THE ROLE OF SYNTHETIC DATA IN TEACHING AND LEARNING STATISTICS

Michael Bulmer and Leonard Coote
School of Mathematics & Physics and School of Business, University of Queensland, Australia
m.bulmer@uq.edu.au

The use of synthetic data is increasingly of interest to teachers and learners of statistics. Real world data has obvious advantages but limitations that may be less obvious. Though we live in a world of "big data," sourcing real world datasets that satisfy the needs of teachers may not be easy. In many situations, synthetic data can offer a viable alternative to real world data for teaching and learning statistics. Our basic premise is that datasets—real world and synthetic—should be relevant to learning objectives, offer students a degree of realism, and be sufficiently rich to encourage learning beyond the immediate context. We offer guidelines for generating synthetic datasets to ensure all three dimensions are satisfied: relevance, realism, and richness.

INTRODUCTION

The use of synthetic data is increasingly of interest to teachers of statistics, business analytics and data science. This paper argues for the use of synthetic data in teaching and provides guidelines for instructors in generating synthetic data. Synthetic data can vary in the extent to which it corresponds to real world processes. At one extreme, synthetic data can be completely artificial and abstract. At the other, it can reflect real world processes and patterns, and problems and contexts.

So-called "real world" data has advantages that may be obvious, and limitations that may be less obvious even to experienced statistics instructors. Real world data represents the world as it is and often addresses real contexts and problems from which students can learn. Achieving realism in teaching is often an important goal held by instructors. Thus, it is relatively easy to see real world data as a gold standard—but the use of real-world data may introduce complications into the teaching of statistics. Real world data may be difficult to access, of low quality, and of limited relevance to teaching objectives.

In many situations, synthetic data can offer a viable alternative to real world data for teaching statistics. This paper specifically attempts two contributions to the literature on teaching analytics/statistics. Firstly, the paper offers a framework for evaluating datasets—real and synthetic. Although we live in a world of "big data," sourcing real world datasets that satisfy the needs of teachers may not be easy. Teachers will benefit from having a clear articulation of the "pros" and "cons" in choosing between real world versus synthetic data and guidelines on the construction and use of synthetic data. The framework emphasizes the dimensions of *relevance*, *realism*, and *richness*.

Second, the paper uses this framework to provide guidelines for those seeking to generate synthetic datasets. Having control over data generation affords pedagogical opportunities that may further improve student outcomes.

MOTIVATING EXAMPLES

Real Data

There is always a cost for incorporating data in teaching, either in obtaining or using it—but this consideration should not outweigh other considerations. Nonetheless, before considering the specific nature of a dataset for use in a course, an instructor must assess the accessibility of the data. Accessibility refers to all the costs of accessing the data, including economic costs and non-economic costs (e.g., cognitive effort and time involved in developing a working knowledge of the data). An example, the Household, Income, and Labour Dynamics in Australia (HILDA) Survey illustrates some of the issues. Briefly, HILDA is a household-based panel study started in 2001 and surveying more than 17,000 households each year. It is a high-quality dataset, funded by the Australian Government, and is used for social and economic policy making. On the surface, the dataset has many attractive properties —the survey uses a wide range of questions reflecting actual behaviors and choices (on economic and social issues from labor market participation through to food choices) with formats that give different types of measurement (e.g., categorical, continuous).

Further, the dataset is available free to the user—but using the data for teaching statistics is not without its costs. Accessing the data requires formal permission from the Australian Data Archive,

In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the Gap: Empowering & Educating Today's Learners in Statistics. Proceedings of the 11th International Conference on Teaching Statistics (ICOTS11 2022), Rosario, Argentina.* International Association for Statistical Education. iase-web.org

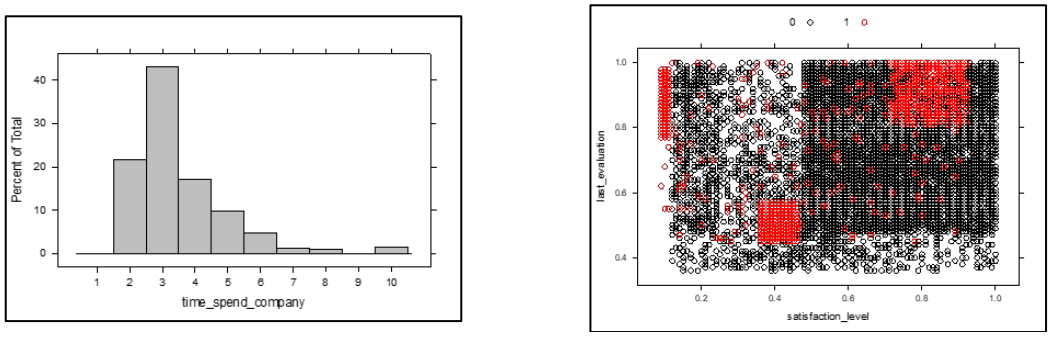
including the enactment of confidentiality agreements. Achieving these permissions can be a non-trivial task in a large class setting. Further, students themselves must come to terms with the data. This may be an important aspect of the task itself, and especially for a course on data wrangling where students are learning to manage, clean, and manipulate data. However, the time and effort required to understand the data may take away time from other tasks more directly relevant to the objectives of a course and, for some courses, the costs of accessing the data may outweigh the benefits of its attractive properties. One of the authors has stopped using HILDA to teach statistics because of the potential to misdirect student focus in learning. Of course, the HILDA example is not unique example of the costs and benefits of working with real world data—there are many such examples and in many such situations the costs may prompt an instructor to consider the use of synthetic data instead.

