IT IS TIME TO INCLUDE DATA MANAGEMENT IN INTRODUCTORY STATISTICS

Robert H. Carver[1] and Mia Stephens[2]
[1]Stonehill College & Brandeis University, Easton & Waltham, Massachusetts, USA
[2]SAS Institute, JMP Division, Cary NC USA
rcarver@stonehill.edu

There has been widespread adoption of real data sets and computational software in the teaching of introductory statistics. To sustain these two developments and to maintain currency with the explosion in freely available data from public sources, it is important for students to learn methods for obtaining, cleaning, organizing and manipulating large datasets from multiple sources prior to analysis. Though experimental studies remain central, the practice of data analytics in many disciplines begins with observational data. This paper begins with a rationale for including foundational concepts of data management such as joining tables, selecting rows, and inserting rows, as well as practice with automated approaches to missing and dirty data. It then goes on to provide illustrative examples of how such topics can be taught in engaging and accessible ways, and to suggest course topics that can be suppressed to open room for these topics.

INTRODUCTION

The triangular relationship of computing technology, real-world statistical applications, and statistics education is a dynamic one that calls upon us frequently to re-evaluate the scope, sequence, and content of the introductory statistics course. Each new generation of hardware and each new iteration of statistical software offer alternatives to traditional methods of statistical practice and education (Tukey, J.W., 1972; IBM SPSS Statistics, Version 21, 2012; JMP, Version 11, 2013). Across a wide variety of public and private sector organizations and disciplines, “Big Data” and “analytics” are being applied to numerous purposes (Economist, The, 2010; Manyika, J. *et al.*, 2012).

Within the statistics education community, calls for reform and adjustment to changing realities are a long-standing tradition. In recent years there has been growing attention given to the importance of using real data, culminating in the GAISE report (Aliaga, *et. al*, 2005) and in the emphasis of computer-intensive randomization methods rather than exclusive reliance on distribution-based approximations (Cobb, G.W., 2007). More recently still we find arguments that, in response to the big data explosion, further changes are needed (Gould, R., 2010; Horton, N.J., *et al.*, 2013). In this paper we take up one of Gould’s (2010) recommendations that we teach about databases, specifically about ways in which statisticians “can create very different datasets from the same database, depending on which subsets are selected, how variables are perhaps combined, and how categories are created or merge from existing variables.” We enlarge the argument some database concepts should be part of the introductory course and suggest practical steps toward successfully revising a course to include such concepts. .

DATA MANAGEMENT AS A FOUNDATIONAL AREA

One of the fortunate phenomena of current times is the widespread availability of large, reliable web-accessible databases providing data on a huge variety of subjects. Many of these databases are assembled and maintained by international agencies and governmental entities within nations. Organizations including the United Nations, The World Bank, the Energy Information Agency all have user-friendly web portals to both time-series and cross-sectional data suitable for use in an introductory course. Moreover, depending on the client disciplines serviced by the introductory course, it may be equally valuable to learn how to navigate the front end of a public database as it is to design and conduct surveys or to carry out experiments.

Acquiring a Data Set

That said, it should be noted that the user interfaces of such public websites vary widely. Users who have some grounding in essential concepts of data structure, metadata, and data queries have an advantage in effectively navigating portals to obtain data of interest for a given study. At a minimum a student venturing into, for example, the UNdata Explorer page (United Nations

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org [© 2014 ISI/IASE]

Statistics Division, 2014). The website connects users to data from diverse U.N. offices and international agencies. Choices are presented in a user-friendly tree structure with informative links as well as a topical search function. Links take users to metadata, footnotes, websites of originating agencies and so on. To the uninitiated visitor the number of choices may be overwhelming. The sense of overload could be mitigated with a rudimentary understanding of the logic of a database query: the user wants to select particular rows (often using a *filter*) corresponding to key fields with certain attributes, and specify particular columns or data series.

The user's analysis plan should inform choices about the layout of the data, particularly with longitudinal data. A given database will default to stacked or unstacked format, but the user should ideally specify the desired layout prior to the data export; this needs to happen at some point in the research process; if we wish to send undergraduates into the jungle of massive, real-time, real-world data, we need to equip them to tame the beast.

Between the Download and Analysis

After obtaining the data in a suitable form—often an Excel workbook—the user is likely to find the typical inconveniences: missing cells, cells that contain footnote references, columns whose meanings or units are ambiguous, data rows that represent subtotals or totals, or other horrors. Prior to analysis, the data need to be cleaned. This is not a new discovery, and a thorough treatment of the issues of missing data goes beyond the scope of an introductory course (Hoyle, 1971; Rubin, 1976). Nonetheless, introductory-level students need to be aware that observational data rarely arrive in complete form ready for analysis. They also need to appreciate the risks that can arise from using misunderstood variables in an analysis. The extensive documentation and metadata associated with public databases provide fertile ground to direct experience with researching the units, meanings, and data collection methods of variables obtained from the web.

