EasyChair Preprint

№ 549

The Role of Statistics Education in the Big Data Era

Ryan H.L. Ip

October 1, 2018

The Role of Statistics Education in the Big Data Era

Ryan H.L. Ip

School of Computing and Mathematics, Charles Sturt University, Wagga Wagga, NSW, Australia
hoip@csu.edu.au

Abstract. With the increasing availability and trendiness of "big data", data science has become a fast growing discipline. Data analysis techniques are shifting from classical statistical inferences to algorithmic machine learnings. Will the rise of data science lead to the fall of statistics? If education is the key to defend statistics as a discipline, what should statisticians teach? This paper aims to provide the current situation of data science and statistics programs within the higher education sector in Australia and some personal thoughts on statistics education in this era.

Keywords: Statistics teaching, Data science, Small data

1 Introduction

Since the emergence of "big data", there has been a long debate on the role of statistics in the modern world flooded with data. I guess most people would agree that we are living in the era of big data. The term "big data" is not only trendy within academia, but is also frequently used by the media, companies and even ordinary people. Recently, *Statistics & Probability Letters (SPL)* published a special issue "The role of Statistics in the era of Big Data" (Volume 136). This special issue collects opinions from statisticians, computer scientists, data miners and other experts from related domains. A key message brought out from the special issue is that - statistics is still a central part of data science (or science, in general), despite the fact that some argue that statistics - especially sampling and modelling - is no longer relevant when sufficient data are obtained [1].

Although there does not seem to have a consensus on the definition of "big data" and the difference between "big data" and "small data", it may be useful to be more specific. Along the lines of [7], the term "big data" is used to refer to datasets with large number of observations (with possibly large number of variables as well) that are usually collected from observational studies, and are even too large to be handled by machines. In contrast, the term "small data" is used to represent datasets that result from designed experiments and are possible to be handled manually, or at most with an ordinary computer. Statistical techniques, at least the classical ones, mainly focus on analysing small data while the focus of "data science" is usually on extracting information from

big data. Despite saying so, the two disciplines are, and will never be, mutually exclusive. As observed by Reid [13], some statistics departments or schools in colleges or universities have renamed themselves to integrate with "data science" or "data analytics". Such a movement is likely to be strategic to secure funding and student number due to the fact that statistics does not seem to have a good fame [8] while "big data" or "data science" are labelled as "sexy" [4]. In fact, I would not be surprised to see statistics, as a discipline, will be merged with, replaced by, or placed under, data science in the next five to ten years. In order to defend statistics as a discipline, [10], [13] and [15], among others, indicate the importance of statistics teaching. In particular, Piecresare Secchi called a debate on "how we should teach the next generation of statisticians" [15, p.11]. It is of course important to teach statistics to the next generation of statisticians. Yet, we must not ignore the fact that it is equally important to teach statistics to the next generation of scientists, businesspeople, educators, among many other domains, as we must rely on these people to rebuild the reputation of statistics.

This paper attempts to enrich the discuss on statistics teaching in the era of big data by describing what is happening within Australia and suggesting some statistical ideas and concepts that should be emphasised by statistics educators to respond to the opportunities and challenges brought by big data.

2 The Rise of Data Science

In Australia alone, 30 out of 42 universities have launched the Master of Data Science (MDataSc) program or similar (including similar names such as Master of Data Analytics or Master of Predictive Analytics, and specialisations/streams under various Master programs such as Master of Information Technology, Master of Science, Master of Computer Science or Master of Business Administration). Where data are available, all these programs were launched in or after 2015. The full list can be found in Table 1. Even if specialisations/streams are excluded, there are still 21 universities on the list. In the opposite, Table 1 shows that only seven universities are offering Master of Statistics (MStat) or Master of Applied Statistics. Combining also Master of Biostatistics, Master of Medical Statistics and various master programs with statistics as an optional specialisation, potential students have 17 options. Considering the "ages" of the two disciplines, the difference is even more striking.

Such a phenomenon can also be observed in other areas. For example, being on of the leaders in statistics in Asia, the University of Hong Kong is launching the Master of Data Science program in 2018. James [10] has reported a similar trend in the United States. While the supply may not necessarily reflect the demand, the abundance of the MDataSc programs indicates, at least, the confidence of student intakes (and hence revenues) from the universities' points of view.

