Contents

A `shiny` New Opportunity for Interaction with Big Data in Undergraduate Education

Karsten Maurer
Iowa State University, Ames, IA, USA

November 9, 2014

Abstract

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 interacting with 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 summary statistics or manageable subsamples from the database for student use. This paper describes a pair of web based applications, the Shiny Database Sampler and the Shiny Database Aggregator, which allows instructors to bypass these barriers using simple JavaScript based tools constructed using R and the R packages `shiny` and `RMySQL`. The Shiny Database Sampler allows instructors and/or students to obtain smaller subsamples from databases, using a variety of random sampling schemes. The Shiny Database Aggregator ...

1 Introduction

This is the intro section from the ICOTS paper, needs to be expanded

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" (Aliaga et al., 2005). There are many software tools and online repositories for instructors to access real data for use in the statistics classroom; such as DASL (DASL Project, 1996) , OzDASL (Smyth, 2011), Journal of Statistics Education Data Archives (American Statistical Association, 2014), CAUSE Web Repository (CAUSE, 2014) and Many Eyes (IBM Corp., 2013). 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 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" (Finzer et al., 2007, p. 1).

general outline of the intro should be something like

- : Motivation of what do we do?
- What has been done? - that's where the part of the lit review comes in

- Why is what we are doing relevant, i.e. where are the holes in the lit review - that can be mixed with the previous item
- Outline of the structure of the rest of the paper

2 Shiny Database Sampler

Exposing students to large data sources is tricky because after a certain size, it is unweildy to transfer, store and access data using 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 and Inc., 2014). 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 and DebRoy, 2012).

This section will begin by describing the design features of the Shiny Database Sampler tool, which allows users to take random samples from databases through a point-and-click online JavaScript interface. After describing interface, examples of how the tool has been integrating into course activities will be detailed. A user experience survey was conducted following a lab activity that used the Shiny Database Sampler. The survey results indicate that on average students found the application easy to use, found that the tool connected them to sampling concepts and felt moderately engaged with the census data that was accessed with the application.

2.1 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 main sections: a main panel and a sidebar panel. The main panel is for displaying information and the side panel is where the user specifies options; however these sections contain diffent options and displays depending on which tab of the application is selected.

When the "Sample and Summarize" tab is selected the Shiny Database Sampler will have the layout displayed in Figure 1. 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 under the "Sample and Summarize" tab displays a data table and a basic summary of each variable in the selected subsample. When first accessing the webpage, a default sample is taken from the Rec Milers database 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. 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.