Synthetic Data

Sites like Kaggle host datasets from a wide range of contexts that can be downloaded to test algorithms and modelling techniques, and many of these datasets are synthetic. For example, there are several datasets related to Human Resource (HR) analytics, such as one uploaded by Benistant (2016). The Kaggle page for this dataset is no longer available but the original upload included the following description:

> Why are our best and most experienced employees leaving prematurely? Have fun with this database and try to predict which valuable employees will leave next. Fields include employee satisfaction level, last evaluation, number of projects, average monthly hours, whether the employee has left … This dataset is simulated. (Benistant, 2016)

The dataset records scores on ten variables for 14,999 employees. Figure 1(a) shows a simple histogram of one of the variables in this dataset, the time spent by an employee at the company. The distribution looks reasonable and is exactly what you might imagine for a variable like this. In contrast, Figure 1(b) shows a scatter plot of the relationship between an employee's score on their last company evaluation and their own satisfaction level with the company. It is immediately clear that there is something strange about the pattern this shows, with sharp rectangular boundaries at odds with what we might expect from such distributions.



(a) (b)

Figure 1. (a) Distribution of time spent at company and (b) relationship between last evaluation score and satisfaction level for employees who have left the company (red) or remained (black)

Based on Figure 1(b), it seems that this dataset would be ideal for teaching a course on classification methods, such as decision trees, with noise added to an otherwise simple pattern. Having a 'correct' answer like this would be very useful, allowing students to know that they had been successful. However, if the learning objective was focused on correlation or linear regression, then the simple nature of the relationship seems less desirable.

The dataset has also been re-uploaded by other users but many of these uploads omit the statement that "this dataset is simulated" and analyses of the data on other sites also treat the data as if it were real: "We have a real-world dataset which is out of an HR database, it contains employees leaving the company" (Borbon, 2017). In an era of concern about 'fake news,' the proliferation of 'fake' data in this way could be alarming. However, it also provides a valuable teaching opportunity in helping students identify such data.

DESIGN FRAMEWORK FOR SYNTHETIC DATA

Based on these examples and our own thinking, we proposed the following framework to consider when choosing data for a learning activity:

Relevance

Does the data have the properties (statistical and otherwise) needed to address the learning objectives of the course? Different objectives will demand data with different properties. For example, a course on correlation and regression would require a dataset with linear relationships as a prerequisite for addressing the learning objectives. A course on data wrangling might call for data with different properties than a course on data visualization.

Realism

Does the data provide insight into a "real world" problem? Aside from the learning objectives, the instructor may have the goal of the students leaning about the world (i.e., to solve a problem—however specific—graduates might encounter themselves). The data itself contains relationships that are representative of relationships found in real world settings. Without realism, specific results from analysis are meaningless.

Richness

Does the data allow learning beyond the immediate context? In using the data, students may develop general knowledge and skills that they can apply to other contexts, problems, and times (including as graduates). A rich dataset can be revisited in other contexts or courses, and enables innovative thinking, going beyond a specific learning objective.

The Framework Applied

The decisions in this framework can be captured by considering separate, and often competing, scales for each of these three dimensions. The Titanic dataset is another example that has become a famous "real world" case study for use in teaching (Dawson, 1995). Figure 2 summarizes the profiles for the Titanic and HR examples for a hypothetical course on linear regression.

