A SHINY NEW OPPORTUNITY FOR BIG DATA IN STATISTICS EDUCATION

Karsten Maurer
Department of Statistics, Iowa State University, USA
karstenm@iastate.edu

As the availability of truly massive data sets proliferates it is enticing to incorporate these data sources into the curriculum of an undergraduate statistics course. Major barriers exist for inclusion of big data due to the computationally intense nature of working with large databases. Difficulties include gaining access to the database, interacting with database management software and obtaining manageable subsamples from the database for student use. This paper describes a web based application, the Shiny Database Sampler, which allows instructors to bypass these barriers using a simple JavaScript based tool. The tool is constructed using R and the R packages Shiny and RMySQL to allow the instructor and/or students to sample observations from a number of different large databases, using selected sampling schemes, for use in the statistics classroom.

INTRODUCTION

Statistics education has been rapidly evolving in the past decade with respect to undergraduate course curriculum and assessment. Technology has played the role as a catalyst for many of these major changes. An important change involves how data is accessed and analyzed in the classroom. The GAISE report laid out six recommendations on how to improve the teaching of introductory statistics; two of which urge statistics instructors to "Use technology for developing conceptual understanding and analyzing data" and to "Use real data" (GAISE, 2005, p. 4). There are many software tools and online repositories for instructors to access real data for use in the statistics classroom; such as DASL (lib.stat.cmu.edu/DASL), CAUSE (www.causeweb.org) and Many Eyes (www-958.ibm.com). These technological tools are wonderful for accessing many real data sets but the majority of the data sets currently available are quite small in scale.

In his paper on graphics for large data, Unwin states that "(t)he definition of large in relation to data is always changing. A data set that required substantial high performance computing one year becomes easily analysable on a laptop a few years later" (Unwin, 1999, p. 129). What constitutes "small data" or "big data" is constantly being redefined in the field of statistics as computation allows us to collect, store and manipulate larger and larger data sets, but what is consistent is the desire to be able to analyze big data. Finzer, Erickson, Swenson and Litwin (2007) argue that in an introductory level statistics curriculum "(w)hat seems to us to be missing are data sets—especially large and highly multivariate data sets—that are ripe for exploration and conjecture driven by the students' intrigue, puzzlement and desire for discovery" (p. 1).

Exposing students to truly massive data is tricky because after a certain size, data is no longer easily transferred and stored to the student's personal computer. This necessitates the use of remote databases and database querying software in order for the students to interact with big data. This is no small task for either student or teacher in most undergraduate statistics courses. The Shiny Database Sampler tool was constructed to streamline this process of accessing data in large databases. It should be noted that the tool is not designed for the user to directly specify a query to the database but instead, as the name implies, allow for manageable subsamples from the large data bases to be obtained and downloaded.

The Shiny Database Sampler is a Javascript based online application created using the `Shiny` package in the R statistical computing language (RStudio Inc., 2013). The `Shiny` package uses specially structured R code files to generate the online graphical user interface that interacts with an R session running on the server. This was used in combination with the `RMySQL` package to allow the R session on the server machine to query the database at the users request via buttons on the graphical user interface (James & DebRoy, 2012).

In this paper we will describe the design of the Shiny Database Sampler tool which allows the user to take random samples from a database through a point-and-click online JavaScript interface. After describing the tool, a few brief examples of applying the tool to course activities will be outlined. This tool was constructed with both the student and instructor in mind, so examples of both student and instructor uses will be described.

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

INTERFACE LAYOUT AND FUNCTIONALITY

The Shiny Database Sampler allows the user to randomly sample subsets from remotely stored SQL databases using a point-and-click graphical user interface. The tool is available online through the link at shiny.stat.iastate.edu/karstenm/. A screenshot of the graphical user interface is shown in Figure 1 below. The interface is broken into two sections: a sidebar panel which contains all the sampling options and controls and a main panel which contains the data table and brief summary of the sampled dataset.