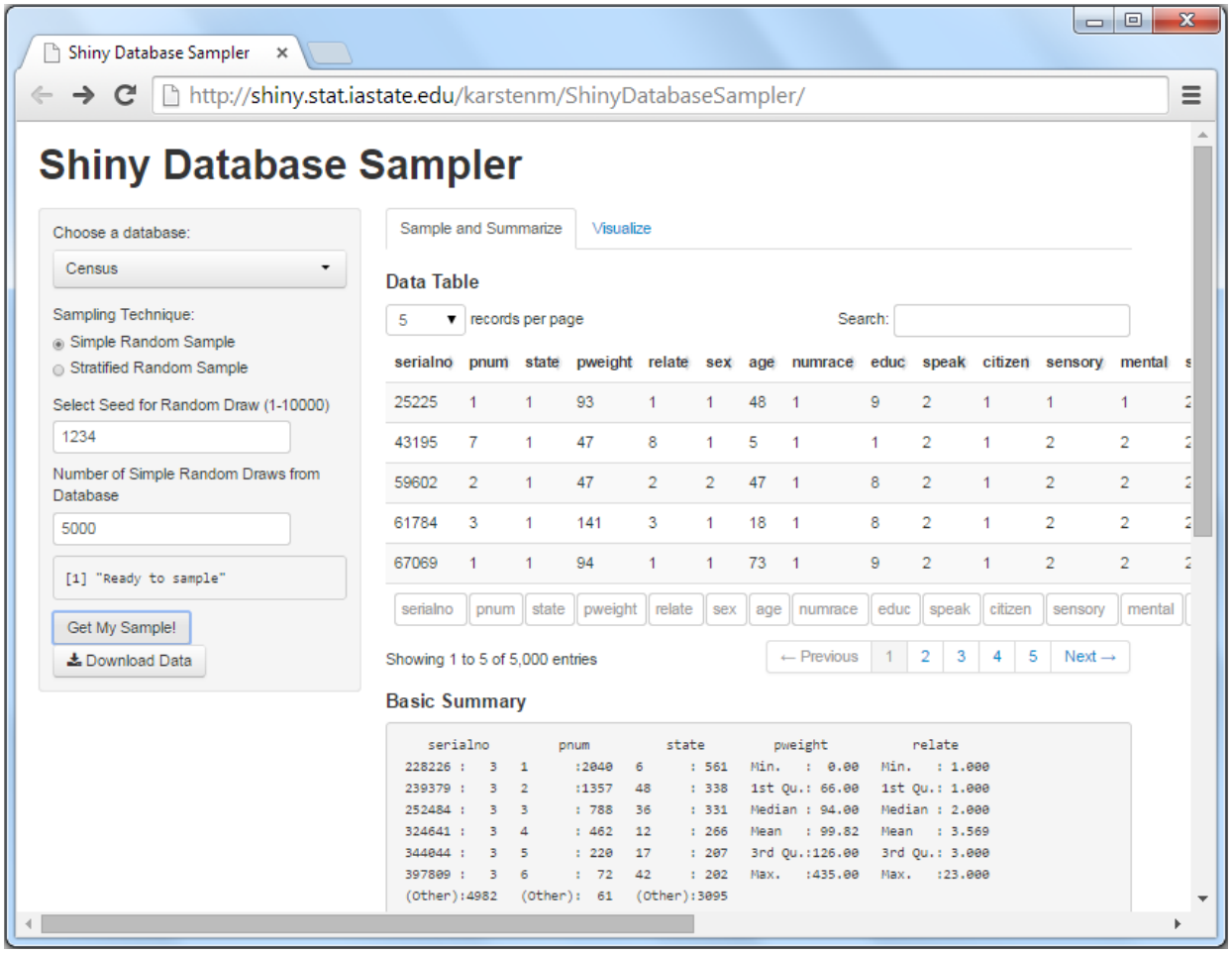

Figure 1: Shiny Database Sampler Layout for "Sample and Summarize" Tab

sampling options

- database: recmiler, accidents, census
- type: SRS, stratified
- sample size

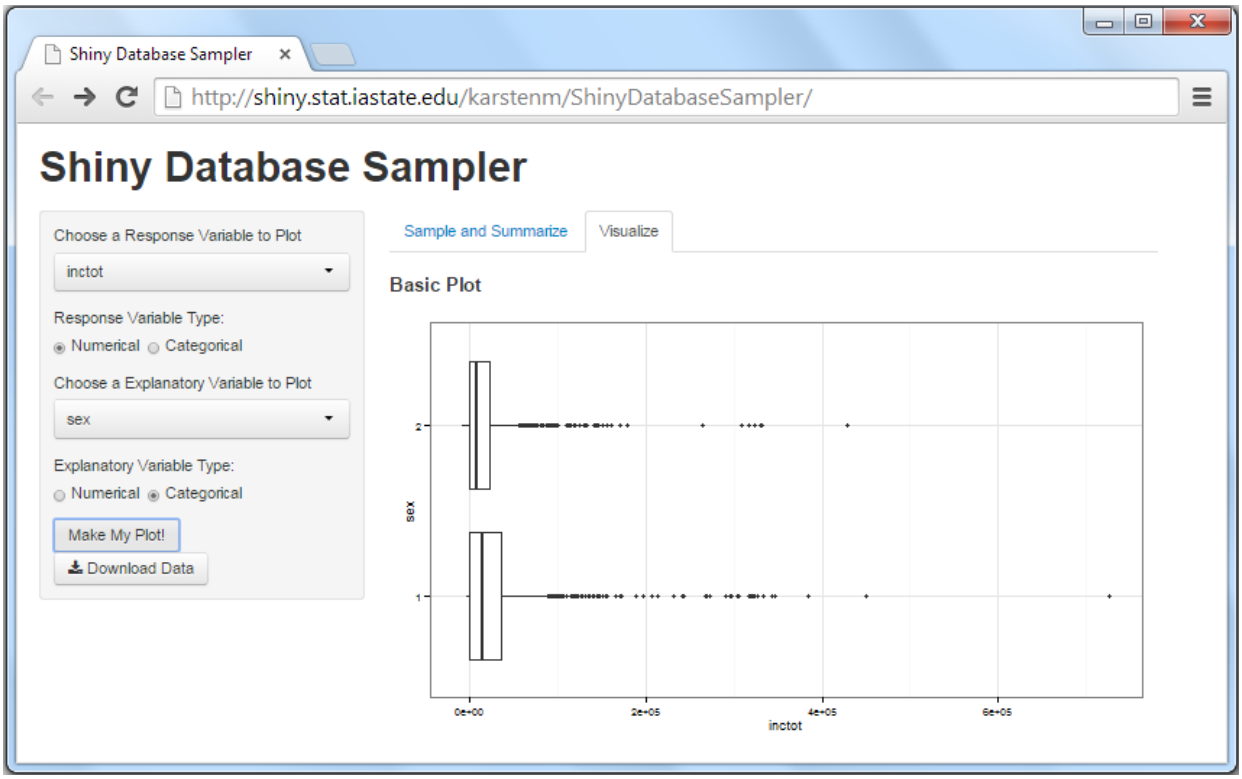

Figure 2: Shiny Database Sampler Layout for "Visualize" Tab

- seed: why seed available

data summaries

- table: sortable, searchable
- summary statistics: basic, broken down by strata if stratified sample taken
- plots: specify 1 or 2 variables and variable types to generate default plot types