In terms of the contents, MStat programs often focus on advanced modelling techniques and probability theories. Typical subjects cover gerenalised linear models, time series models, longitudinal analysis, measure theory, multivariate

analysis, survival analysis, non-parametric statistics and Bayesian techniques. Besides, MStat programs often involve some mathematics subjects such as partial differential equations and real analysis. Notably, data mining is also a commonly seen subject in MStat programs. On the other hand, MDataSc programs often focus on algorithmic learning. Typical subjects cover database management, programming, cloud computing, data mining, machine learning, visual or graphical analytics and artificial intelligence. In addition, MDataSc programs usually involve one or more statistics subjects.

Although the contents of MDataSc and MStat are different, it is likely that an interested potential student will not enrol in both programs. With a larger number of options and a "sexier" title, perhaps data science has won the battle if we consider the market as a battlefield. Thus, the rise of data science may in turn lead to the shrinkage of statistics as a discipline.

Is it a threat? Should statisticians, especially the young ones who have no plan to retire in the next ten years, be worried? The answer is perhaps "yes, but not now". As aforementioned, most MDataSc programs involve one or more statistics subjects at the moment, which indicates statistical thinking and techniques are still a central part, or a foundation, of many techniques in data science. Statistics subjects may still serve as "service subjects" like in all other disciplines such as education and science. Moreover, a statistician may still claim to be a data scientist until there is a proper definition and a clear distinction. At the moment, statistics can still sail along the data science boat.

Nonetheless, statisticians should step up and act fast to defend the discipline and rebuild the brand name, before it is too late. Apart from research, which is the main perspective of the papers in the special issue in *SPL*, statisticians should grasp the opportunity of teaching statistics, even at the introductory levels, to stress the importance of statistics in this era. Quality teaching is certainly required [21] but the context is also important.

3 What should we teach?

In a world full of big data, what should statisticians offer to the students, especially the undergraduates who do not major in statistics, to help them appreciate statistics? Admittedly this is a challenging task. This section is devoted to list several important statistical concepts or topics that shall be emphasised. The list is far from being complete but hopefully it can shed some light and promote further discussions. Moreover, some of the topics may be more suitable for introductory level subjects while some are more suitable for advanced subjects.

3.1 Small data are still prevalent

First and foremost, educators should let students know that small data are still commonly seen and analysed. Despite the fact that many people are talking about big data and big data are becoming more and more available, such datasets are not as prevalent as one may perceived.

Articles in the first issues in 2018 for four journals: *Crop and Pasture Science*, *Australian Education Researchers*, *Rural and Remote Health* (only original research articles were reviewed) and *Journal of Diabetes Investigation* were reviewed. These are leading journals in their respective subject areas in Australia according to *Scimago Journal & Country Rank* [14]. The review focused on the sample sizes considered in these articles. Purely qualitative studies, case studies, letters to editors and editorials were excluded. Following [11], sample sizes over 10,000 were considered to be "large". Out of the 44 relevant articles reviewed, 38 (84%) of them considered samples of sizes less than 10,000. Among the four journals, articles published in *Crop and Pasture Science* considered large datasets the most, which mainly came from genomic studies. Other large datasets considered in other journals arose from meta-analyses. In the opposite, none of the articles in *Australian Education Researchers* considered large datasets as the student numbers in the studies were often limited. Although the list in Table 2 is by no means complete and representative, it shows that small data are still commonly analysed in the literature even though we are living the era of big data.

It is not hard to understand why small data are still prevalent in the literature. The existence of big data does not mean the data are available to everyone (scholars included) since the collection of big data often involve a huge cost. Even if the data can be collected with a low cost, various reasons such as privacy issues limit the availability. Since small data are still widespread, students from various disciplines need to learn the right tool – statistics. I am not denying the importance of machine learning techniques, but statistical techniques are simply more relevant in analysing small data.

3.2 Quality beats quantity

Given that small data are still prevalent, educators should emphasize the quality of data rather than the quantity while teaching statistics. It is always better to have small data with high quality than big data with low quality [7]. Quality datasets should at least be representative to the desired target population. As aforementioned mentioned, big data often come from observational studies rather than designed experiments. Without a properly designed collection process, there is a risk that the sample does not represent the target population [5]. The analysis results may thus be biased. Such a bias may not be obvious and is often hard to quantify. From this angle, collection of data is of high importance and experimental design would be the most relevant topic for students.

