BIG DATA IN THE INTRO STATS CLASS: USE OF THE AIRLINE DELAYS DATASET TO EXPOSE STUDENTS TO A REAL-WORLD, COMPLEX DATASET

Nicholas J. Horton[1], Benjamin S. Baumer[2] and Hadley Wickham[3]
1: Department of Mathematics, Amherst College, Amherst, MA USA
2: Department of Mathematics and Statistics, Smith College, Northampton, MA USA
3: Rice University and RStudio, USA
nhorton@amherst.edu

Students in the introductory statistics course need exposure to bigger datasets and more complex questions to be able to make sense of the increasingly data-centric world that they will inhabit. In this talk, I will describe how the airline delays dataset (150 million records on all commercial flights in the US from 1987 to 2012) can be integrated into the introductory statistics course. This large dataset is introduced early in the semester through a model eliciting activity due to Garfield and colleagues that leads students to undertake informal inference when comparing the on-time performance of two airlines servicing a pair of airports. Later in the course, students are able to assess the performance of their comparison rule by repeatedly sampling from the underlying population of flights, as well as visualizing the population using straightforward commands in R to access the database using simple SQL commands. The techniques are facilitated by use of R Markdown. In addition to providing insight into flight delays, the activity helps expose students to the power of statistics to make decisions in the face of uncertainty.

INTRODUCTION:

In a world awash in data, there is a pressing need for people who are able to extract actionable information from data. As statistics educators, we play a key role in helping to prepare the next generation of statisticians and data scientists. Gould (2010) cautioned that many of our courses do not answer the statistical questions students want to address.

Nolan and Temple Lang (2010) stress the importance of knowledge of information technologies, along with the ability to work with large datasets. Relational databases, first popularized in the 1970's, provide fast and efficient access to terabyte-sized datasets (Tahaghoghi and Williams, 2006). Connections between general-purpose statistics packages such as R and database systems can be facilitated through use of SQL (structured query language). Such interfaces are attractive as they allow the exploration of large datasets that would be impractical to analyze using general purpose statistical packages (Ripley, 2001).

The use of SQL within R is straightforward once the database has been created. An add-on package (such as RMySQL or RPostgreSQL) must be installed and loaded, then a connection made to a local or remote database. This makes it possible to realistically analyse a large dataset stored in a database in an introductory course. Students can use this setup to address questions that they find real and relevant (Gould, 2010), such as airline delays. It is not hard to find motivation for investigating patterns of flight delays. Ask students: have you ever been stuck in an airport because your flight was delayed or cancelled and wondered if you could have predicted it if you'd had more data? This dataset, which contains nearly 150,000,000 observations corresponding to each commercial airline flight in the United States between 1987 and 2012, was utilized in the ASA Data Expo 2009 (Wickham, JCGS, 2011). The ASA Data Expo 2009 website (http://stat-computing.org/dataexpo/2009) provides full details regarding how to download the Expo data (1.6 gigabytes compressed, 12 gigabytes uncompressed through 2008), set up a database, add indexing, and then access it from within R and RStudio.

This opportunity to make a complex and interesting dataset accessible to students in introductory statistics is quite compelling. In the first course, this was introduced through use of the "Judging Airlines" model eliciting activity (MEA) documented by the CATALST Group (2009). This MEA requires no technology, but guides students to develop ideas regarding center and variability using small samples of data for pairs of airlines flying out of Chicago (see Figure 1).

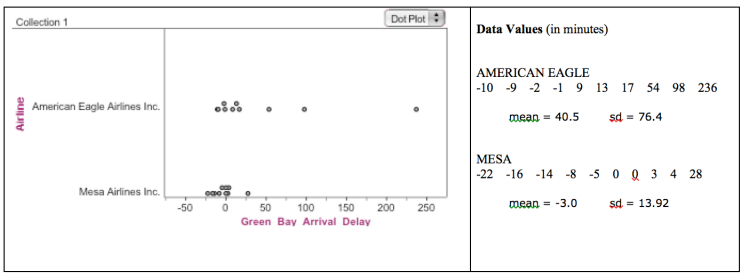

Figure 1: Sample data from two airlines flying from Chicago to Green Bay (with graphical depiction and summary statistics). Students are asked to develop rules to determine which airline is more reliable.