2.2 Applications

- Used in introductory statistics course for lab and course project
- Project Use: students pick question then run mock survey using sampler tool
- Lab Use: description of lab (perhaps relocate paragraph 2 of next section here)

2.3 Student Response Survey Following Lab Application

As discussed above, the Shiny Database Sampler was designed for student use on course assignments. We were interested to learn student opinions about using the tool. Specifically we wanted know if students find the tool easy to operate, if they see the connection to sampling concepts and if they find the data engaging. These topics are selected to assess the quality of the Shiny Database Sampler with respect to the interface design and educational value. slow down - here we will need about one paragraph each for the topics. Start by a more formal definition of what you mean by the topic - give a reference - and how it helps with student

learning. **citation on HCI for why ease of use is important**. The GAISE guidelines recommend that technological tools should be used to help teach statistical concepts and that the use of real data is important for student engagement, hence we focus on these topics (Aliaga et al., 2005).

Student responses were collected in an anonymous survey following a group lab assignment that required students of Stat 104, Introduction to Statistics, at Iowa State University to use the Shiny Database Sampler tool. Six sections of Stat 104 students were surveyed. The students were informed that the survey was not required and that no penalties or rewards were affiliated with its completion. Of the 320 students attending lab, 265 completed the survey.

2.3.1 Lab Overview

The lab that utilized the Shiny Database Sampler was designed for students to think critically about sampling approaches, then use the tool allowed to treat the large database as a population from which to obtain survey data. Students were asked to consider the following pair of hypothetical situations:

1. Suppose that our goal is to estimate the mean age of all US residents. Similar to polling organizations we have a budget that allows us to survey around 1000 people. To collect our sample we decide to take a simple random sample of 1040 US residents.

2. Suppose now that our goal has changed. Now we wish to investigate the association between age and state of residency. We want to compare the median ages for different states. We still have a budget that allows us to survey around 1040 people. To collect our sample we decide to take a stratified random sample of 20 residents from each state in the United States plus the District of Columbia and Puerto Rico.

XXX what you describe below is called 'good friction' when designing computer interfaces, because you need to intentionally slow users down sometimes, and get them to make important decisions first. I am fairly certain that there is a corresponding counter part in the educational literature. Look into literature that talks about how to get students engaged into material - that should be a similar concept to good friction. In each scenario students were asked to discuss the choice of sampling scheme, and in particular to identify potential problems. The students used the Shiny Database Sampler tool to obtain a sample from the database containing a 1% microsample of the 2010 U.S. Census, from which they estimated mean and median age of U.S. residents. This lab was written to ensure that sampling concepts were the primary focus, with the Shiny Database Sampler acting in a supporting role. In order to avoid (sporadic) clicking of buttons to obtain samples without ever stopping to consider why the sampling approach matters, we intentionally designed the assignment to invite students to carefully consider sampling options *before* using the tool. The entire lab assignment can be found in Appendix A.

2.3.2 Survey Description

After completing the lab assignment, students were asked to fill out a survey consisting of twelve statements (referred to as *items* in the following, see Table 1 for an overview). For each statement, students were asked for feedback on their level of agreement on a Likert scale from strongly disagree to strongly agree. The twelve items were designed to assess student opinion within three topics of four items each: ease of use, connection to sampling concepts, and engagement with the census data. We will refer to these as the Ease, Concept and Engagement item sets. For each group of four items, two were worded positively and two were worded negatively. Introducing negation with half of the items was done to reduce the response bias associated with *acquiescence*, the tendency to respond positively irrespective of the item content due (Furnham, 1986). Responses were scored as -2 (strongly disagree), -1 (disagree), 0 (Neutral), 1 (agree), 2 (strongly agree). Responses for negatively worded items were reverse-scored for the purposes of analysis,
From Table 1 we see that all response averages are positive after reverse-scoring. With the Ease items this indicates that students tend to find the tool relatively easy to operate. For frame of reference, we assume that students are comparing the difficulty of use with other educational technologies and webpages they have encountered in the past; in particular the `JMP` software used previously on their Stat 104 labs and homework.

Topic Set	ID	Item	Polarity	Mean	SD
Ease	1	*I found the web tool easy to use*	+	0.84	0.76
	2	*The layout of the web tool was intuitive*	+	0.63	0.74
	3	*Using the web tool was difficult*	−	0.77	0.88
	4	*Learning to use the web tool was hard*	−	0.81	0.85
Concept	1	*The web tool helped me understand sampling concepts*	+	0.80	0.78
	2	*I understand sampling ideas less after using the web tool*	−	0.83	1.01
	3	*Sampling techniques are clearer after using the web tool*	+	0.58	0.73
	4	*The web tool made me less sure how to randomly sample*	−	0.89	0.87
Engagement	1	*I did not enjoy working with the Census data*	−	0.38	1.01
	2	*I thought the Census data was boring*	−	0.23	0.97
	3	*Knowing that the Census data was from real people made it more interesting*	+	0.82	0.84
	4	*I liked analyzing the Census data*	+	0.28	0.86

Table 1: Survey questions and response summaries for all items *after* Reverse-Scoring (RS)

Students also tend to respond to Concept items in a manner that is affirmative that the tool connects them to sampling concepts. Students' responses are near to neutral for most items about engagement with the census data, with the exception of Engagement item 3. The phrasing of this question seems to have led students to reconsider their engagement level and led to a consistently more positive attitude.