3.3 Estimations rather than hypothesis tests

Classical statistics always focus on hypothesis tests. From my personal experience, statistics educators (including myself) tend to teach students to follow a routine (often a five- to seven-step process) in conducting hypothesis tests and to compare the p-value with a pre-specified threshold (usually 5%). In this way, students, especially whose mathematical abilities are not strong, can at least get some marks in the assessments. However, such a practice of teaching is likely

to lead to confused ideas about hypothesis tests and various bad consequences such as "p-hacking" [17] and misinterpretation and over-reliance of p-values [19]. Moreover, in the big data era, when the sample size is large enough, everything will be statistically significant [11].

In fact, the usage of hypothesis tests and p-values have long been doubted by many scholars (see, e.g. [2]). Various remediations have been suggested, ranging from lowering the significance level [9] to declaring an end for p-values [6]. Completely moving away from hypothesis tests will not come to a success unless there is a radical change in the way how statistics is taught [18]. The need to focus on estimation and practical significance rather than hypothesis tests and statistical significance while teaching statistics has never been greater due to the emergence of big data.

3.4 Importance of assumptions

Almost all classical statistical techniques were built upon a set of assumptions. Some assumptions are strict while some are relatively weaker. Even for small data, diagnostic checks are rarely mentioned or reported in the literature. Statistical results may simply become invalid when one or more assumptions are not met. The situation escalates when big data are analysed [3].

"Assume assumptions are met" is a common phrase in various statistics texts and lecture notes. Such an assumption is too optimistic to be true in practice. Statistics educators should instead encourage students to critically challenge the assumptions more often. Examples where assumptions were not satisfied and the consequences should appear more frequently in texts and teaching materials.

3.5 Dependency structures

Related to the previous point, independence is often an important assumption in classical statistical techniques. Yet, observations in large datasets are rarely truly independent [12]. The effect of dependencies among observations would be substantial. For example, in spatial analysis, ignoring the positive correlation among observations often leads to underestimation of the standard error and thus narrowing any confidence intervals formed [16]. In practice, the dependency structure may be far less obvious and far more complicated. As described by Wit, while we have never been closer to the long-fancied situation, namely $n \to \infty$, the "bigness" of data actually only leads to the "death" of classical asymptotic results [20].

Thus, when time and resources permit, statistical techniques which handle dependent variables should be included in the syllabus. Topics may include time series, spatial statistics and multivariate statistics. It does not mean results relying on the assumption of independence should not be taught. Undoubtedly these results form the foundation of more complicated theories. However, it is perhaps time to rethink the curriculum and spare time for the above topics.

4 Conclusion

To defend the shrinking discipline, education is the key. Quality teaching and state-of-the-art contents that respond to the needs of the era are needed. Hopefully this short paper would motivate more discussions on what should be taught and how should we teach.

Acknowledgement

Helpful comments from three reviewers are greatly appreciated.

References

1. Anderson, C.: The end of theory: The data deluge makes the scientific method obsolete (2008). URL https://www.wired.com/2008/06/pb-theory/. Last Accessed: 26 Jul 2018
2. Berger, J.O., Sellke, T.: Testing a point null hypothesis: The irreconcilability of P value and evidence. J. Am. Stat. Assoc. **82**, 112–122 (1987)
3. Cox, D.R.: Big data and precision. Biometrika **102**, 712–716 (2015)
4. Davenport, T., Patil, D.: Data scientist: The sexiest job of the 21st century. Harvard Bus. Rev. **90**, 70–76 (2012)
5. Dunson, D.B.: Statistics in te big data era: Failures of the machine. Stat. Probabil. Lett. **136**, 4–9 (2018)
6. Evans, S.J., Mills, P., Dawson, J.: The end of the p value? Brit. Heart J. **60**, 177–180 (1988)
7. Faraway, J., Augustin, N.: When small data beats big data. Stat. Probabil. Lett. **136**, 142–145 (2018)
8. Gal, I., Ginsburg, L.: The role of beliefs and attitudes in learning statistics: Towards an assessment framework. J. Stat. Educ. **2**, 2 (1994). DOI 10.1080/10691898.1994.11910471
9. Ioannidis, J.: The proposal to lower p value threshold to .005. J. Am. Med. Assoc. **319**, 1429–1430 (2018)
10. James, G.: Statistics within business in the era of big data. Stat. Probabil. Lett. **136**, 155–159 (2018)
11. Lin, M., Lucas Jr, H., Shmueli, G.: Too big to fail: Large samples and the p-value problem. Inform. Syst. Res. **24**, 906–917 (2013)
12. Meinshausen, N., Bühlmann, P.: Maximin effects in inhomogeneous large-scale data. Ann. Statist. **43**, 1801–1830 (2015)
13. Reid, N.: Statistical science in the world of big data. Stat. Probabil. Lett. **136**, 42–45 (2018)
14. SCImago: SJR–SCImago Journal & Country Rank. URL http://www.scimagojr.com. Last Accessed: 26 Jul 2018
15. Secchi, P.: On the role of statistics in the era of big data: A call for a debate. Stat. Probabil. Lett. **136**, 10–14 (2018)
16. Sherman, M.: Spatial Statistics and Spatio-Temporal Data: Covariance Functions and Directional Properties. John Wiley & Sons (2011)