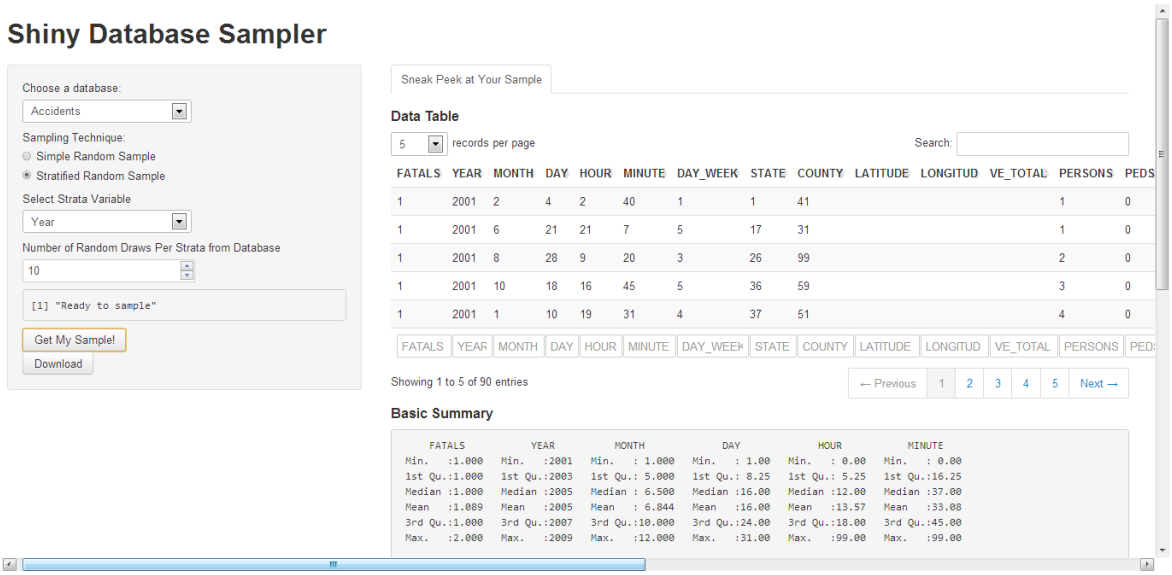

Figure 1: Screenshot of the Shiny Database Sampler tool

The sidebar panel contains several fields and buttons for selecting and executing a sampling plan. At the top of the sidebar is a dropdown menu to select the database table from which the user wants to take a random subset. The current version of the tool allows users to access workout data from an Iowa State fitness club called the RecMilers (www.recservices.iastate.edu/fitness/recmilers), the 2001-2009 Fatality Analysis Recording System accident data from the National Highway Traffic Safety Administration (www.nhtsa.gov/FARS) and the Public Use Micro Sample data from the 2000 United States Census (www.census.gov/). After choosing the database, the user can choose between taking a simple or stratified random subsample of data from the database. If the user chooses simple random sampling then all that remains is selecting a sample size; whereas if the user chooses a stratified random sample the stata variable and number of samples per stratum need to be specified. Once the sampling setup is ready, the user may click the “Get My Sample!” button and the randomly selected subsample of the database will be obtained and displayed in the main panel of the interface. Lastly, the side panel contains the button to download the selected subsample to a local drive on the user’s computer. The data will be downloaded as comma separated values (csv) file to the default download folder on the user’s computer.

The main panel of the Shiny Database Sampler interface displays a data table and a basic summary of each variable in the selected subsample. When logging into the webpage a default sample is taken and displayed until a sample of the users choosing is selected. The data table is searchable, sortable and expandable which makes it easy for the user to take a quick peek at the variable names and values that have been selected. The basic summary statistics for each variable are also displayed in the main panel below the data table; those familiar with R programming will quickly recognize this as the verbatim output of the `summary` function in R. The displays in the main panel of the Shiny Database Sampler are not intended to be the location for any extensive analysis of the sampled data but instead a quick check that the data that were sampled are what the user intended to select.

In these ways the layout is designed for the user to select actions in the side panel and view the results in the main panel. The interface allows students and teachers to obtain random subsets from existing databases without needing to learn a programming language.

APPLICATION EXAMPLES

With the understanding of what the tool is designed to do, instructors can likely think of several applications for the particular courses they teach. Below are three examples of possible applications for the tool in statistics education. The examples discussed below are by no means an exhaustive list of the teaching applications for the Shiny Database Sampler but are intended to exhibit its versatility as an education tool.

First is an example of the use as a teaching resource. Statistics instructors regularly include examples of graphical and analytical procedures in lecture materials, for which they are on a never-ending quest to find more data sources. The Shiny Database Sampler can be used alongside other online data repositories as yet another source of data for lecture, lab and homework materials. In particular it may be useful as a data source when teaching topics that pertain to large data; such as when teaching about over plotting in scatterplots when the number of observations becomes large. A course instructor can quickly log onto the webpage and obtain a subsample from an available database to be used for in class examples.

Secondly, the tool could be used to generate many similar, but unique, data sets for students to use in course projects. In lower level undergraduate courses, it is often ideal to have course projects utilize the same data set for all students, or groups of students, so that instruction and grading can be coherent and uniform. It has been argued that technology should be used to generate sufficiently different versions of the primary data set to minimize problems with plagiarism (Shutes, 2009). With the Shiny Database Sampler it would be relatively simple to generate a different random subsample from the same database for each student or group of students who are doing a class project. This provides the same variables and number of observations for each student or group of students. The benefit is that there is little chance of obtaining identical datasets, which thus necessitates analysis that is unique to the dataset and could not be copied directly from anyone else in the class.

The last example is to have the students access the Shiny Database Sampler themselves to obtain a dataset for a course project or lab assignment. An upper level statistics course likely has students that are quite capable of selecting a random subsample using the tool then export the csv file to then be imported to whichever statistical software program is being taught in class for analysis.

CONCLUSION

The Shiny Database Sampler was designed as tool for both students and instructors in undergraduate statistics to be able to have access to big data in the classroom. The tool allows the user to select a random subsample from an existing database using a flexible random sampling scheme within a point-and-click web based interface. The randomly selected subsample may then be viewed and exported to the user's computer for use in the statistics classroom.

Work is currently being done to improve the efficiency of the sampling and querying algorithms. Future work will include the expansion to more than three of database choices, running a case study to evaluate the use of the tool in a classroom setting and expanding the tool to contain options for graphically exploring selected samples.

REFERENCES

Finzer, W., Erickson, T., Swenson, K., & Litwin, M. (2007). On getting more and better data to the classroom. *Technology Innovations in Statistics Education, 1*(1).

GAISE (2005). *Guidelines for assessment and instruction in statistics education (GAISE) college report*. The American Statistical Association (ASA). Retrieved December 20, 2013 from http://www.amstat.org/education/gaise/GaiseCollege_Full.pdf

James, D., & DebRoy, S. (2012). *RMySQL: R interface to the MySQL database.* http://CRAN.R-project.org/package=RMySQL

R Development Core Team (2013). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

RStudio Inc. (2013). *shiny: Web application framework for R.* http://CRAN.R-project.org/package=shiny

Shutes, K. (2009). A note on using individualised data sets for statistics coursework. *Technology Innovations in Statistics Education*, *3*(4).

Unwin, A. (1999). Visualising large data sets. *Sistemi Complessi e Statistica Computazionale in Venice*, 129-136.