2.3.3 Assessment of Internal Consistency for Item Topic Sets

The goal for this survey is to use the responses to sets of items to infer student opinions about the underlying topic of each set. It is reasonable to aggregate the responses over entire questions sets if we can show that items within each set are measuring the same latent topic. We use fluctuation diagrams and Cronbach's α (Cronbach, 1951) to assess this internal consistency.

A fluctuation diagram visually displays a contingency table of a pair of variables as the area of blocks on the bivariate grid of all possible response values. A diagonal heavy fluctuation diagram indicates strong agreement or *internal consistency* between responses of the two items. Figure 3 contains fluctuation diagrams for all item pairs within topic sets. We notice that most pairs of responses fall heavily along the diagonal and are primarily in the upper right of each diagram. This indicates that most items within sets have strong agreement and that the response values are generally neutral to positive for all items after reverse-scoring. For the item pairs in the Concept topic set we see that fluctuation diagrams have slightly larger off diagonal trends than items within the other two sets, which indicates a lower level of internal consistency for Concept items than in the other two topic sets.
Cronbach's α measures internal consistency between a set of responses by comparing the sum of individual variances to the variance of the sum of the responses. It is defined as follows

$$\alpha \cdot (K-1)/K = 1 - \sum_{i=1}^{K} \mathrm{Var}\left(Y_i\right) \Big/ \mathrm{Var}\left(\sum_{j=1}^{K} Y_j\right), \tag{1}$$

where Y_i denotes the response on the i^{th} survey item ($i = 1, ..., K$), and K is the number of survey items considered for internal consistency. Generally, $K = 4$ for the item sets of this survey. Cronbach's α reaches a maximal value of 1, if there is perfect agreement between items (i.e. all responses to the same item set are identical). In the case that items sets are independent, the internal consistency is measured as $\alpha = 0$. Cronbach's α is negative int the situation of consistent disagreement between responses and will approache negative infinity if there is perfect disagreement between items. See appendix C for details on the bounds for α. Nunnally and Bernstein (1978, p. 265) propose that an α of 0.7 or above should be considered as an indication of "modest reliability". George and Mallery (2003) provide the commonly used extended scale,

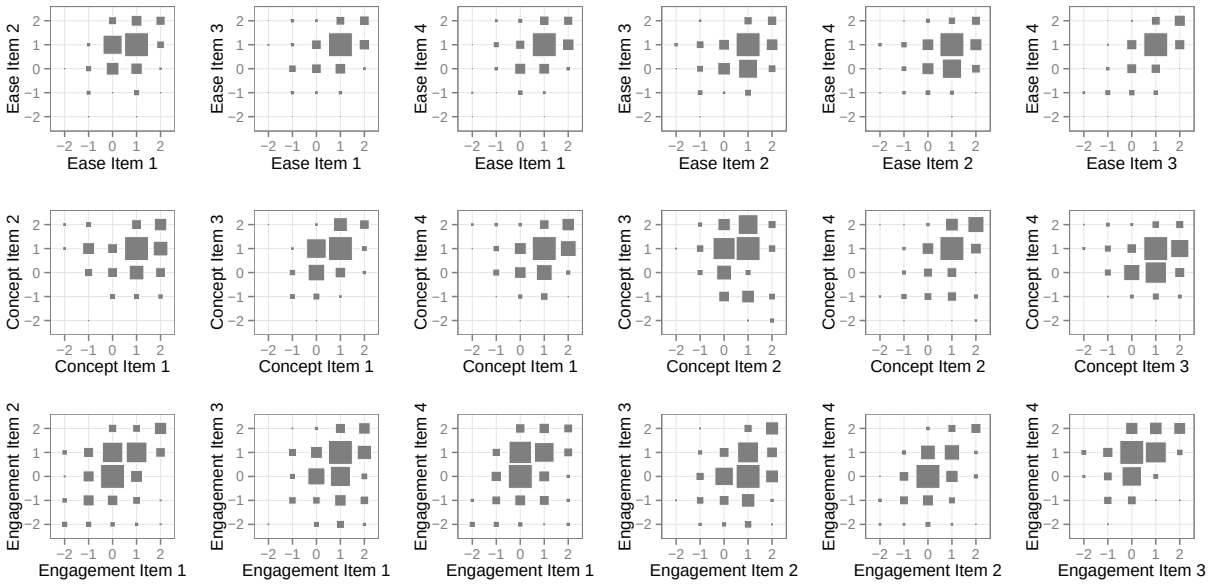

Figure 3: Fluctuation Diagrams of All Item Pairs within Topic Sets

displayed in Table 2, for interpreting internal consistency based on Cronbach's α.

