**Analysis and Interpretation in Quantitative Historical Social Science**

Spring 2015

Shanxi University

Cameron Campbell, James Lee, Matt Noellert

This course covers basic issues in study design, research method, and interpretation of results for quantitative social science research. The goals of the course are twofold. The first is to introduce basic principles and techniques for the management and analysis of social data, and interpretation of results. The second is to make students aware of the key challenges researchers face when trying to draw conclusions about social processes and relationships from archival and survey data.

The course will be divided into five modules, each focused on a specific topic in social research: demographic behavior, education, bureaucracy and careers, wealth and landholding and intergenerational mobility. The module topics are broadly similar to those introduced in the Fall 2014 class on “Big Data and New Methods.” The difference is the focus in this class on teaching the specific domain knowledge and quantitative methods necessary for students to produce their own research on these topics and to sensitize them to issues of comparative interpretation. Reading and discussion will focus on exemplary published studies from China or other settings that illustrate such considerations, and showcase appropriate methods.

Each module will last for 2-4 weeks. It will include a substantive portion involving reading and discussion of published studies that exemplify common empirical approaches to the study of the topic, with consideration of research design and methods for analysis. There will also be a methodological portion involving structured exercises using existing datasets. These exercises will focus on familiarizing students with approaches in STATA for manipulating data to prepare it for analysis, and tabulating or graphic the data.

**Module 1  
Demographic Behavior**

***Substance: When does a relationship represent cause and effect?***

This module will focus on fundamental issues in interpretation of descriptive results that commonly arise in social science. The focus is on distinguishing between situations in which it is plausible to claim that an observed relationship reflects cause and effect, and situations in which an observed relationship may be an artifact of other issues such as omitted variable bias, selection effects, and reverse causality. The discussion will be illustrated with examples from descriptive results on community, family and life course influences on individual demographic behavior, including marriage, migration, reproduction, and mortality. The readings consist of relevant contemporary and historical studies that present and interpret results on relationships of demographic behavior to other variables.

***Method: Basic operations in STATA for managing and describing data***

This module will focus on basic operations in STATA for managing and describing data, and guidelines for the preparation of professional-looking tables and graphs that are suitable for presentation or publication. These include:

* Opening and saving datasets via use and save
* Moving data from or to Excel files via import excel and export excel
  + destring following an import to make sure a variable is numeric
* Missing values in STATA
  + mvencode, mvdecode
  + General issues to consider when dealing with missing values
* Labeling variables, and labeling values for categorical variables
* Keeping track of results via log
* Using do files to automate the execution of sequences of commands
* tabulate for basic descriptives
  + Moving results from Excel and then into Word for presentation or publication
  + Guidelines for adding titles, headings, and other content to make sure tables are presentation or publication ready.
  + Calculating and displaying percentages as a way of clarifying results
* table for advanced descriptives
  + Calculating the mean or some other measures for one variable, according to values of other variables.
* graph for producing figures
  + line
  + scatter

For exercises, students will work with the already released CMGPD-LN and CMGPD-SC datasets. Exercises will demonstrate the use of commands in STATA to organize the data to produce variables for community and household context and the individual life course that can be used in an analysis of determinants of individual demographic behavior. Exercises will require students to construct and interpret tabular or visual representations of relationships between demographic outcomes and the community, household, and individual variables they have prepared.

***Reading***

*Required*

Campbell, Cameron and James Z. Lee. 2004. “Mortality and household in Seven Liaodong populations.” Chapter 10 in Bengtsson, Tommy, Cameron Campbell, James Lee, et al. 2004. *Life Under Pressure: Mortality and Living Standards in Europe and Asia*, *1700-1900*. MIT Press, 293-324. Published in Chinese as 托米·本特森,康文林,李中清等. 2008. 压力下的生活：1700～1900 年 欧洲与亚洲的死亡率和生活水平. 北京： 社会科学文献出版社. Translated by 李霞 and 李恭忠.

*Recommended*

Wang Feng, Cameron Campbell, and James Z. Lee. 2010. "Agency, Hierarchies, and Reproduction in Northeastern China, 1789 to 1840." Chapter 11 in Noriko Tsuya, Wang Feng, George Alter, James Z. Lee et al. *Prudence and Pressure: Reproduction and Human Agency in Europe and Asia, 1700-1900*. MIT Press, 287-316.

Chen Shuang, Cameron Campbell, and James Lee. 2014. “Categorical Inequality and Gender Difference: Marriage and Remarriage in Northeast China, 1749-1913.” Chapter 11 in Lundh, Christer, Satomi Kurosu, et al. *Similarity in Difference: Marriage in Europe and Asia, 1700-1900*. MIT Press, 393-438.

Campbell, Cameron and James Lee. 2010. "Fertility control in historical China revisited: New methods for an old debate." *History of the Family*. 15:370-385.

Hao Dong and James Z. Lee. 2014. “Kinship Matters: Long-Term Mortality Consequences of Childhood Migration, Historical Evidence from Northeast China, 1792-1909.” *Social Science & Medicine.* 119: 274-283 (DOI: http://dx.doi.org/10.1016/j.socscimed.2013.10.024).

