A Shiny New Opportunity for Big Data in Statistics Education

Random Sampling from Large Databases Using the Shiny Database Sampler Tool

Overview

- Data Tables
- Web Tool Design
- Lab Assignment Application
- Course Project Application
- Conclusions
- Future Work

Data Tables

- Data tables stored in a MySQL Database
  - RecMilers Fitness Group (exercise logs)
  - US Fatal Car Accidents (2001-2010)
  - Census Bureau information on US residents (2010 Public Use Micro Sample)
- Big, Real, Rich, Interesting Data: There is only one problem…
- How can undergraduate students access it?!

Web Tool Design

Three Design Goals

1. Easy accessibility using an online point-and-click environment
2. Allow students to take appropriately sized subsets from the databases to practice methodology learned in class
3. Emphasize the role of random sampling techniques as the mechanism for selecting subsets of data

shiny.stat.iastate.edu/karstenm/ShinyDBSampler

User Interface

Running Javascript in Web Browser

Shiny

RMySQL

Server Machine

Running R Session

Data Tables

Stored in MySQL Database

Web Tool Design

Demonstration

shiny.stat.iastate.edu/karstenm/ShinyDatabaseSampler

Lab Assignment Application:

- Group lab assignment
- Following lecture unit on random selection
- Presents hypothetical scenarios for students to design surveys
- Use the Shiny Database Sampler tool to conduct the surveys

Lab Assignment Application:
Surveying American Residents

Example Scenario:

- Goal is to investigate the association between age and income for all US residents.
- Have a budget that allows us to survey around 2000 people.
- Decide to take a simple random sample of 2000 US residents.

Lab Assignment Application:
Surveying American Residents

Example Problem:

- Your colleague Bob claims that we are wasting our budget to get only 2000 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 to Bob why the random selection is important.

Lab Assignment Application:
Surveying American Residents

Example Problems:

- Another colleague, Jill, asks why we do not stratify by state when we take the sample so that each state is well represented. Explain why this idea would not create a representative sample to pursue our goal.

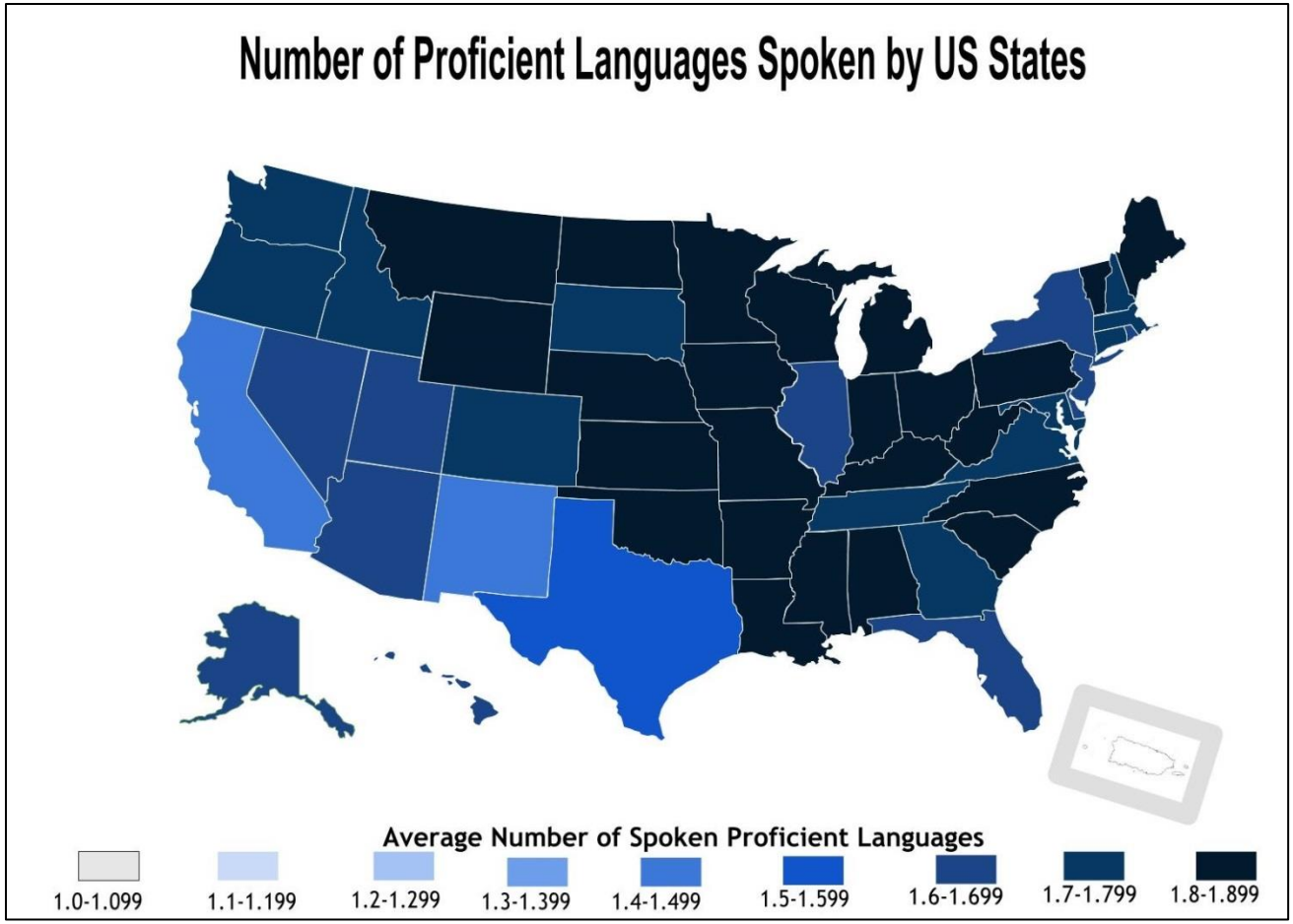
- Export the data, then use JMP to create a linear regression line. Interpret what the slope tells us about the association between age and income.

Course Project Application

- Group Project
- Requires students to collect data to answer a research question of their choosing
- Focus on bivariate associations
- Data tables on Shiny Database Sampler given as an option for data collection

Course Project Application

Example: One group investigated geographic association with linguistics using a state stratified sample of PUMS data



Conclusions

- Connects students to real, big data
- Ease of use with point-and-click environment
- Application speed declines with multiple users
- Provides students with rich data to explore and analyze for course projects
- Labs utilized deliberate questions aimed to motivate when to use certain sampling schemes

Future Work

- Survey students on user experience
- Provide access to new databases
- Add cluster sampling functionality
- Add plotting support for geographic data
- Server and Database adjustments to handle higher web traffic

Acknowledgment

- Dr. Heike Hofmann (major professor)
- Alan Kansakar (server support)
- Summer 2014 Group 2: Mariah, Dez, Mark and Will (student project example – linguistics map)

Questions

What questions do you have?