Internal Consistency	Range
Excellent	$[0.9, 1.0]$
Good	$[0.8, 0.9)$
Acceptable	$[0.7, 0.8)$
Questionable	$[0.6, 0.7)$
Poor	$[0.5, 0.6)$
Unacceptable	$(-\infty, 0.5)$

Table 2: Extended Scale for Cronbach's α (George and Mallery, 2003).

'are available' is a bit vague. Could you re-phrase this and say what distribution α follows under normality of the item sets? Is this better? the construction of the degrees of freedom for the F distribution are quite convoluted, but I have tried to boil down the primary origins

Under the assumption of Gaussian data, the distribution of alpha is approximately F_{ν_1,ν_2} distributed, where $\nu_1 = n - 1$ and ν_2 is based on a function of the eigenvalues from the quadratic linear combination of the roots of the variance matrix. (Kistner and Muller, 2004). Thus distributionally based confidence intervals are available for Cronbach's α, but we are not entirely willing to assume Normally distributed responses to our survey and thus have elected to bootstrap the intervals instead.

Table 3 displays the point estimates and 95% central bootstrap intervals for Cronbach's α for each item set from the student survey. The intervals were created using quantiles of Cronbach's α values from each item set based on 10,000 bootstrap resamples. The results indicate modest levels of internal consistency for Ease and Engagement item sets, and a lower level for the Concept item set. This is in agreement with the findings based on the fluctuation diagrams in Figure 3.

Set	Estimate	95% Confidence Interval
Ease	0.70	(0.613 , 0.759)
Concept	0.53	(0.410 , 0.637)
Engagement	0.72	(0.643 , 0.776)

Table 3: Cronbach's α Estimates for each item set with 95% central confidence intervals based on 10000 bootstrap samples

2.3.4 Assessment of Polarity Issues

We next turn our attention to the polarity of the survey items; specifically we consider that positive and reverse-scored negative items may elicit a different responses.
The survey contained six unique item pairs based on topic and polarity combinations. Figure 4 displays compares the distribution of responses from positive and negative item pairs within topics. We see strong similarity between positive and reverse-scored negative items in response distributions with the Ease and Engagement item sets. The Concept item set however displays a noticeable difference in response distributions from each polarity. In particular, we see that students are more neutral toward the positively worded questions. This polarity difference in student responses may explain the lower internal consistency measured by Cronbach's α. This might be partly due to the problem that the negation of positive constructs may be linguistically counter-intuitive (Friborg et al., 2006). For instance, students may not interpret the statement "It is not less clear" as equivalent to the statement "It is more clear".

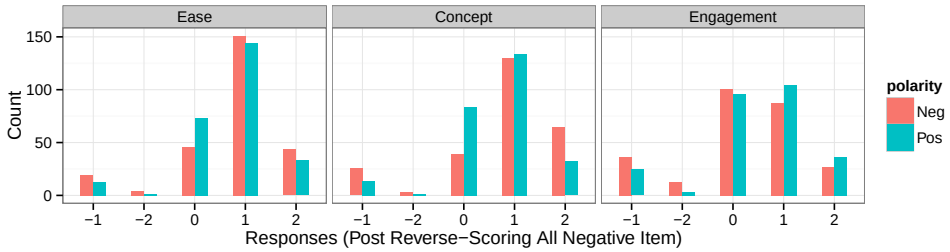

Figure 4: Item set response distributions

To assess whether responses from positive and reverse-scored negative items can be reasonably grouped together within topic sets we turn to principal component analysis. We decompose the item pairs averages for student responses from the six topic and polarity combinations. The component variances and factor loadings from this decomposition are found in Table 4. We argue that the data could be reasonably reduced to four principal components because each of these components explains over 10% of the variance and together they explain 87.4% of the total variation. The uniformly aligned factor loadings for Component 1 reflect the general tendency toward student agreement to all items on the survey. The factor loadings for Components 2 and 3 displayed in Figure 5 show similar projections for positive and negative item scores for Ease and Engagement pairs but a dramatic separation in the positive and negative item scores for the Concept set.