**Module 2  
Education**

***Substance***

The focus of this module will be on education, especially the role of social origins in access to postsecondary education, choice of major, and transition to employment. Discussion will focus on practical issues in the measurement and comparison of education such as harmonization of measures of type and amount of schooling to facilitate comparison across different settings or time periods, and coding of college major to facilitate analysis. Readings will consist of contemporary studies on these issues that exemplify novel approaches to addressing key methodological issues in studying the determinants or effects of education, most notably, selection effects and omitted variable bias.

***Methodology***

The focus will be on manipulation of variables to prepare them for analysis. This includes recoding them to produce categories that can be compared across datasets, or producing new variables from existing ones for the same purposes. Students will carry out basic exercises in STATA using the extracts from the Shanxi dataset and the Republican-era higher education dataset. These exercises will focus on introducing techniques in STATA for recoding raw data on education and family background, and tabulating or graphing the resulting harmonized data.

Specific topics to be covered include:

* generate to create new variables from existing ones
* replace to modify the contents of existing variables
* recode to convert numeric variables to categorical variables, or transform existing categories
* Basic string variable operations using index and substr and other functions to produce basic numeric variables reflecting the presence or absence of particular text in a string.
* bysort combined with generate and egen to produce new variables describing characteristics of sets of observations, for example, household-level variables constructed from data in individual records.

***Reading***

TBD

**Module 3  
Occupations and Careers**

***Substance***

The focus of this module will be on occupations and careers. For the study of occupations, the emphasis will be on comparisons of occupational structure over time or space. For the study of careers, the focus will be on examinations of careers at the individual level, in particular the influence of individual and other characteristics on trajectories. Readings will consist of consist of contemporary and historical studies that address two relevant major design and methodological issues: 1) how to categorize or index occupations to produce variables that can be used in comparisons across time or space, 2) approaches for studying career trajectories over the life course, including changes in employer or occupation.

***Methods***

Exercises will introduce students to procedures in STATA for categorizing occupations, employers, industrial sectors, and other variables that are originally open-ended. Exercises will also introduce students to basic approaches for describing changes in occupation over time, and representing career trajectories. For their exercises, students may make use of extracts from the *jin shen lu* and from our Republican higher education dataset.

Specific topics to be covered include:

* subinstr and other advanced string operations for cleaning up raw text
* merge as a way of assigning categories or other values to string variables to prepare them for analysis
* Elementary operations using local macros and foreach to process raw text
* Operations for converting back and forth between ‘wide’ data with one record per person or household into ‘long’ data that is organized as one record per time period, or one record per event.

***Reading***

TBD

**Module 4  
Wealth and landholding inequality**

The focus on this module will be on inequality in landholding and other forms of wealth. The emphasis will be on techniques for describing wealth inequality in tabulations or graphs, and summarizing trends over time and difference across space with Gini coefficients and other indices. Readings will include empirical studies of trends and patterns in wealth and landholding inequality.

Exercises will introduce students to procedures in STATA for tabulating and graphing wealth distributions, plotting Lorentz curves, and computing Gini and other indices. Students will make use of the consolidated Shanxi dataset, and the CMGPD-SC.

*Reading*

TBD

**Module 5  
Intergenerational Mobility**

The focus of this module will be on the study of associations in social and economic outcomes across generations. Wealth, landholding, education, and occupation will receive particular attention. Readings will introduce basic approaches and key findings, with an emphasis on historical and comparative studies. Discussion of methods will include approaches to dealing with mobility that distinguish effects of structural changes in education and labor force opportunities from actual social fluidity.

Exercises will introduce students to approaches in STATA for arranging various types of social data to prepare it for analysis of associations across generations. Datasets to be used will include the CMGPD-LN and CMGPD-SC, and the consolidated Shanxi datasets. Students will also produce and interpret simple tabulations

**Grading**

*Exercises – 20%*

Structured exercises will introduce students to specific capabilities in STATA for organizing, recoding, and tabulating or graphing data. Basic exercises using the already released CMGPD-LN and CMGPD-SC will walk students through their first use of the relevant STATA commands, and make sure that they can reproduce . More advanced exercises will introduce the students to the use of these newly learned commands to work with extracts from the Shanxi and higher education datasets. Students may be asked to offer written interpretations of tabulations or graphs.

*Article Reviews (written) – 25%*

Students will select two published research articles relevant to the topics covered in the modules and write reviews of them. The reviews should introduce the data, methods, and conclusions. The review should conclude by discussing ideas for using the Shanxi data to pursue questions raised in the article.

*Article Reviews (presentation) – 10%*

Students will make a presentation to the class on one of the articles they have selected. Their presentation will summarize and critique the article, and answer questions about it.

*Class Participation – 10%*

*Final project (written) – 25%*

Students will carry out a final project using one of the datasets provided in class.

*Final project (presentation)* – 10%

Students will make a presentation based on their final project.

**Suggested resources**

* UCLA has an extensive set of online resources to help with learning STATA: <http://www.ats.ucla.edu/stat/stata/>
* STATA now appears to make most of its documentation available online: <http://www.stata.com/support/documentation/>