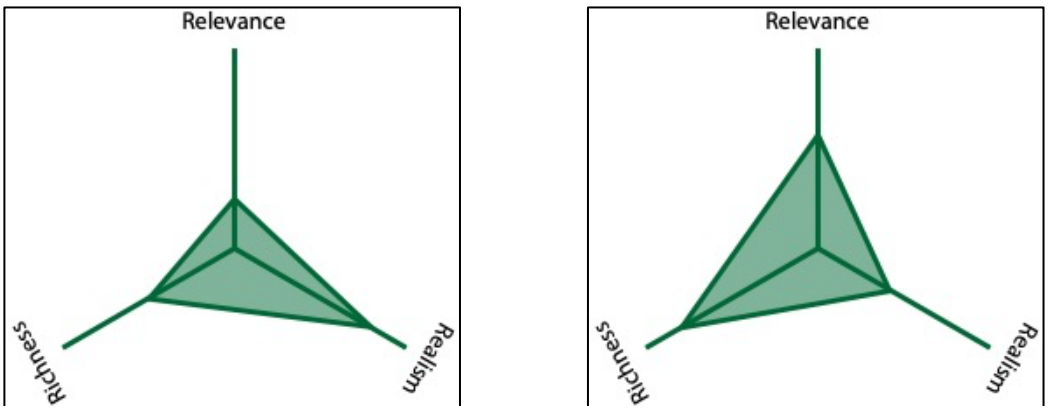


Figure 2. The framework for evaluating datasets applied to the Titanic (left) and HR (right) examples

GENERATING SYNTHETIC DATA

The above framework is a useful tool for a consumer of synthetic data, such as when looking for existing data from Kaggle or other repositories. However, it also informs design when producing new synthetic data. There is a long history of simulation methods for generating data from a range of statistical models (such as Ross, 1990), and R has standard functions to help with this. For example, data for a multiple regression study can be created by first generating values for explanatory variables from given distributions and then using a linear model with added noise to generate the response.

However, here we focus on identifying important pedagogical opportunities that come from having control over the generation process.

For example, an immediate benefit of generating synthetic data is that each learner can be provided with their own version of the data. This is a simple deterrent to plagiarism because each version would also have different answers, meaning that learners need to share processes and understandings rather than numerical results. It also offers deeper affordances because learners have an ownership of their data and a resulting sense of consequentiality, better aligning the incentive of grades with being able to successfully analyse their own data.

The design of synthetic data also needs to be aligned with the new directions of the discipline. Even in 1991, Halley notes that "poor response rates, coding errors and sampling errors are absent from simulated data; thus, students and instructors receive a ready supply of "clean" data for demonstrating statistical concepts and principles" (Halley, 1991, p. 518). This is still frequently desirable—very few examples in even the newest introductory textbooks contain missing values—but there is also a need to teach students how to deal with the deep messiness of real data. There are similarly times when it is useful for simulated data to have a clear 'right answer,' often tied to a particular solution technique, but generally we feel that the analysis of good synthetic datasets should not be too easy. Datasets need to be generated by simulations that can produce interesting multivariate relationships, emphasizing the 'Richness' dimension, so that students can gain experience with multivariate thinking.

CONCLUSION

The core idea of the diagrams of Figure 2 is that datasets should be relevant to learning objectives and offer realism and richness. The framework is useful for classifying existing data, and for focusing the development and design of new synthetic data. Synthetic datasets can be designed with all three properties in mind, following the guidelines outlined in the paper. We have had experience with real world data lacking in all three respects.

There is no single best dataset for everyone. Industry, students, and instructors may have different perspectives on what makes an appropriate dataset for learning statistics—with each elevating different dimensions of the diagram of Figure 2. Thinking like economists, we could remind ourselves that industry is the primary source of demand for our programs in statistics, business analytics and data science. Thus, let's start with an industry perspective on data for learning statistics. For example, industry may highlight *richness* above other dimensions. They may be seeking students who can solve a variety of problems beyond their immediate training. Students may place greater emphasis on *realism*. They may be driven by a desire to learn about and acquire the knowledge and skills needs to solve real world problems. Arguably, instructors place most importance on *relevance*—graduates have the skills that they say they have according to the leaning objectives of the courses and programs they complete. The issue of accessibility, which we think is crucial, needs to be carefully considered in conjunction with these other dimensions.

REFERENCES

Benistant, L. (2016). *Employee profile* [Data set]. HR Data for Analytics. Retrieved October 25, 2017 from https://www.kaggle.com/ludobenistant/hr-analytics [Retrieved July 11, 2022 from https://www.kaggle.com/datasets/jacksonchou/hr-data-for-analytics]

Borbon, B. (2017, September 27). *HR analytics with decision trees*. RPubs by RStudio. https://rpubs.com/brenborbs/312639

Dawson, R. J. (1995). The 'unusual episode' data revisited. *Journal of Statistics Education*, *3*(3). https://doi.org/10.1080/10691898.1995.11910499

Halley, F. S. (1991). Teaching social statistics with simulated data. *Teaching Sociology*, *19*(4), 518–525. https://doi.org/10.2307/1317899

Ross, S. M. (1990). *A course in simulation*. Macmillan.