SCOPING STATISTICAL ANALYSIS SUPPORT

A Data Library Project

ABSTRACT

This report examines the current and future statistical analysis needs of postgraduate researchers across the University of Edinburgh. Supported by an ISG Innovation Fund grant, the Data Library surveyed postgraduate researchers and interviewed members of staff across multiple schools. The results reveal demand for in-person, instructor-led training on the use of R, Python and SPSS for conducting statistical analysis. The report concludes by making a number of recommendations on how the Data Library and others can develop and improve statistical analysis training at the university.

Cindy Nelson-Viljoen
Diarmuid McDonnell

31/07/2017

Introduction

Information Services (IS) have produced a long-term strategy and 10-year plan in which Research IT and Data Science feature prominently as part of the Strategic Vision. As part of the Research Data Service, the Data Library service will contribute by providing local researchers with support for finding, accessing, using and managing numeric datasets for the purposes of research and teaching. The Scoping Statistical Analysis Support project aims to increase visibility and raise the profile of the service by: understanding how statistical analysis support is conducted across University of Edinburgh Schools; scoping existing support mechanisms and models for students, researchers and teachers; identifying services and support that would satisfy existing or future demand.

The objectives of the project were as follows:

- To survey University of Edinburgh research students to find out what statistical analysis support is known to be available, what statistical analysis support they need, and what tools and resources they use.
- To engage with and interview faculty (researchers and teachers) focusing on perceived needs for quantitative methods support for their own research and that of their students.
- To report on findings from both survey and interviews to inform future planning of statistical analysis support within IS and the University.
- To liaise with internal and external stakeholders including researchers, data producers, data service providers, support and administrative staff.
- To support and promote the Data Library service as appropriate.

To meet these objectives, the project employed a mixed methods design:

- A survey of research students across the University to find out what statistical analysis support is known to be available, what statistical analysis and data manipulation support they need, and what tools and resources they use.
- Identify Schools and researchers across the university that engage in statistical analysis, and to engage with and conduct (half-hour) interviews with faculty (researchers and teachers) focusing on their perception of the needs for quantitative methods support for their own research and that of their students.

The report is structured as follows. First, the results of the survey are presented, followed by the findings arising from interviews with members of staff. A number of recommendations are made and the report concludes with a list of appendices containing information on current statistical analysis training provision.

1. Survey Results

1.1. Student awareness and use of existing statistical training and support

54% of respondents are based in a school in CAHSS, 36% in CSE, and 9% in MVM. A majority of respondents (51%) reported that they were unaware of whether their school provides statistical analysis training and support to postgraduate students; 24% indicated that their schools did provide support, and 25% reported that their school did not provide support.

Table 1 presents the percentage of respondents that are aware of and use existing statistical training and support currently provided at the university. While a significant minority are aware of the limited number 1:1 consultancy sessions provided through the Institute for Academic Development (IAD)[1], the most commonly used statistical training resource is Lynda.com (16%).

22% of respondents indicated that they engage with statistical support and training provided by organizations other than the University of Edinburgh. Examples of these include statistics and probability courses at visiting institutions, and training and programming workshops in R undertaken during internships.

Table 1. Awareness and use of existing statistical training

	% of respondents	
Statistical training	Awareness	Use
Statistics Consultancy 1:1 Session with IAD	32	6
Lynda.com	31	16
Introductory Statistics for Life Scientists – Level 1	28	9
Introductory Statistics for Life Scientists – Level 2	20	2
Statistics for the Terrified	19	9

Note: Column percentages do not add to 100, as respondents were able to select more than one response. Number of respondents = 90.

1.2. Student use of and interest in statistical software packages

Student use of and interest in statistical software packages is outlined in Table 2. Interest and use of open-source software such as R or RStudio and Python are favoured over closed-source packages such as SPSS and Matlab. The relatively high percentage of SPSS use (27%) and interest (33%) is likely due to its prominent use in statistical analysis in the social sciences. 'Other' software packages include GenStat, Graph Pad Prism, MPlus, PAST and QGIS.

[1] http://www.ed.ac.uk/institute-academic-development/postgraduate/doctoral/courses/course-list

Table 2. Use of and interest in statistical software packages

	% of respondents	
Statistical training	Use	Interest
R or RStudio	43	60
SPSS	27	33
Python	23	41
Matlab	20	37
Other	14	3
Stata	13	9
MLwiN	4	1
ArcGIS	3	6
Minitab	2	7
Amos	0	2
SAS	0	4

Note: Column percentages do not add to 100, as respondents were able to select more than one response. Number of respondents = 90. Other = GenStat, Graph Pad Prism, MPlus, PAST, QGIS.

