: Preparing Undergraduates for the Work Force in a Statistical Consulting Capstone. *The American Statistician, 69*(4), 300-306. (DOI:10.1080/00031305.2015.1077731)

Taback, N. (2018). Do you have experience? Incorporating experiential learning opportunities into statistics education is messy but important. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, 8-13 July, 2018), Kyoto, Japan. Voorburg*, The Netherlands: International Statistical Institute.

The World Values Survey Wave 7 (WVS-W7). (2017-2021). Retrieved from https://www.worldvaluessurvey.org/WVSContents.jsp

Tuckman, B.W. (1965). Developmental sequence in small groups. *Psychological Bulletin*, *63*(6), 384-399.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, PA: Addison-Wesley.

Visscher-Pleijers, A. J. S. F., Dolmans, D. H. J. M., De Leng, B. A., Wolfhagen, I. H. A. P., Van Der Vleuten, C. P. M. (2006). Analysis of verbal interactions in tutorial groups: a process study. *Medical Education*, *40*(2), 129-137.

THE EFFECTS OF NUDGING ON IN-SEMESTER STUDENT LEARNING BEHAVIOUR AND EMOTIONS: A CASE STUDY OF STUDENTS AT RISK

KAUR, Charanjit and TURSUNALIEVA, Ainura
Department of Econometrics and Business Statistics,
Monash University, Caulfield East, Vic,
charanjit.kaur@monash.edu
ainura.tursunalieve@monash.edu

Student retention plays a significant role in higher education. The primary focus of the current discourse on attrition is on efforts made at the institutional level. Given that the main driver of attrition is a lack of student engagement and support, we argue that shifting the focus to individual units of study will improve completion rates more sustainably. In this paper, we introduce a case study of using personalised emails to improve completion rate in a core statistical unit that is part of a graduate programme. We sent personalised emails to students who were at risk of failing in order to nudge them towards putting more effort and achieving better academic performance. These emails are customised based on in-semester workshop attendance and assessment performance results. We then analysed student performance and engagement in learning before and after the interventions. We have also captured emotional sentiments from student replies. Preliminary findings suggest that prompting students to alter their learning behaviour early in the semester is a more effective preventive strategy for improving completion rates.

Key words: *student retention; attrition; personalised emails; nudge; learning behaviour; completion rates*

1 INTRODUCTION

Attrition is an increasingly important problem faced by Universities around the world. It refers to "the proportion of students in a particular year who neither graduate nor continue studying at the same institution in the following year" (Grebennikov & Shah, 2012). Data shows that one in ten students drop out of their course within the first year of enrolment (OECD, 2019). In Australia, four-year completion rates for commencing domestic bachelor students between 2005 and 2019 is a mere 44.5% on average (Department of Education & Employment, 2019). Various factors drive attrition. A critical factor frequently identified is the lack of engagement (Tinto, 2003). As a result, Universities are paying greater attention to addressing the need to enhance the successful completion of degree programs (Damgaard & Nielsen, 2018). In fact, the rate of successful completion is often used as an indicator of students' success and higher rates of satisfactory learning outcomes, thereby indicating a "healthy higher education system" (TEQSA, 2020).

The need to improve completion rates requires the early identification of those at risk of non-completion (TEQSA, 2020). This raises the need to prompt students to make decisions that promote better learning behaviours. One of the strategies that allow for better decision-making is nudging. Nudging is a scientific concept originating from behavioural science and economics. It was popularised by Thaler and Sunstein in 2008 in their book titled "Nudge: Improving decisions about health, wealth, and happiness" (Thaler & Sunstein, 2008). Nudging in education is a relatively new concept. It is based on the notion that indirect reinforcements can prompt students to make better decisions on their learning more than direct instruction. The techniques used in nudging can vary substantially. It is more widely applied at the school level. In cases where it involves young students, it is usually aimed at their parents in the form of text message reminders about literacy activities at home (York, et al., 2019) or weekly messages on students' performance (Kraft & Rogers, 2015). Nudging has also been widely used at the institutional level to encourage applications for federal student aid (Page, et al., 2020) or to motivate first year-students at risk (Corrigan, et al., 2015). However, there is also evidence that, if not carefully implemented, nudging can have adverse effects on those at risk (Carroll, et al., 2009); (Damgaard & Gravert, 2018); (Handel, 2013); (Rogers & Feller, 2016). Therefore, we argue that more research is essential to optimise the benefits of nudging in individual classrooms.