	Principal Component	1	2	3	4	5	6
Variances	Prop. of Var	0.457	0.181	0.135	0.101	0.071	0.055
	Cumu. Prop. of Var	0.457	0.638	0.773	0.874	0.945	1.000
Loadings	Pos. Ease	-0.292	0.169	-0.509	-0.036	-0.055	0.789
	Neg. Ease	-0.444	-0.390	-0.464	0.429	0.372	-0.335
	Pos. Concept	-0.304	0.330	-0.389	-0.350	-0.537	-0.487
	Neg. Concept	-0.408	-0.641	0.250	-0.585	-0.067	0.116
	Pos. Engaged	-0.367	0.523	0.207	-0.321	0.663	-0.083
	Neg. Engaged	-0.569	0.164	0.519	0.496	-0.355	0.087

Table 4: Summary Statistics from Principal Component Analysis with Six Topic/Polarity Item Pairs

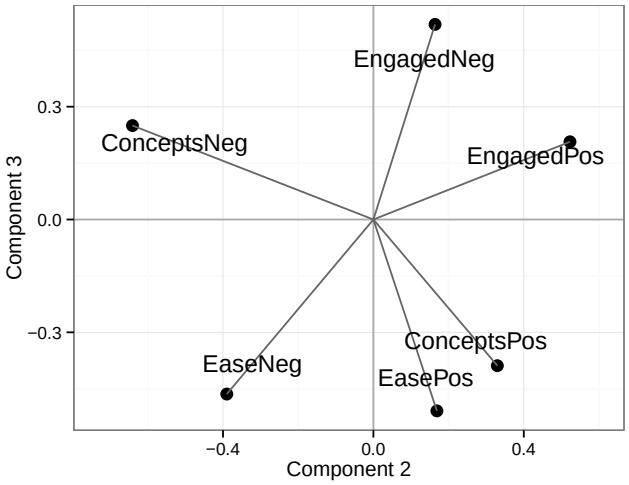

Figure 5: Item Pair Loadings on Components 2 and 3 from the Principal Component Analysis with Six Topic/Polarity Item Pairs

This principal component analysis, with all topic and polarity combinations, suggests that we can reduce the dimensionality by combining the positive items with the reverse-scored negative items for Ease and Engagement topics. This leaves only the Concept item set separated based on polarity for final analysis. The decision to combine the responses for Ease and Engagement items also aligns with the higher internal consistency for these item sets as displayed in Cronbach's α values and fluctuation diagrams in Figure 3. Thus, we will carry forward with the final analysis using four resulting item sets: Ease, Positive Concept, Negative Concept and Engagement.

2.3.5 Assessment of Orthogonality

The next major consideration is whether the item sets are truly measuring different latent topics and are not redundant. The ability of the survey to separately measure the topics of Ease, Concepts and Engagement can be assessed through the orthogonality of the responses from different item sets. To check the orthogonality of the sets we conduct another principal component analysis; this time on the average responses for each student from the four item sets – Ease, Positive Concept, Negative Concept and Engagement. Items sets will be considered highly orthogonal if the principle component analysis cannot reduce the dimensionality from the four sets.

	Principal Component	1	2	3	4
Variances	Prop. of Var	0.515	0.249	0.134	0.102
	Cumu. Prop. of Var	0.515	0.764	0.898	1.000
Loadings	Ease	-0.430	0.144	-0.307	0.837
	Pos. Concept	-0.408	0.562	-0.520	-0.497
	Neg. Concept	-0.619	-0.745	-0.101	-0.227
	Engagement	-0.515	0.330	0.790	-0.031

Table 5: Principal Component Analysis with Final Four Item Sets

Table 5 displays the proportion of variance explained by each of the four principal components and also the loadings for each item set mean that compose each component. The first principal component has similar loadings from all item sets, which we can interpret as the general tendency toward positively scored responses on all items. The second, third and fourth principal components create separation for mean responses of the Negative Concept item set, the Engagement item set and the Ease item sets, respectively. The variances in Table 5 reveal that over 10% of the variation is explained by the fourth component, thus it is necessary to retain all four principal components. This inability to reduce dimensionality implies that average student responses from the four item sets are largely orthogonal. Based on the separation in the loadings and the orthogonality of the principal components, we conclude that the average response scores from the four item set have interpretability as measurements of unique latent topics.

2.3.6 Survey Assessment Results

In the analysis of student responses, we found that the internal consistency, assessed with Cronbach's α and fluctuation diagrams, is acceptable for interpreting the combined item responses that measure of Ease of Use and Engagement with the census data. We did not have the same certainty with the measures of Concept and therefore split the Concept items into two sets: the Positive and Negative Concept item sets. This split is supported by the initial principal component analysis of the six topic and polarity item pair scores. The follow-up principal component analysis on the combined responses for each of the four resulting item sets indicated that the factors were all fairly orthogonal. This ensures us that the survey was effective at eliciting unique characteristics of the user experience.

The barcharts found in Figure 4 show that the distribution for each item set is heavily skewed to the left, with the majority of students having the neutral to positive responses. The small bump at the far left of each distribution indicates that there was a small minority of students that expressed negative views. The response distributions indicate that on average students found the application easy to use, found that the tool connected them to sampling concepts and felt moderately engaged with the census data that was accessed with the application.

3 Shiny Database Aggregator

Brief intro the functionality goals for the tool

3.1 Layout and Functionality

To be determined

3.2 Applications

To be determined

4 Conclusions and Future Work

References

Aliaga, M., Cobb, G., Cuff, C., Garfield, J., Gould, R., Lock, R., Moore, T., Rossman, A., Stephenson, B., Utts, J., Velleman, P., and Witmer, J. (2005), "Guidelines for Assessment and Instruction in Statistics Education: College Report," .