1.3. Student use of and interest in training formats and secondary data

Table 3 describes the use of online formats for statistical analysis training; respondents indicated a preference for self-paced tutorials/courses and online discussion boards/forums. Table 4 illustrates the use of in-person training formats; respondents expressed a preference for the use of textbooks. When asked to select one preferred training format, respondents were strongly in favour of in-person training opportunities such as one-to-one consultations (see Table 5). Just under 50% of respondents stated that they use secondary data in their research. Very few respondents have used DataShare, Finding Data Portal or other data repositories.

Table 3. Use of online formats for statistical analysis training

			% of respondents	
Training format	Never	Rarely	Sometimes	Frequently
Instructor-led tutorials/courses	51	12	27	4
Self-paced tutorials/courses	32	13	39	12
Discussion boards/forums	27	13	21	33
Textbooks	27	27	21	12
Videos	29	17	21	16

Table 4. Use of in-person formats for statistical analysis training

	% of respondents			
Training format	Never	Rarely	Sometimes	Frequently
Instructor-led tutorials/courses	43	24	23	4
Self-paced tutorials/courses	56	13	20	4
One-to-one consultations	57	18	10	7
Textbooks	31	15	35	13

Table 5. Interest in formats for statistical analysis training

	% of respondents	
Training format	In Person	Online
Instructor-led workshops	79	-
Instructor-led tutorials/courses	72	46
One-to-one consultation/training	57	-
Self-paced tutorials/courses	44	53
Textbooks	26	38
Other	0	0
Videos	-	47

2. Interview Findings

A number of academics in each of the 20 schools across the university were contacted to participate in the interviews. Academics were selected based on their (potential) involvement in providing statistical training in their school; a particular effort was made to contact Director's of Research or Heads of Postgraduate Research/Teaching.

2.1. College of Medicine and Veterinary Medicine

2.1.1. Royal (Dick) School of Veterinary Studies - Roslin Institute

The Royal (Dick) School of Veterinary Studies and Roslin Institute currently provides online and in-person statistical training and support to its research community. The School is supported by in-house statisticians that provide consultancy and training, as well as opportunities to collaborate on projects. Online support is provided (and restricted) through the School intranet and includes PowerPoint presentations, videos and self-learning resources (e-learning) on topics such as basic statistical modelling, experimental design, data summary, introductory statistical analysis and mixed modelling. In-depth 5-day statistical training courses are run every October by in-house statisticians, and

complimented by ad-hoc sessions throughout the year. Though the postgraduate community is varied, additional statistical support is provided to these students on an ad-hoc basis by postdoctoral researchers. Statistical training and support is currently well-established within the school, and there are no plans to delegate statistical training and support to external providers. Minitab, SAS, Genstat, SPSS and R are the most common statistical software packages used by researchers in the School.

2.1.2. Edinburgh Medical School: Molecular, Genetic and Population Health Sciences – Usher Institute

Statistical support within the School of Molecular, Genetic and Population Health Sciences courses is mainly provided through research supervision. The school's Masters program (open to PhD students) includes three modules addressing topics in statistics: introductory statistics, statistical modelling, and practical applications of statistical techniques. Staff in the school are generally equipped with the necessary statistical skills needed to support postgraduate researchers, and the school has a number of in-house statisticians. The school developed an online statistical course for Life Sciences researchers and this was made openly available to other students through IAD (see appendices) to make it accessible for all university of Edinburgh students (though the examples are drawn from bioscience/biometric research). This is a non-credit bearing course run once in each semester; 1st semester includes the fundamental principles of inferences; 2nd semester considers study design, applications etc. There are no plans to develop further training and/or support resources within the school.

2.2. College of Science and Engineering

2.2.1. School of Physics and Astronomy

Within the School of Physics and Astronomy, research staff (postdoctoral researchers, fellows and academics) are usually proficient in statistical analysis techniques before appointment to their role; therefore, they are capable of supporting postgraduate research students. Statistical training workshops are run on an ad-hoc basis to develop expertise in specific and/or specialized statistical theory throughout the academic year, though not necessarily by providers based at the School or University of Edinburgh. The EPCC (formerly the Edinburgh Parallel Computing Centre, a supercomputing centre based at the University of Edinburgh) provides a range of training programs to members of the School of Physics and Astronomy. PhD students are further supported by, and members of, the Scottish Universities Physics Alliance (SUPA) Graduate School; they are required to complete 40 hours of technical courses and 20 hours of core skills training in topics such as advanced statistical physics, introductory and advanced data analysis. Staff and students that are members of wider collaborations such as CERN (the European Organisation/Council for Nuclear Research) participate in summer schools, workshops and analysis retreats. Within the School most researchers are part of small groups, with ad-hoc discussions and problem solving between team members.