Given the short duration of individual semesters in a degree program, there is a need for an in-depth analysis of the effects of nudging in individual units of study. Adopting optimal intervention

strategies in order to encourage positive learning behaviours can be better realised if we have a clear understanding of how nudging can be implemented in individual classrooms where students encounter different content being presented based on varying use of pedagogy, technology and space. The challenge of promoting positive learning behaviour is exacerbated by the complexity of varying levels of literacy and numeracy amongst students (Tishkovskaya & Lancaster, 2012). Ultimately, students feel disengaged and this adversely affects their learning outcomes.

There is a need to use customised nudges aimed at specific groups of students at risk within a single unit as universal nudges can have very heterogeneous effects (Allcott, 2011).. This paper contributes to the existing literature by providing an exploratory analysis of learning behaviour changes as a result of nudging within a single unit. We implemented nudging via a personalised communication system that allows educators to provide targeted and personalised feedback to at-risk students. In the next section of this paper, we describe the data collection and methodology used. In section 3 we discuss the results and the final section refers to our conclusion.

2 DATA COLLECTION AND METHODOLOGY

The sample consists of students enrolled in a core introductory Statistics unit offered to non-specialist graduate students. Most of the students enrolled in this unit have never studied statistics or will not be studying any further statistical units. The data collection process started with a review of the outline of the assessment design for the unit. There are two major assessments as well as weekly online quizzes and workshop participation. The assessment design and timeline of nudges implemented in this unit is depicted in Figure 1 below.

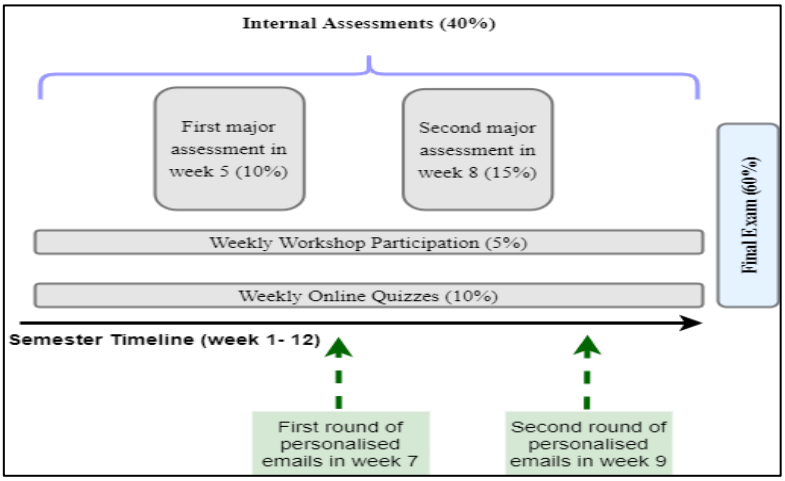


Figure 1: Assessment design and timeline for nudging

Based on the overall performance, we identified 14% of students were at-risk of failing the unit. All of these students were sent a personalised email with recommendations as a form of nudging to motivate them to improve their learning behaviour. We used the Student Relationship Engagement System (SRES) to send those emails (Liu, et al., 2017). The emails aimed to provide an opportunity for at-risk students to reach out to educators to seek help, thereby promoting dialogue between educators and students. They act as a nudge for students to improve their learning behaviour through recommendations made by educators via targeted feedback and support strategies. By providing strategies that navigate students' decision-making, the lecturer acts as a "choice architect". According to nudge theory, a "choice architect" influences decision-making by "organising context in which people make decisions" (Thaler, et al., 2013, p. 428). Depending on when the nudges were implemented, we divided at-risk students into two homogenous groups.

- ***First Nudge Group***: This group consists of students who received the personalised email after failing the first major assessment for the semester. They passed the second major assessment.
- ***Second Nudge Group***: This group consists of students received the personalised email after failing the second major assessment. They passed the first major assignment.

Based on the groupings above, we further divided students into those who failed and passed the final exam for the analysis. It should be noted that only students who read their emails were included in our analysis. In general, we found that almost 90% of students who received the emails read them. As can be seen from Figure 2, all those who passed the final exam have a greater tendency of reading the personalised emails. On the other hand, for those who failed the final exam, there is a higher proportion of students who did not read their emails.