17. Simmons, J.P., Nelson, L.D., Simonsohn, U.: False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol. Sci. **22**, 1359–1366 (2011)
18. Sterne, J.: Teaching hypothesis tests – time for significant change? Statist. Med. **21**, 985–994 (2002)
19. Wasserstein, R., Lazar, N.: The ASA's statement on p-values: Context, process, and purpose. Am. Stat. **70**, 129–133 (2016)
20. Wit, E.C.: Big data and biostatistics: The death of the asymptotic Valhalla. Stat. Probabil. Lett. **136**, 30–33 (2018)
21. Zieffler, A., Garfield, J., Alt, S., Dupuis, D., Holleque, K., Chang, B.: What does research suggest about the teaching and learning of introductory statistics at the college level? A review of the literature. J. Stat. Educ. **16**, 2 (2017). DOI 10.1080/10691898.2008.11889566

University[1]	MStats or similar	MDataSc or similar (Year commenced)[2]
Adel.	MBiostatistics	MDataSc (2018)
ANU	MStat	MAppDataAnalyt (2016)
Aust.Cath.	–	–
Bond	–	Under MBA (2017)
Canb.	–	MBusAnalyt (2015)
Carnegie Mellon (Aust.)	–	Under MIT (Unknown)
C.Qld.	–	MDataSc (2017)
C.Darwin	–	MDataSc (2018)
C.Sturt	–	–
Curtin	–	MPredAnalyt (2017)
Deakin	–	MDataAnalyt (2016)
Divinity	–	–
E.Cowan	–	–
FedUni	–	–
Flin.	–	–
Griff.	–	–
James Cook	–	MDataSc (2017)
La Trobe	MStat	MDataSc (2016)
Macq.	MStat	MDataSc (2018)
Melb.	Under MSc	MDataSc (2017)
Monash	MBiostatistics	MDataSc (2016)
Murd.	–	–
Newcastle(NSW)	MMedStat	Under MIT (Unknown)
Notre Dame Aust.	–	–
Qld.	MBiostatistics	MDataSc (2017)
Qld. UT	MBiostatistics	–
RMIT	MStat	MDataSc (2017)
S.Aust.	–	MDataSc (2018)
S.Cross	–	–
S.Qld.	Under MSc	Under MSc (2015)
Sunshine Coast	–	MInfCommunTech (2017)
Swinburne UT	MStat	Under MIT (2015)
Syd.	MBiostatistics	MDataSc (2016)
Tas.	–	Under MIT (2017)
Technol.Syd.	Under MSc	MDataScInn (2015)
Torrens (Aust.)	–	–
UNE	–	MDataSc (2018)
UNSW	MStat	Under MIT (2015)
Vic.(Melb)	–	–
W.Aust.	MMathStatSc	MDataSc (2017)
W'gong	MStat	Under MCompSc (2018)
W.Syd.	–	MDataSc (2016)

[1] Where available, the abbreviations are based on the list of Institution & Qualification Abbreviations from The University of Queensland. url: http://qual.app.uq.edu.au/institutions/. Last assessed: 25 July 2018.
[2] The year of commencement is based on the handbooks, new presses and/or brochures of the university.

Table 1. Master programs on statistics and data science offered by universities in Australia.

Subject Area	Journal	$n \leq 10,000$	$n > 10,000$
Agriculture	*Crop and Pasture Sci.*	5	3
Education	*Aust. Educ. Res.*	2	0
Health	*Rural Remote Health*	7	2
Medicine	*J. Diabetes Invest.*	23	2

Table 2. Number of articles with small ($n \leq 10000$) or large ($n > 10000$) sample sizes in the first issues of four journals in 2018.