2.2.2. School of GeoSciences

Structured statistical training and support is provided on a limited scale via online videos and e-learning resources. Students are encourage to learn software packages and statistical analysis in an 'organic process of learning' while they are doing their research. Research students are supported predominantly by their supervisors and within independent student-led discussion/supports groups that meet to discuss statistical analysis techniques and software (predominantly Python and R). In-depth courses on environmental data are run by NERC (Natural Environment Research Council) in Glasgow, and are free for students on taught degree programmes in the School of GeoSciences. There are currently no plans to develop further statistical training and support within the School. Software packages used predominantly within the School of GeoSciences include R, Version Control systems and Python.

2.2.3. School of Mathematics

Training in statistical analysis is predominantly provided in taught courses within the School of Mathematics. The School does not provide specialized in-house training or consultancy for postgraduate research students. Postgraduate students are encouraged to sit-in on lectures (on a voluntary basis) and be involved in tutoring. PhD students receive 4 to 22 weeks of statistical training via the SMSTC (Scottish Mathematical Sciences Training Centre). Though the school does not have the current resources, they are interested in developing an online basic statistics course for MSc students that could also be used throughout the university. This notion is supported by the success of the School's recently launched MOOC (Statistics: Unlocking the World of Data). This six-week online course is open to all University of Edinburgh staff and students, and explores the ideas and methods behind the statistics encountered in everyday life.

2.3. College of Humanities and Social Science

2.3.1. School of Health in Social Sciences

The School of Health in Social Sciences does not provide in-house statistical training to PhD or masters by research postgraduate students. However, psychology conversion program students and some Masters students do develop proficient, but non specialized, statistical analysis skills through programme-based statistical training. Support for postgraduate research students is provided mainly through thesis supervision, access to statistical courses run by the School of Social and Political Science, and online resources via Learn. A specialized training course in statistical modelling is occasionally provided on demand, and run by experts outwith the university. The school has no plans to develop any statistical training and support programs. Software predominately used within the school: SPSS, Dedoose, Nvivo and Qualtrics.

2.3.2. School of History, Classics and Archaeology

Traditionally the School of History, Classics and Archaeology did not provide focused statistical analysis training and/or support. Statistical support is provided mainly through thesis supervision. Archaeology courses, however, are currently being re-designed to include statistical theory and practice in undergraduate and postgraduate courses. These include compulsory courses for MSc students (voluntary for PhD's) on introductory statistics, how to design and test hypotheses, and what test to use and how to use them. The focus of these courses is not about specific software packages, but rather to make students think quantitively. Archaeology is also developing 2nd year undergraduate and postgraduate courses about data science and data visualization and will include topics such as basics statics, how to plot data and detect patterns. The courses will include lectures and tutorial practicals with hand-outs and exercises. Software used include SPSS and R. Though the school does not provide consultancy, recently appointed staff may be utilised as a potential source of statistical support for postgraduate students.

2.3.3. School of Law

Though no official training is provided to postgraduate research students, they can seek assistance from the School of Social and Political Science. Support within the school is provided predominantly via thesis supervision. Postgraduate students are however required to complete a 'training needs analysis' within their first few weeks of starting their program to identify what skills are needed. Any additional training needs identified are then provided on an ad-hoc basis as there is no 'critical mass' within the school itself to invest in in-house training. The packages used most often by researchers in the school are SPSS, R and STATA.

2.3.4. School of Social and Political Science

The school of Social and Political Science runs a number of university-wide undergraduate and postgraduate courses in statistical analysis: one such offering is Statistical Literacy, an introductory course that currently runs during the 2nd semester (will move to 1st semester for academic year 2018/2019). For postgraduate students, the school runs a core quantitative data analysis course during semester one, and includes training in the use of SPSS. During semester two, the course covers intermediate inferential statistics. The school also hosts Q-Step, a centre intended to equip students with quantitative skills via specialist degree programs, internships and quantitative courses for all students regardless of degree program.

3. **Summary**

A central theme that immerged with staff interviews is that the diverse and large research profile and interests of postgraduate students make bespoke statistical training and support for student cohorts problematic. The level of statistical knowledge and experience among the student body varies greatly between disciplines and amongst students enrolled on the same course. Students also do not engage in the use of datasets from disciplines other than their own, as it seems to act as a barrier for student engagement. Many Schools also do not have the student numbers to justify the design of School/subject specific statistical courses. Additionally, workload prevents many staff from fully engaging in statistical support for their research students. Staff commented that statistical training and support should be provided by PhD supervisors and included in job descriptions and work schedules.

Staff favour online training (e-learning) via videos, power-point presentations and self-paced courses as they are more cost and time effective than intensive in-person courses, though admittedly not as effective. Additionally, staff question the accessibility of training courses run once a year as many postgraduate researchers have complex timetables and do not necessarily start their program at the beginning of an academic year. Furthermore, staff criticised the uncoordinated nature of online and in-person support currently provided within the university, with many staff unaware of the services listed in the Appendix.