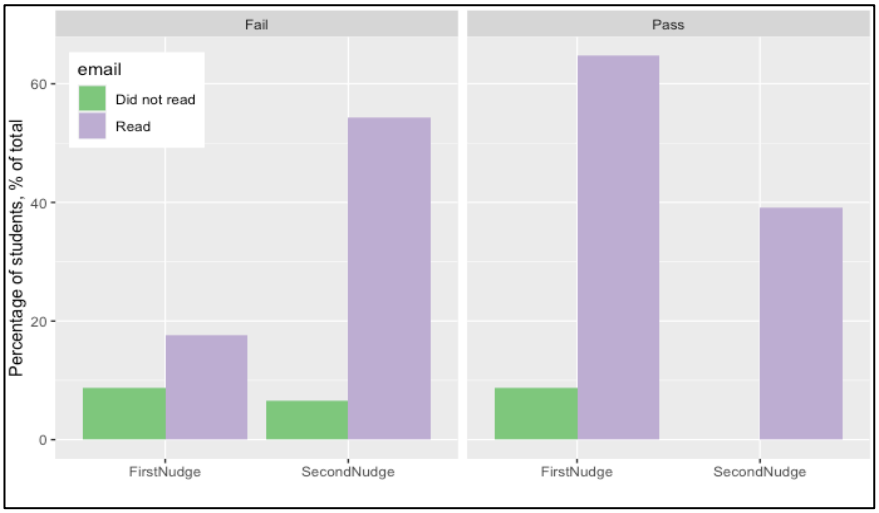


Figure 2: Proportion of students who read personalised emails sent

When collecting data on learning behaviours, we employed a mixed-method design whereby we collected qualitative and quantitative data on academic performance and engagement in weekly in-semester assessment tasks such as workshop participation and quiz performance, average number of logins into an online learning system referred to as MyLab as well as sentiments based on email responses received from students. Changes in these indicators are used as warning signals that suggest a change in learning behaviours. Data was collected from a sample of 28 students from the *First Nudge Group* and 43 students from the *Second Nudge Group*. The results of the analysis are discussed in the following section.

3 DATA ANALYSIS AND RESULTS

In general, students who received the first nudge performed better in the exam, as evidenced in Figure 3. The results show that students from the first nudge group achieved a higher average final exam mark. The results show that while 78% of those who received the first nudge passed, only 42% of those who received the second nudge passed the final exam. In order to investigate if a similar difference of performance and participation was present within the in-semester assessments and activities, we looked at various indicators of learning behaviours as discussed in the following sub-sections.

3.1 Weekly Workshops and Quizzes

The first aspect of behaviour we looked at is weekly workshop participation and quiz performance. After receiving a nudge, those who failed either of the two major assessments but ultimately passed the final exam on average participated in more workshops after the nudge. Although there is also an improvement in participation for those who failed the final exam, this improvement is very small. As for weekly quiz performance after the nudge, it was found that there was an improvement in the performance of those nudged earlier in the semester, regardless of whether they passed or failed the final exam. Nudging late in the semester is either not beneficial or results in a weaker performance.

Grouping	Final Exam Performance	Change in Average Mark	
		Workshops	Quizzes
First Nudge	Passed	4.9%	10.7%
	Failed	1.4%	10.4%
Second Nudge	Passed	4.9%	1.0%
	Failed	1.2%	-0.2%

Figure 3: Change in Average Workshop Participation and Quiz Marks

3.2 Moodle Logins for MyLab

The second aspect of behaviour we looked at is participation in an LMS platform referred to as MyLab. Participation in this platform is not compulsory, and the activities do not contribute to the overall mark. We compared the average number of logins for the two groups based on their performance in the final exam. The results obtained indicate that all students who were nudged spent less time on MyLab after the nudge. The highest decline is for those who failed the first major assessment and the final exam. For this group, the average number of MyLab logins decreases by 22 percentage points. All other groups also reduced their logins to MyLab by four to five percentage points.