American Statistical Association (2014), "Journal of Statistics Education (JSE) Data Archives," `http://www.amstat.org/publications/jse/jse_data_archive.htm`, accessed: 09/03/2014.

CAUSE (2014), "Consortium for the Advancement of Undergraduate Statistics Education Resourses," `https://www.causeweb.org/resources/`, accessed: 09/03/2014.

Cronbach, L. (1951), "Coefficient alpha and the internal structure of tests," *Psychometrika*, 16, 297–334.

DASL Project (1996), "The Data and Story Library (DASL)," `http://lib.stat.cmu.edu/DASL/`, accessed: 09/03/2014.

Finzer, W., Erickson, T., Swenson, K., and Litwin, M. (2007), "On Getting More and Better Data into the Classroom," *Technology Innovations in Statistics Education*.

Friborg, O., Martinussen, M., and Rosenvinge, J. (2006), "Likert-based vs. semantic differential-based scoring of positive psychological constructs: A psychometric comparison of two versions of a scale measuring resilience," *Personality and Individual Differences*, 40, 873–884.

Furnham, A. (1986), "Response Bias, Social Desirability and Dissimulation," *Personality and Individual Differences*, 7, 385–499.

George, D. and Mallery, P. (2003), *SPSS for Windows Step by Step: A Simple Guide and Reference, 11.0 Update (4th Edition)*, Boston, MA: Allyn & Bacon.

IBM Corp. (2013), "Many Eyes," `www.ibm.com/manyeyes`, accessed: 09/03/2014.

James, D. A. and DebRoy, S. (2012), *RMySQL: R interface to the MySQL database*, r package version 0.9-3.

Kistner, E. and Muller, K. (2004), "Exant Distributions of Intraclass Correlation and Cronbach's Alpha with Gaussian Data and General Covariance," *Psychometrika*, 69, 459–474.

Nunnally, J. and Bernstein, I. (1978), *Psychometric Theory (Third Edition)*, New York, New York: McGraw Hill.

RStudio and Inc. (2014), *shiny: Web Application Framework for R*, r package version 0.9.1.

Smyth, G. (2011), "Australasian Data and Story Library (OzDASL)," `http://www.statsci.org/data`, accessed: 09/03/2014.

Unwin, A. (1999), "Visualising Large Data Sets," *Sistemi Complessi e Statistica Computazionale*.

A Appendix: Lab Assignment

For this activity you will be using a tool called the Shiny Database Sampler to take a random sample of United States residents from US census data. The census data is the Public Use Microdata Sample (PUMS) which is a 3 million person subset of the entire Census data. For this activity we treat our samples as though they are selected from the full census records.

We are going to explore how these random sampling plans relate to the goals of a sample survey. The tool will allow you to define either a simple random sampling plan or a stratified random sampling plan. In the following two scenarios we will explore the advantages and disadvantages of these two sampling plans. Access the tool at `http://shiny.stat.iastate.edu/karstenm/ShinyDatabaseSampler`.

Scenario 1: Suppose that our goal is to estimate the mean age of all US residents. Similar to polling organizations we have a budget that allows us to survey around 1000 people. To collect our sample we decide to take a simple random sample of 1040 US residents.

(a) Is this study and example of an experiment or an observational study? Explain your answer.

(b) Your colleague Bob claims that we are wasting our budget to get only 1040 people using random sampling. He says that we could get 20000 responses to the survey if we invested that money into a mailing campaign in Minneapolis. Explain why the random selection is important.

(c) Another colleague, Jill, asks why we do not stratify by state when we take the sample so that we get 20 people from each of the 50 states along with Puerto Rico and the District of Columbia. Explain why this idea would not create a representative sample to pursue our goal.

Now that we have decided on our sampling plan, let's go collect our data. The Shiny Database Sampler needs to be told 4 pieces of information in order to collect census records the way you want. (1) Choose the database called "Census", (2) select the "simple random sample" option, (3) enter a random seed, any number between 1 and 10000, you can do this by rolling a 10-sided die 4 times and (4) lastly tell it that we want "1040" random draws. Once you have drawn your samples the page will display basic summary statistics for the variables in the census.

(d) Report the 5-number summary and sample mean age.

(e) Use the 5-number summary to construct a box plot of age.

(f) Go to the "Basic Plots of Your Sample" tab. Choose age as your Response Variable to Plot. What type of variable is this? By clicking on Make My Plot? a histogram of the sample of ages will be displayed. Describe the shape of the data distribution of age.

(g) Is the relationship between the sample mean and sample median consistent with your description of shape? Explain briefly.

(h) If our goal was to not only estimate the mean age of all the U.S. residents but also come up with estimates of the median age of all residents in each of the 50 states, plus the District of Columbia and Puerto Rico what is a drawback of using the simple random sample of 1040? Hint: Set the Data Table to display 100 records per page and go to the page that has "states" 10 and 11 (Delaware and the District of Columbia).