A ‘Centre for Statistics’ is currently being developed by staff at the School of Mathematics. The network will focus on research and aims to join researchers across the University through interdisciplinary research events and seminars. Though the centre will based at the School of Mathematics, it will not be a School centre but rather an inclusive central system for statisticians from all University of Edinburgh Schools to network and collaborate. As it stands, the centre will not have the resources to provide consultancy or statistical training to students. The centre aims to launch in autumn semester 2017.

4. Recommendations

4.1. The Research Data Service and statisticians/staff throughout the University should join and interact with the proposed 'Centre for Statistics'. Additionally, the Research Data Service should work with the Centre to develop statistical training courses that are run by the Research Data Service. It is proposed that the Research Data Service should invest in CPD workshops in consultation with the Centre for Statistics.

4.2. As both staff and students indicate that training should be provided more regularly, it is proposed that statistical training in the most commonly used software packages such as R, Python and SPSS should be run quarterly.

4.2.1. The training should focus on data management, teaching syntax and the use of software packages to do statistical analysis. Two levels of training are proposed: Introductory (for students who have some knowledge of statistics and no/little experience with statistical packages) and Intermediate (for students who have a good understanding of statistics and require a more in-depth course on using software in statistical analysis). Suggested topics include simple descriptive statistics; frequency tables, correlations, regression, interval tables, cross tabs, contingency tables etc.

4.2.2. It is proposed that the training is reviewed to ensure quality, and that the courses are designed in such a way to allow for different datasets to be 'plugged-in' by instructors. As it has been noted by staff that students do not want to use datasets from different disciplines it is further suggested that each training session run should be focus on for example 'Social Sciences', 'Humanities' and 'Life Sciences'.

4.2.3. The development of online training via the Research Data Service is not advised as there is currently various online statistical courses and resources available to University of Edinburgh researchers.

4.3. It is proposed that the Research Data Service provides a Helpdesk service, either online or in-person, to signpost research students and staff to the relevant statistical courses and resources available throughout the university (see Appendix).

5. **Appendix**

- UoE MOOC: Statistics: Unlocking the World of Data https://www.edx.org/course/statistics-unlocking-world-data-edinburghx-statsx
- UoE Q-Step http://www.q-step.ed.ac.uk/
    - Mathematics for Social Science
    - Introduction to Statistics for Social Science
    - Doing Social Research with Statistics
    - Designing and Doing Social Research
    - Analytical perspectives in Social Policy
    - Statistical modelling in Social Sciences
- UoE Digital Scholarship http://www.digital.hss.ed.ac.uk/
    - Have an ever-changing suite of workshops and events (e.g. **Statistical Literacy: What You Need to Know and Why**).
    - Drop-in sessions aimed at researchers and PGs with computational/technological questions and problems for which they'd like advice. Each session will have a different specialist on hand.
- UoE IAD http://www.ed.ac.uk/institute-academic-development/postgraduate/doctoral/courses/online-courses
    - Introductory Statistics for Life Scientists - Level 1
    - Introductory Statistics for Life Scientists - Level 2
    - Statistics Consultancy 1:1 Session
    - Data management training
- UoE Information Services
    - **Statistics for the Terrified** is a self-paced, online, basic statistics tutorial, written in plain English, with only a tiny bit of mathematics. It explains common statistical concepts using common sense terminology and explanations, and includes plenty of interactive exercises and animated illustrations.
- UoE Research Data Service http://www.ed.ac.uk/information-services/research-support/research-data-service/training
    - Handling data using SPSS
    - Introduction to NVivo
    - NVivo: Beyond the basics - Queries
    - Introduction to visualising data in ArcGIS
    - Introduction to visualising data in QGIS
- UoE miscellaneous

    - AQMeN is a quantitative research and training network based at the university. Individuals can access free training materials, including those from previous workshops - http://aqmen.ac.uk/QMresources
    - Lynda.com – can be accessed through MyEd.
    - Data Carpentry (see also Software Carpentry) develops and teaches workshops on the fundamental data skills needed to conduct research; materials are available from previous workshops - http://www.datacarpentry.org/workshops/
    - EPCC is an international centre for excellence in high-performance computing for over 25 years - https://www.epcc.ed.ac.uk/education-training
- BioSS (Scotland) - http://www.bioss.ac.uk/training/courses.html -
    - They provide consultancy and some training (their list is quite extensive but the popular ones are listed below)
    - Basic statistics courses
    - Getting started in R
    - Regression and curve fitting
    - Grahpical methods for Multivariate data
    - Statistical Methods for Repeated Measures data