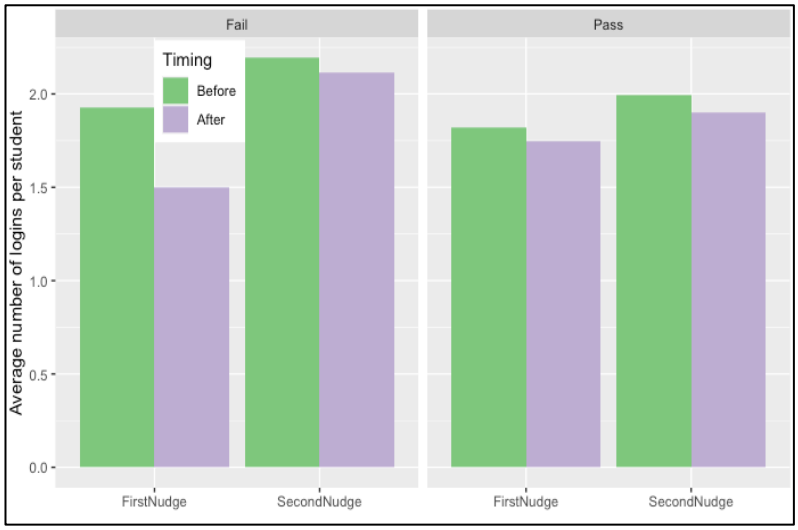


Figure 4: Average Number of MyLab Logins

3.3 Sentiment Analysis from Email Replies

The third aspect of behaviour we analysed is based on information collected from email responses received from students who were nudged. We used a TM package in R to search for negative and positive emotions. Overall, we found that there were more positive emotions for both groups that were nudged. We also found that those who were nudged earlier in the semester displayed more emotions as compared to those who were nudged later in the semester.

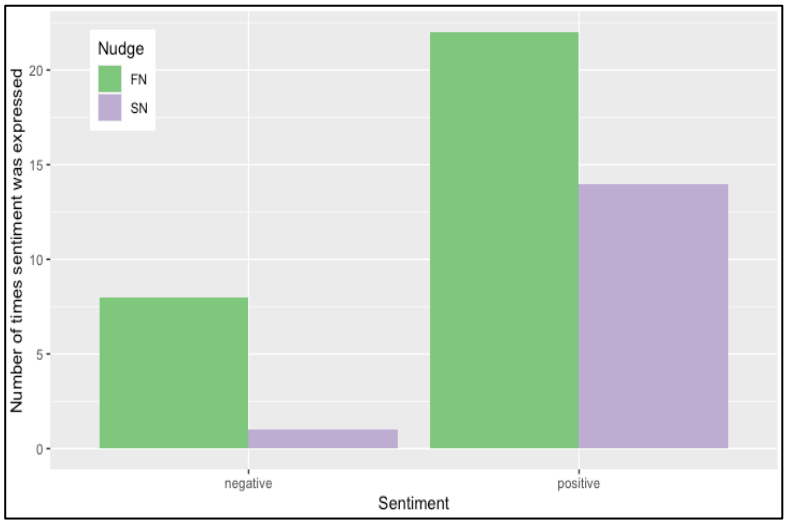


Figure 5: Sentiments of First Nudge and Second Nudge Group

4 CONCLUSION

As retention becomes increasingly important to Universities and higher education policymakers, it is important to understand the role that nudging plays in assisting at-risk students. Our research provides feedback on the effects of nudging for those at-risk within a single unit. The results provide a preliminary understanding of the differences in learning behaviour across different groups of those at-risk, depending on when the nudge is implemented. This is motivated by the fact that optimal nudging practices within a single unit can be adapted across multiple units within a study program.

There are several important points that we can draw from the findings. Firstly, the results provide evidence of the effectiveness of early intervention. Research shows that all students at risk, regardless of their propensity to succeed or otherwise, will show some indication of improved effort (Miguéis, et al., 2018). This is supported in our findings whereby after receiving a nudge, at-risk students re-prioritised their learning towards activities that count in their overall marks. They spent less time on MyLab and chose to attend more workshops, thereby indicating re-prioritisation of time spent on learning activities.

Our findings also indicate that nudging is effective if students are given enough time to steer their learning behaviour to prepare better for subsequent assessments. Those who were nudged earlier in the semester improved their performance on subsequent weekly assessments more than those who were nudged later, regardless of their performance in the final exam. This is evident in the post-nudge improvements in weekly quiz performance. This is not surprising from the educational standpoint given that early nudging allows those at risk more time to improve, especially since desired learning behaviour is often difficult to achieve quickly (Ruggeri, 2018).

Our results are consistent with most research findings that indicate offering targeted help for a specific group of students leads to positive student outcomes (Goh, et al., 2012). Based on the data, almost twice as many students who were nudged earlier in the semester passed the final exam. This is further supported by the higher proportion of positive emotions displayed by all at-risk students when nudged. It shows the desire to succeed in the unit. It would however be unrealistic to expect that all those nudged would complete the unit. Daamdard and Nielsen, 2018, argue that there are various behavioural barriers that need to be considered such as self-control, cognitive ability and default biases. We found that approximately 20% of students who were nudged earlier in the semester, still failed the final exam. Although nudging promotes better learning behaviour, as evidenced by their improved quiz performance, those who are nudged earlier may also need follow-up support mechanisms to help them through a major end-of-semester assessment. This calls for further investigation into the need for a follow-up nudge for those who fail a major assessment within the first 6 weeks of the semester, as is the case with our *Early Nudge Group*.