If one source provides all of the needed data for the project, the student might begin the analysis at this stage. On the other hand, if the variables of interest reside in different databases, the student will confront the need to *join* data from two or more sources and to reconcile differences in the labeling and structure of the two source databases. With a small data set, this could be accomplished with a careful copy-and-paste operation. However with real data, it's often a more complex task, as illustrated in the next section of this paper.

Before stepping through an example, we note that many users will conduct their analyses using statistical software other than Excel. Standard packages such as SPSS, SAS, Minitab, JMP and R all make it quite easy for a new user to import data form Excel or open Excel files directly. Despite the well-documented obstacles with using out-of-the-box Excel for statistical analysis, Excel is an excellent platform for cleaning and preparing data for analysis. However, when it comes to joining multiple tables for a study, Excel falls short. In contrast, the statistical packages just cited all anticipate the need to merge tables and provide interfaces to make the task accessible even to introductory students. In the following example, we'll use the JMP (JMP, Version 11, 2013) to demonstrate an approach that accomplishes the goal without overloading the course schedule.

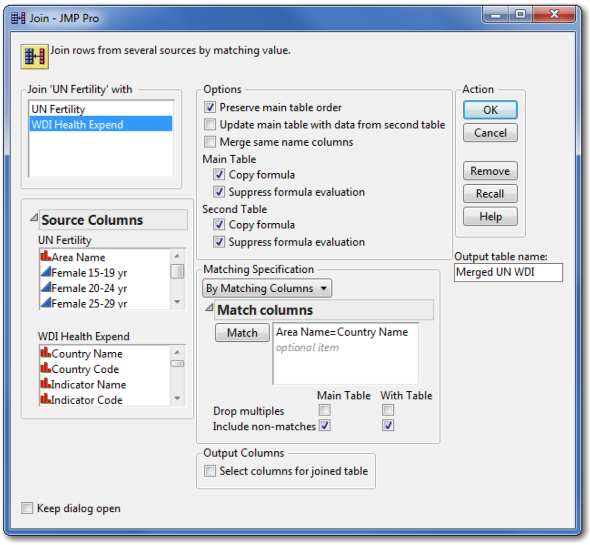
AN EXAMPLE

When conducting an observational study, it is common to assemble data from multiple sources into a single data table for analysis. Consider a study using data from, for example, the UN Statistics Division's Gender Info database and The World Bank's World Development Indicators (WDI). Specifically, we have per capita health expenditures (current USD) from the World Bank, and Female Fertility Rates for different age groups from the UN. The UN download lists 195 distinct countries with available data from the series of interest; the WDI lists 217 countries. In the UN table, the column listing countries is titled "Area Name" and in the WDI table it is "Country Name."

Moreover, some nations are identified by slightly different names in the two tables. For example, the UN lists "Hong Kong SAR of China" and the WDI lists "Hong Kong SAR, China." The UN table includes "Bolivia (Plurinational State of)"; the WDI data lists "Bolivia." A human reader would readily conclude that the two refer to the same country, but that conclusion is not necessarily obvious to software. For these reasons alone, the copy and paste approach would require considerable care to insure accurately aligning countries and data values. With approximately 200 rows the task would be fussy; with 20,000 rows it would be unrealistic to carry off. Confronting such a task, students would benefit from the few database management concepts cited above.

The essential concept is *key fields*, a simple idea that is traditionally not covered in introductory statistics. The key field, in a database, is the field that links a one type of record to another. At the introductory level, it is not necessary to go deeply into different types of keys, even into different types of table joins. As a first exposure to the world of databases, it is probably sufficient to comprehend that keys allow for accurate, automated merging of data from multiple sources.

Among the table functions provided in JMP, there is a Join command that opens the dialog shown below. It lists several options and customizations, among which is the option to identify the shared key column. Though it does not use the database terminology of inner and outer joins, check boxes in the Matching Specification panel allow for these operations. At a minimum, the user only needs to specify the tables to merge and the columns to treat as keys—well within the reach of a novice student.


Join - JMP Pro

Join rows from several sources by matching value.

