Workshop proposal

Frances Tong

2009-07-21

1 Introduction

We are heading into an age where we are being overwhelmed with publicly available data, easily obtainable from sites such as data.gov, google public data search, and various biological repositories, etc. However, the large majority of people, even if highly educated, is not equipped to deal with using such data in any easy and efficient manner.

While programs such as Gapminder and Google's public data plotting capabilities are a great way for the public to get started on viewing and interpreting data, in some ways they prevent students from ever having to learn how to deal with raw data on their own. Now is a good time for for all students, whether they are future medical doctors, sociologists, or lawyers, to be comfortable with jumping right in to find and process relevant data to see it the way they want and to use it to answer whatever burning questions or hypotheses they may have.

2 Problem and data

An important area on which to base this project is human rights. Statistical work in this field has been what past ASA President I. Richard Savage has called a hard-soft problem: soft as in difficult to define, and hard as in difficult to solve. This all the more shines light on the need for more qualified statisticians to be interested and working on such problems that will benefit humanity. From 2001, the Millenium Development Goals have highlighted eight pressing matters that the world should make progress on improving: 1) poverty and hunger; 2) universal primary education; 3) gender equality; 4) child mortality; 5) maternal health; 6) HIV/AIDs, malaria and other diseases; 7) environmental sustainability; and 8) global partnership for development. There is one very important issue that if prevented, can dramatically affect at least six of these goals: child marriage. Young girls who are forced into marriage to strangers before they turn 18 face a life of poverty, lack of education, domestic abuse, and risks of contracting HIV and early pregnancy complications that may result in the death of their babies and them. Accurate measurements are needed to determine which areas require the most help, what factors are important in prevention, and if progress is being made. Some data sources are available from UNICEF http://www.unicef.org/sowc09/statistics/statistics.php and the Millenium Development Goals site: http://mdgs.un.org/unsd/mdg/Data.aspx

3 Learning objectives

Learn how to:

- ask thoughtful questions
- search for relevant data
- clean and reformat raw data
- combine data from different sources

- recode factors
- deal with missing values or sparse datasets
- design a database
- load and retrieve data from database
- plot data in meaningful ways
- write a user friendly program to analyze data
- think about and analyze issues outside their daily lives

4 Target audience

undergraduate students and graduate students in other departments that use data and hopefully senior high school students (AP Stats, less coding requirement)

5 Computation techniques

Although they may not have to deal with large datasets in their future, having students learn how to plan, implement, and access a reasonably sized database filled with data that they find and process will expose them to the mindset that they have the potential to tackle any large and messy data sources they may encounter.

The goal is for students to create something like Gapminder, where they are able to code something for the benefit of others to view their data and promote better understanding of various issues. Students will install R packages that enable dynamic plots and creating GUIs that are elegant and easy to use. Ultimately they will also learn how to share their work online so that anyone can access it.