Scenario 2: Suppose now that our goal has changed. Now we wish to investigate the association between age and state of residency. We want to compare the median ages for different states. We still have a budget that allows us to survey around 1040 people. To collect our sample we decide to take a stratified random sample of 20 residents from each state in the United States plus the District of Columbia and Puerto Rico.

(i) Explain in general why collecting a stratified random sample is a better plan than a simple random sample for answering this question.

Now that we have decided on our new sampling plan, let's go collect our data. The Shiny Database Sampler will need to be told 5 pieces of information in order to collect census records the way you want this time. (1) Choose the database called "Census", (2) select the "stratified random sample" option, (3) enter a random seed, any number between 1 and 10000, you can do this by rolling a 10-sided die 4 times, (4) select "state" as strata variable and (5) lastly tell it that we want "20" random draws from each state, plus the District of Columbia and Puerto Rico.

It will take a minute or two to collect these data. It is sifting through millions of records and randomly selecting them from within state groups after all! Once you have drawn your samples you can take a peek at your data set in the main panel of the webpage. You will be able to answer the following questions using the summaries provided on the webpage.

You will notice that the summaries are all broken down by state, but the states are not given names, they are given a code number. This is done on the census to save computer storage space (saving a "19" is much smaller than "Iowa"). A list of all the state codes is available at `https://www.census.gov/geo/reference/ansi\_statetables.html` (Click on FIPS Codes for the States and the District of Columbia).

(j) Report the mean and 5-number summary for the age of the sample from the state of Iowa (`state = 19`).

(k) Report the mean and 5-number summary for the age of the sample from the state of Alaska (`state = 2`).

(l) Compare the distribution of ages in Alaska and Iowa using the values from parts j and k.

(m) Making comparisons as we have done above would become tedious if we wanted to compare ages between all pairs of states in the country. What would be a good way to visually display this information so aid in making these comparisons? Explain your answer.

B Appendix: Database Descriptions

To be detailed when databases updated and origins better known.

C Appendix: Cronbach's α Properties

Recall the form of Cronbach's α from equation (1):

$$\alpha = (K/(K-1))\left(1 - \sum_{i=1}^{K} \text{Var}\,(Y_i) \Big/ \text{Var}\left(\sum_{j=1}^{K} Y_j\right)\right),$$

Claim 1: Perfect agreement in items leads to $\alpha = 1$

Proof: Let $Y = Y_1 = Y_2 = ... = Y_k$, thus having perfect agreement.

$\Rightarrow \text{Cov}(Y_i, Y_j) = \text{Var}\,(Y) = \sigma_y^2 \quad \forall i \neq j$

$\Rightarrow \text{Var}\left(\sum_{j=1}^{K} Y_j\right) = \sum_{i=1}^{K} \text{Var}\,(Y_i) + \sum_{i \neq j} \text{Cov}(Y_i, Y_j) = K\sigma_y^2 + K(K-1)\sigma_y^2$

$\Rightarrow \alpha = (K/(K-1))\left(1 - \sum_{i=1}^{K} \text{Var}\,(Y_i) \Big/ \text{Var}\left(\sum_{j=1}^{K} Y_j\right)\right) =$
$(K/(K-1))\left(1 - K\sigma_y^2 \Big/ K\sigma_y^2 + K(K-1)\sigma_y^2\right) =$
$(K/(K-1))\,(1 - 1/K) = (K/(K-1))\,((K-1)/K) = 1$

Claim 2: For independent items $\alpha = 0$

Proof: Let $Y_1 = Y_2 = ... = Y_k$ be independent
$\Rightarrow \sum_{i=1}^{K} \text{Var}(Y_i) = \text{Var}\left(\sum_{j=1}^{K} Y_j\right)$

$\Rightarrow \alpha = (K/(K-1))\left(1 - \sum_{i=1}^{K} \text{Var}(Y_i) \Big/ \text{Var}\left(\sum_{j=1}^{K} Y_j\right)\right) =$
$\alpha = (K/(K-1))\left(1 - \text{Var}\left(\sum_{j=1}^{K} Y_j\right) \Big/ \text{Var}\left(\sum_{j=1}^{K} Y_j\right)\right) =$
$\alpha = (K/(K-1))(1-1) = 0$

Claim 3: Perfect disagreement in items leads to $\alpha = -\infty$

Proof: Let $K = 2$ and $Y_1 = -Y_2$, thus having perfect disagreement.

$\Rightarrow \text{Var}(Y_1 + Y_2) = \text{Var}(Y_1 - Y_1) = \text{Var}(0) = 0$
$\Rightarrow \alpha = (K/(K-1))\left(1 - \sum_{i=1}^{K} \text{Var}(Y_i) \Big/ \text{Var}\left(\sum_{j=1}^{K} Y_j\right)\right) = (2/1)(1 - 2\sigma_y^2/0) = -\infty$