Finding Data for Historical Research

So you've read Stephen Ruggles, and now you want to engage in some quantitative history of your own. Here are some basics of this kind of humanities data analysis.

Sources of data

General Sources

There are a few general-purpose websites that act as central repositories for data. They are generally affiliated with disciplines or government bureaus. If you are interested in education history since 1960, say, there is an incredible amount of information on the Department of Education's site; if you are interested in historical shipping records, NOAA (the weather organization) keeps large numbers online.

For population-based studies and things like election results, two sites are particularly important.

- The Inter-university Consortium for Political and Social Research has an enormous quantity of data related to historical political behavior; American and European demographic surveys; and so forth.
- IPUMS, the Integrated Public Use Microdata Series, is a project directed by Steven Ruggles that collects very large amounts of census data. In recent years, it has begun releasing *full-count* census data from around the turn of the century.

Specific sets

Other sources, often the ones of the greatest historical interest, may only exist as a single file somewhere.

- The trans-atlantic slave trade database contains information about slave voyages.
- A great deal of library data, both about books and including information about their language, is available from a number of different sources, including the Bookworm project I work on; see me if you want some help creating an extract of some sort.
- Along similar lines: Data for Research, from JSTOR, allows you export both wordcount and metadata information about journal articles that can be useful for the history of scholarship.
- The IMDB contains far more information about the history of movies and television than cultural historians have used.
- A random site we'll deal with more later; information about 19th century whaling ship crew members in Massachusetts is available from the Whaling Crew List Database of the New Bedford Whaling Museum.

But as these examples show, the sort and sources of data you can find may be varied and sundry, indeed. How can you find data adequate for what you want? The same way you find archival sources. Ask around among experts. Google indiscriminately, then discriminately, and then indiscriminately again. Find a source that would use the data if it existed, and then plumb its footnotes. There's no single way to find it.

Create your own

Sometimes the data you want exists but not in a standard digital form. So if there's something you know where to find the precise numbers for that isn't in a standard format like a csv, let me know. Maybe your love of regular expressions can carry you through into a clean copy.

Note that as with all primary sources, you **have to work from what already exists**. Saying "I'd really like a dataset that shows the property holdings of every settler in Kansas in the 1880s" is fine, but unless the data was collected by the census, you're unlikely to get it without extraordinary grant funding.

Tools for analysis

How do you crunch numbers?

Spreadsheets

The easiest way, and not a bad one, is to use a spreadsheet.

There are two places you may encounter limits to spreadsheets.

1. They make it extremely to *share your work*. You can share results: but if you have a special correction, it tends to get lost. This can be catastrophically embarassing: just ask Reinhart and Rogoff. But it also makes it difficult to update your research. If you made a mistake at the beginning, it's hard to wade through to the end.
2. They can't handle extremely large collections of data. You'll know this one when you encounter it.

Scripted tools for data analysis.

If you take my "Humanities Data Analysis" class, you will the language R. This is the most widely tool for statistical analysis nowadays, and beats SPSS or Stata or whatever you learned in your intro statistics class in college, if you took one. And it's free. The downside is that it's basically writing computer code. But that's essentially true of everything else.

If you know some programming already, there are extremely well developed tools for analysis in the Python language. Hardcore scientists sometimes use languages like Julia or Matlab; people with money to spend on computer programs love Mathematica.

Special online tools

There are a variety of sites online that you can upload your data to and perform particularly types of analysis, particularly making a chart. These suffer many of the same shortcomings as spreadsheets, and you'll usually use them *in addition to* rather than *instead of* the traditional forms of analysis. One particularly good resource is Palladio, from Stanford, which is explicitly designed for "Humanities data visualization:" it's up to you to decide if there is such a thing. Tableau is a commercial data visualization platform that has some free products that can produce professional-looking, interactive visualizations.

Cleaning

Cleaning data is hard, and falls below the level of skills appropriate for this course. Regexes are a first and nearly necessary precondition to a lot of data cleaning, which is one of the reasons we suffered through them in week 1.

There is one great freely available, interactive tool for data cleaning: "Open Refine." It is *well* worth using, even for certain sorts of administrative (rather than research) work.