Later in the course, students return to the informal "rule" they developed in an extension to determine whether to make the call about one airline being more reliable than the other. Their rule can be

automated, and then carried out on a series of random samples from the flights from that city on that airline within that year. This allows them to see how often their rule picked an airline as being more reliable. Finally, students can summarize the population of all flights, as a way to better understand sampling variability. This process reflects the process followed by analysts working with big data: sampling is used to generate hypotheses that are then tested against the complete dataset.

The computation for the comparison of their informal "rule" and analyses of the distribution of the population values requires some coding. It would not be feasible to have students run these commands without some support. The provision of an instructor-provided R Markdown template (Baumer et al, 2014) facilitates the use of R for this purpose.

In a second course, more time is available to develop diverse statistical skills. This includes more sophisticated data management and manipulation, such as the calculation of the weekly count of flights over this period (reprising the display from Wickham (2009)) with additional years of data. This can be undertaken with a single SQL SELECT statement and some modest post-processing in R. Figure 2 displays the pattern, which has many interesting aspects.

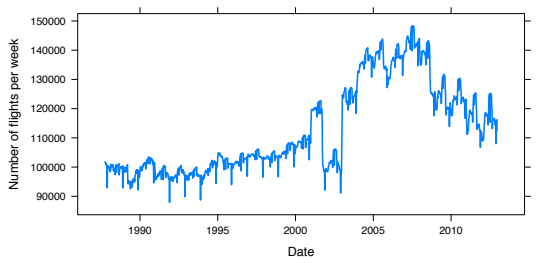

Figure 2: Display of weekly counts of commercial flights in the United States over the period 1987 to 2012 (total n=148,562,493). There is a strong seasonal pattern and clear impact of 9/11.

Other data wrangling and manipulation capacities can be introduced and developed using this example, including data joins/merges (since there are tables providing additional (meta)data about planes and airports). Use of a database to access this rich dataset helps to excite students about the power of statistics as well as introduce tools that can help energize the next generation of data scientists.

CONCLUSION

Nolan and Temple Lang argue that students need the facility to express statistical computations. In addition, there have been other calls for an increased use of computing in the statistics curriculum at the undergraduate level (American Statistical Association, 2000). In an era of increasingly big data, we agree that this is an imperative to develop in students, beginning with the introductory course. Some in the data science world argue that statistics is only relevant for "small data" and "traditional tools." We believe that the integration of these precursors to data science into our curricula—early and often—will help statisticians be part of the dialogue regarding Big Data and Big Questions (Davidian, 2013).

This work was partially supported by Project MOSAIC, US NSF (DUE-0920350). Examples and more information can be found at http://www.amherst.edu/~nhorton/airlines.

REFERENCES

American Statistical Association (2000). Curriculum guidelines for undergraduate programs in statistical science, http://www.amstat.org/education/curriculumguidelines.cfm.

Baumer, B., et al. (2014) R Markdown: Integrating a reproducible analysis tool into introductory statistics, conditionally accepted, *Technology Innovations in Statistics Education.*

CATALST Group (2009), Judging airlines Model Eliciting Activity (SERC), http://serc.carleton.edu/sp/library/mea/examples/example5.html.

Cobb, G. W. (2007). The introductory statistics course: a Ptolemaic curriculum?, TISE 1(1), http://www.escholarship.org/uc/item/6hb3k0nz.

Davidian, M. (2013). Aren't we data science? *Amstat News*, July 1, http://magazine.amstat.org/blog/2013/07/01/datascience/.

Gould, R. (2010). Statistics and the modern student. *ISR*, 78(2):297-315.

Nolan, D. & Temple Lang, D. (2010), Computing in the statistics curricula, *The American Statistician*, 64, 97–107.

Ripley, B.D. (2001). Using databases with R. *R News*, 1(1), 18-20.

Tahaghoghi, S.M.M & Williams H. E. (2006). *Learning MySQL*. Sebastopol, CA: O'Reilly Media.

Wickham, H (2009). ASA 2009 Data Expo, *JCGS*. 20(2):281-283.