Join 'UN Fertility' with
UN Fertility
WDI Health Expend

Source Columns
UN Fertility
Area Name
Female 15-19 yr
Female 20-24 yr
Female 25-29 yr

WDI Health Expend
Country Name
Country Code
Indicator Name
Indicator Code

Options
☑ Preserve main table order
☐ Update main table with data from second table
☐ Merge same name columns
Main Table
☑ Copy formula
☑ Suppress formula evaluation
Second Table
☑ Copy formula
☑ Suppress formula evaluation

Matching Specification
By Matching Columns
Match columns
Area Name=Country Name
optional item

	Main Table	With Table
Drop multiples	☐	☐
Include non-matches	☑	☑

Output Columns
☐ Select columns for joined table

Output table name: Merged UN WDI

☐ Keep dialog open

Joining the tables in this way results in a new table with 252 rows—making it clear which rows do not match in the two tables, either because of differences in naming or missing data values. The user can then make judgments about further editing country names for consistency.

At this point, the student is nearly ready to proceed with the intended analysis. After the join, further data quality inspection (automated or manual) is critical. The exercise of locating suitable data, successfully downloading the relevant series, initially cleaning it, joining two tables and finally reconciling mismatched observations would require thought, judgment, and time. Where in an introductory statistics course might an instructor find the space to add the instruction necessary to carry out an exercise like this?

MAKING ROOM FOR THESE TOPICS IN THE COURSE

If we limit the additional concepts and skills to the few cited earlier, a short reading assignment, combined with perhaps 60 minutes of class time and an out-of-class hands-on assignment should suffice. Assuming that most instructors already find the course dense with objectives, it remains only to figure out what can be jettisoned to make space. Ultimately the decision should depend on the nature of the course and the disciplinary needs of students, but we might suggest several topics to consider. If the course now covers use of printed probability tables and also uses software, the tables can easily go. If the course covers computation of several forms of confidence intervals or *t*-tests, dropping some of this coverage could free the needed space If the instructor currently delivers lectures on expository topics that are better left for homework or active discovery, the time might be found there. In that same vein, the data management skills cited in this article can all become part of a “flipped classroom” strategy (Bergmann, 2012).

CONCLUSION

We have concurred with the argument that current developments in freely available, large-scale databases elevate the need to view some concepts of data management as within the scope of the introductory statistics course. As the applications of big data analytics become more widespread and software becomes increasingly capable of seamlessly handling additional types of data, we will need continuously recalibrate the scope of introductory statistics. Among the many changes on the horizon, this paper has focused on a small set of data management tools and sketched an approach that can integrate data management into the introductory syllabus.

REFERENCES

Aliaga, M., Cobb, G., Cuff, C., Garfield, J., Gould, R., Lock, R., Moore, T., Rossman, A., Stephenson, R., Utts, J., Velleman, P., & Witmer, J. (2005). *Guidelines for assessment and instruction in statistics education: College report.* American Statistical Association: Alexandria VA. Retrieved December 22, 2013 from http//www.amstat.org/education/gaise/.

Bergmann, J., & Sams, A. (2012). *Flip Your Classroom.* Washington DC: International Society for Technology in Education.

Cobb, G.W. (2007). The introductory statistics course: A Ptolomaic Curriculum? *Technology Innovations in Statistics Education, 1*(1). www.escholarship.org/us/item/6hb3k0nz.

Economist, The (2010). Data Deluge: Special Issue, February 27, 2010. The Economist Newspapers Ltd.

Gould, R. (2010). Statistics and the modern student. *International Statistical Review, 78*(2), 297-315.

Horton, N. J., Baumer, B. S., Kaplan, D.T., & Pruim, R. (2013). *Precursors to the Data Explosion: Teaching how to compute with data*. Conference presentation at Joint Statistical Meetings, Montreal, Quebec. Abstract retrieved from http://www.amstat.org/meetings/jsm/2013/onlineprogram/AbstractDetails.cfm?abstractid=307064

Hoyle, M. H. (1971). Spoilt data--an introduction and bibliography. *Journal of the Royal Statistical Society. Series A (General), 134*(3), 429-439.

IBM SPSS Statistics Version 21. IBM Corporation, Armonk, NY, 1989–2012.

JMP, Version 11. SAS Institute Inc., Cary, NC, 1989–2013.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs,R., Roxburgh, C., & Byers, A.H. (2011). Big data: The next frontier for innovation, competition, and productivity. Retrieved August 27, 2011 from http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581-592.

The World Bank (2013). *World Bank Open Data.* Accessed December 7, 2013 from *http://data.worldbank.org/*

Tukey, J. W. (1972). *How Computing and Statistics Affect Each Other.* Paper presented at the Babbage Memorial Meeting: Report of Proceedings, London.

United Nations Statistics Division (2014). *UNdata.* Retrieved January 2, 2014 from *http://data.un.org/Explorer.aspx*