Within a large group of students across various units, it would be possible to assess the effects of nudging on learning behaviour across various disciplines. The results will provide the framework for behaviour modification strategies that enable improved learning behaviour in the longer term. This will

ultimately improve learning outcomes across all units, thus ensuring higher completion rates. This task is planned for future semesters.

REFERENCES

Allcott, H. (2011). Social norms and energy conservation. *Journal of public Economics, 95*(9-10), 1082–1095.

Carroll, G. D., Choi, J. J., Laibson, D., Madrian, B. C., & Metrick, A. (2009). Optimal defaults and active decisions. *The quarterly journal of economics*, *124*(4), 1639–1674.

Corrigan, O., Smeaton, A. F., Glynn, M., & Smyth, S. (2015). Using educational analytics to improve test performance. In: *European Conference on Technology Enhanced Learning* (pp. 42-55). Springer, Cham.

Damgaard, M. T., & Gravert, C. (2018). The hidden costs of nudging: Experimental evidence from reminders in fundraising. *Journal of Public Economics*, *157*, 15–26.

Damgaard, M. T., & Nielsen, H. S. (2018). Nudging in education. *Economics of Education Review*, *64*, 313–342.

Department of Education, S. & Employment. (2019). Completion Rates of Higher Education Students - Cohort Analysis, 2005-2019. s.l.:s.n.

Goh, T.-T., Seet, B.-C. & Chen, N.-S. (2012). The impact of persuasive SMS on students' self-regulated learning. *British Journal of Educational Technology*, *43*(4), 624–640.

Grebennikov, L., & Shah, M. (2012). Investigating attrition trendsin order to improvestudent retention. *Quality Assurance in Education*.

Handel, B. R. (2013). Adverse selection and inertia in health insurance markets: When nudging hurts. *American Economic Review*, *103*(7), 2643–82.

Kraft, M. A., & Rogers, T. (2015). The underutilized potential of teacher-to-parent communication: Evidence from a field experiment. *Economics of Education Review*, *47*, 49–63.

Liu, D. Y.-T., Bartimote-Aufflick, K., Pardo, A., & Bridgeman, A. J. (2017). Data-driven personalization of student learning support in higher education. In: *Learning analytics: Fundaments, applications, and trends* (pp. 143–169), Springer, Cham.

Miguéis, V. L., Freitas, A., Garcia, P. J. V., & Silva, A. (2018). Early segmentation of students according to their academic performance: A predictive modelling approach. *Decision Support Systems, 115*, 36–51.

OECD (2019). Education at a Glance 2019. In: s.l.:OECD Publishing, Paris.

Page, L. C., Castleman, B. L., & Meyer, K. (2020). Customized nudging to improve FAFSA completion and income verification. *Educational Evaluation and Policy Analysis, 42*(1), 3–21.

Rogers, T., & Feller, A. (2016). Discouraged by peer excellence: Exposure to exemplary peer performance causes quitting. *Psychological science*, *27*(3), 365–374.

Ruggeri, K. (2018). *Behavioral insights for public policy: Concepts and cases*. Routledge.

Tertiary Education Quality and Standards Agency (TEQSA). (2020). Good practice note: improving retention and completion of students in Australian higher education.

Thaler, R. H., & Sunstein, C. R. (2008). Nudge: Improving decisions about health, wealth, and happiness. New Haven CT: Yale University Press.

Thaler, R. H., Sunstein, C. R. & Balz, J. P. (2013). Choice architecture. In: *The behavioral foundations of public policy* (pp. 429-439). Princeton University Press.

Tinto, V. (2003). Promoting student retention through classroom practice. *Enhancing student retention: Using international policy and practice*, 5–7.

Tishkovskaya, S., & Lancaster, G. A. (2012). Statistical education in the 21st century: A review of challenges, teaching innovations and strategies for reform. *Journal of Statistics Education*, *20*(2).

Weeda, M., n.d. Nudge effectiveness over different educational levels.

York, B. N., Loeb, S., & Doss, C. (2019). One step at a time the effects of an early literacy text-messaging program for parents of preschoolers. *Journal of Human Resources*, *54*(3), 537–566.

EFFECTIVE QUESTIONING IN "INTERACTIVE LECTURES": AN ALTERNATIVE APPROACH

FIJN, Paul W. T.
The University of Melbourne
paul.fijn@unimelb.edu.au

Many statistics educators are making use of readily available technologies to incorporate interactive questioning within more traditional lectures. These recent technologies allow efficient participation for large numbers of students in real-time and simultaneously by allowing anonymous or crowd-sourced answers, minimising the embarrassment for asking "stupid" questions or giving "wrong" answers. Past research has focused on the different modes of questioning (open-ended, multiple choice, continuum, visual and short answer); I propose an alternative classification based on the intended purpose of the question (knowledge, discussion, participation, self-evaluation and feedback). Through better understanding of the purpose of a question, it is possible to improve the phrasing to foster more engagement and productive interaction within lecture environments. This work draws primarily on experience from a second-year introductory statistics course (Analysis of Biological Data) which is taught using a flipped classroom model, with one-hour interactive lectures each week. These question development methods have also been applied successfully in more traditional lecture environments for large (200+) undergraduate and postgraduate statistics classes.

INTRODUCTION

Traditional lectures focused on a teacher presenting content to students are gradually being replaced by more interactive lectures centred on active learning by students. This shift is aligned with strong evidence that active learning is highly effective in undergraduate STEM fields (Freeman *et al.*, 2014). Additionally, advances in technology have made anonymous responses and instantaneous feedback within a lecture environment more convenient and more flexible, facilitating its use.

Interactive lectures fall broadly into two categories in the current literature: 'fully interactive' sessions supported by content delivery outside of the classroom (either pre-reading or pre-recorded videos), as with the 'flipped classroom' model; and by incorporating regular interactivity into more standard lectures involving content. Mazur (1997) has been a strong proponent of the former, almost exclusively using multiple choice questions and think-pair-share peer learning activities throughout the lecture. Abd Rahman & Masuwai (2014) propose the CDEARA (Connect, Deliver, Engage, Apply, Reflect, Assess) model which provides a structure for incorporating regular interactivity within a standard content-delivery lecture. Both Mazur and Abd Rahman & Masuwai are primarily concerned with teaching physics; their structures and methods are readily applied to statistics education.

Regardless of the model being implemented, the effectiveness of an interactive lecture has been strongly linked to the questions that are asked (Larson & Lovelace, 2013). Much previous discussion has focused on the format of questions – multiple choice, short answer, continuum, visual, open-ended – used in interactive lectures. All of these question formats can be used successfully in different ways. For example, multiple choice questions can readily assess impressions, predictions and preconceptions; these can then be used for short peer discussions as to why they chose the particular options they selected (see e.g. Mazur, 1997). Continuum questions, where students mark a point on a continuous scale (for example 'which is more important to report?' with a scale from p-values to confidence intervals) can lead to fruitful discussions as to why some students value/prefer one aspect over another. Since all question formats can form part of an effective interactive lecture, an alternative framework for designing questions is proposed: one based on the intended purpose of the question.

This paper draws primarily on the experiences of the author in a second-year statistics subject for students with a biology (no mathematics required) background, Analysis of Biological Data. This subject is taught using a 'flipped classroom' model, with fully interactive lectures supported by pre-recorded videos and short quizzes that students complete prior to the interactive lecture. Similar question development has been used in other large (200+) statistics classes for undergraduate and postgraduate subjects, also taught by the author, but using a CDEARA model.

QUESTIONS FOR AN INTENDED PURPOSE

Subtle differences in the ways in which questions are asked can have large impacts as to the responses (or non-responses) of students. Most educators are familiar with this, via re-phrasing a question if there is no response after a substantial wait-time. Interactive lectures are often more dependent on pre-prepared questions, partly due to the difficulty in implementing them within the technology during the class itself, and accordingly designing effective questions is more important. For a student-centred learning approach, it is proposed that questions need to consider what types of thought, behaviour and learning the questions are designed to encourage and reinforce.

A classification into five main categories is suggested: *knowledge, discussion, participation, self-evaluation* and *feedback*. These can be defined as follows:

> *Knowledge* questions are designed to identify students understanding of various concepts, topics, and threshold concepts. They can also prompt students to consider the relationships between concepts in their phrasing. Some examples include: what does the standard error measure? What is the platonic world (word cloud response)? Compare and contrast Type I error and Power.
>
> *Discussion* questions typically do not have a single correct answer and are designed to prompt discussion (and deepen understanding) of relevant concepts. For example, what is more important? (Continuum response) p-values confidence intervals.
>
> *Participation* questions are best used to promote interest or engagement among students. These questions are designed to encourage students, particularly those who have not prepared for the class, to engage. These can include collecting data (pulse rate, confidence level on a Likert scale, lecture location) which is then dynamically incorporated into the lecture, or visual questions which do not rely on knowledge (that will then be linked to a concept). Also, short multiple-choice quizzes – especially in an anonymous/semi-anonymous competitive format – encourage all students to participate.
>
> *Self-evaluation* questions encourage students to reflect on their learning and identify areas to improve. This can include ratings scales where students rank their ability to understand and use notation, or an open-ended question where students nominate which concept they have the most difficulty with.
>
> *Feedback* questions can provide information to teachers as to student perceptions of the classes and also can be linked to common errors on assessment leading to peer discussion on common misunderstandings (that is, feedback from teachers to students). Some examples could be: what still confuses you about linear models (open-ended)? How could this response <example of common error from assessment> be improved?

These question purposes can be considered in conjunction with the revised taxonomy for teaching, learning and assessment (Anderson & Krathwohl, 2001), to ask questions which require students to remember, understand, apply, analyse, evaluate, and create. To varying extents, these can be applied to all five of the question purposes above.

IMPLEMENTING PURPOSEFUL QUESTIONS

Designing good questions is inevitably an iterative process. Importantly, the ways in which the questions are integrated into the lecture can also have a large impact on how effective they are. Barriers to participation can undermine otherwise excellent questions, and imperfect questions can still be effectively used to promote student discussion and peer learning.

The use of QR codes, as an efficient way to direct students to the software, enables quick access to the questions. In an online environment, posting a link in chat or using features built into the video streaming software is also effective. This minimises both the effort required from students, and the start-up time involved in running an interactive component in an otherwise traditional lecture. Accordingly, teaching time can be spent in more productive discussion arising from the activity than in the setup itself.

Participation questions are ideally suited to initial engagement with a new topic or motivating an example through collecting some relevant data within the class. In biological statistics subjects,

asking students to measure their pulse and answer if their pulse is above/below 75 beats per minute, or in an online environment "Are you wearing anything on your feet?" can both give data for discussing proportions. Short multiple-choice quizzes in a competitive environment often reward those who answer quickly (e.g. *Kahoot!*, *PollEverywhere* competitions). These also operate either anonymously or semi-anonymously (students can select a name; or where only top performers are listed) and students can participate with low stakes for selecting incorrect responses, particularly in large cohorts.

Knowledge questions can effectively be used in many ways. For difficult concepts, a single multiple-choice question (with a correct response) can be used to prompt paired/small group discussion between peers, if there is a variety of responses. Deliberately ambiguous questions, or questions about very fine distinctions can also be used similarly to promote peer learning in a lecture environment (especially when used as a think-pair-share exercise). For example, giving a very sparse description of a study design, and asking "Is this a random sample?" or an open-ended response question "What is the platonic world?". Also, simpler understanding questions (potentially as part of a poll) can be used to gauge students' knowledge of a topic. *Discussion* questions are ideally suited to stimulating a small-group (e.g. breakout rooms, in an online environment) exploration of a particular concept, or the relationship and connections between different parts of the course. An interpretation of analysis/results given by a biologist could be followed by the question "Is the biologist right? Why/why not?". These can then be drawn together as a whole class discussion, or revisited in assessment tasks.

Self-evaluation and *feedback* questions can both be used to check students understanding, confidence and engagement with the course; albeit with a different focus. Some self-evaluation and feedback questions used effectively previously are "How well do you think you understand/can use notation?" or "What still confuses you/do you find difficult about linear models?". These can also provide students with some agency in their learning, if they are invited to supply their own questions as part of their feedback (e.g. "What questions do you have about linear models?"). They can also be used to glean much-needed information on what is and is not effective for their learning, helping with iterating questions (and which topics in particular need more focus) for future cohorts within the same course.

CHALLENGES WITH INTERACTIVE LECTURES

There are some pervasive challenges with implementing interactive lectures. Commonly cited difficulties are perceived student attitudes, academic workload, and the need to cover content (see e.g. Borda *et al.*, 2020). There is some research which shows that students frequently perceive interactive (peer and or flipped) learning as less effective, in spite of the evidence this is a false perception (see e.g. Burke & Fedorek, 2017). Resistance from academics in relation to increased workload is usually focused on implementing a flipped classroom model, which can be extremely labour-intensive to implement initially. More widespread use of the CDEARA framework can increase interactivity within lectures without the workload implications. Finally, the amount of content covered is almost necessarily reduced (Borda *et al.*, 2020), however this is counteracted by the fact that students typically retain much more knowledge from active learning (Freeman *et al.*, 2014).

Truly open-ended questions (or if students are invited to ask their own) can also provide challenges, in that it is difficult to prepare for such lectures: while some responses can be predicted, frequently there will be some which would be best explained in conjunction with some prepared images, for example. This can be partially overcome by using the activity in order to contribute to future resources or lectures, rather than answering or responding to the responses within the class where the student response occurs.

Perhaps the largest challenge is in learning to develop useful questions, and effective ways in which to use them. Team-teaching combined with a community sharing effective strategies is the best way to accelerate this process.

EFFECTIVENESS

The author has used this framework to develop interactive questions and activities for lectures in a variety of large statistics subjects, taught using both a flipped learning and CDEARA model, over several years. The effectiveness of the questions has anecdotally been observed, through student participation rates during interactive lectures, strong positive feedback on student surveys ("interactive lecture questions, it boosted my understanding of how to use the statistics", [response to Tell us one thing that is good] "the active interaction and activities in lectures", "Very enjoyable when having the

lesson"), and responses to feedback questions. Currently, research is being conducted to determine specifically the most engaging and beneficial aspects of Analysis of Biological Data, the primary subject discussed. This concrete data on engagement, as measured by both behaviour and self-reporting on cognitive engagement, will help to elicit whether these positive anecdotes are grounded in genuine results.

REFERENCES

Abd Rahman, N. & Masuwai, A. (2014). Transforming the Standard Lecture into an Interactive Lecture: The CDEARA Model. *International Journal for Innovation Education and Research*, *2*(10), 158-168.

Anderson, L.W., & Krathwohl, D.R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Longman.

Borda, E., Schumacher, E., Hanley, D., Geary, E., Warren, S., Ipsen, C. & Stredicke, L. (2020). Initial implementation of active learning strategies in large, lecture STEM courses: lessons learned from a multi-institutional, interdisciplinary STEM faculty development program. *International Journal of STEM Education, 7*(1*),* 1-18.

Burke, A.S. & Fedorek, B. (2017). Does "flipping" promote engagement?: A comparison of a traditional, online, and flipped class. *Active Learning in Higher Education*, *18*(1), 11-24.

Freeman, S., Eddy, S.L., McDonough, M., Smith, M.K., Okoroafor, N., Jordt, H. & Wenderoth, M. P. (2014). Active learning boosts performance in STEM courses. *Proceedings of the National Academy of Sciences*, *111*(23) 8410-8415.

Larson, L. R., & Lovelace, M. D. (2013). Evaluating the efficacy of questioning strategies in lecture-based classroom environments: Are we asking the right questions? *Journal on Excellence in College Teaching*, *24*(1), 105-122.

Mazur, E. (1997). *Peer instruction: A user's manual.* Upper Saddle River, N.J: Prentice Hall.

MY EXPERIENCE IN TEACHING STATISTICS TO INTERNATIONAL STUDENTS IN AUCKLAND, NZ

Sawsan Al-Shamaa
Business Administration Programmes
Auckland Institute of Studies, Auckland, NZ
sawsana@ais.ac.nz

Teaching Statistics for international students pursuing a bachelor's degree in Business and MBA from New Zealand through a private institute in Auckland is a challenging experience for a statistician with 40 years' experience in teaching applied statistics in business degrees.

Teaching statistics to adult students with no background in statistics and mathematics in a limited time is difficult. This is especially the case when students have to apply and interpret statistical outputs to make real life decisions.

This oral presentation will discuss the teaching and learning challenges regarding the above. It will further elaborate the strategies that have been used to assist the learners understand the statistical concepts and interpretation of statistical outputs within the 15 hours teaching.
The presentation will cover the author's experience from 1998 to present. This period will be divided into two stages; developing statistical courses, teaching the courses to undergraduate and postgraduate students.

Solving statistical problems using Excel and other software is easy for MBA students but the challenge for them is to interpret the output and make decision accordingly.

In conclusion, most international students doing MBA degrees become very anxious using statistical application in real life.

In: A. A. Bilgin and S. Budgett (Eds.). Proceedings of the Tenth Australian Conference on Teaching Statistics, July 2021 [Virtual]