TEACHING BASIC STATISTICS USING INTEGRATED GLOBAL CENSUS AND SURVEY DATA

Miriam L. King, Lara Cleveland, and Kristen Jeffers
Minnesota Population Center
225 19th Avenue South, University of Minnesota, Minneapolis, MN 55455
kingx025@umn.edu

We describe the IPUMS databases of consistently coded census and survey microdata from around the world, which are available for free on the Internet (www.ipums.org) to educators and students, and we discuss the advantages of using these real-world data to teach statistics. Based on feedback from statistics educators at the ICOTS 9 conference and at meetings of the International Association for Statistical Education, we developed exercises, posted online, with 9 modules on teaching basic statistics using global census and health survey data. With integrated data from nearly 100 countries, teachers can easily modify these exercises to suit the interests of their students, vary course content every semester, and prepare students to evaluate inconsistencies and cope with imperfections in real-world data.

BACKGROUND

The IPUMS project at the University of Minnesota provides census and survey data from around the world, integrated (i.e., consistently coded) across time and space. IPUMS integration and documentation makes it easy to study change, conduct comparative research, and analyze individuals within their family and community context. From a user-friendly web interface, researchers can see at a glance the variable available for each country and year, and can download a customized dataset with just the information needed for their research project. Data and services are available free of charge, at www.ipums.org.

Along with U.S. data on health, labor force participation, and population and housing, IPUMS contains a wealth of census and survey data from around the world. The IPUMS-International database covers 85 countries and over 300 censuses from the 1960s to the present, with information on individuals' demographic characteristics, education, labor force participation, migration history, and housing, among other topics. IPUMS-DHS, an integrated version of Demographic and Health Survey data for low- and middle-income countries, includes data from over 100 health surveys from Africa and Asia, from the 1980s to the present. IPUMS-DHS offers several thousand consistently coded variables on the health and well-being of women and children, including fertility and family planning, infant and maternal health, domestic violence, household decision-making, marriage and sexuality, sexually-transmitted diseases, vaccinations, nutrition and malnutrition, school attendance, and health behaviors. IPUMS-PMA integrates high-frequency, nationally representative surveys about family planning, water, sanitation, and health in Africa and Asia to monitor progress toward Family Planning 2020 goals.

Funded by the U.S. National Science Foundation and the National Institutes of Health, IPUMS has attracted over 120,000 users since its inception in the mid-1990s, primarily from social science, health science, and computer science fields.

Beginning with an IPUMS workshop at the ICOTS 9 conference, and continuing with participation in meetings of the International Association for Statistics Education, IPUMS staff members have begun to reach out to statistics educators as a new audience for these free online data from around the world.

SPECIAL NEEDS OF STATISTICS EDUCATORS

Discussions with statistics educators were eye-opening for the IPUMS staff members, who are primarily trained in the social sciences. We learned the following:

- While social and health scientists use statistical packages such as Stata, SAS, and SPSS, statistics educators tend to use Excel and the open-source R programming language for analysis;
- While social and health scientists are interested in particular substantive *topics*, statistics

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

educators need specific *types of data* suited to teaching particular techniques (e.g., continuous variables for OLS regression analysis);
- Statistics educators welcome concrete examples of how to use real-world data to teach basic statistics concepts and techniques.

ONLINE EXERCISES USING IPUMS DATA TO TEACH BASIC STATISTICS

In response to these special needs of statistics educators, IPUMS hired two Fellows to create exercises for teaching basic statistics using IPUMS-International and IPUMS-DHS data. Erez Garnai, a doctoral student in Sociology with experience teaching statistics to undergraduate Sociology majors, and Stephanie Chen, an undergraduate majoring in Social Statistics, followed the topics covered in a basic statistics textbook to create a series of 9 exercises based on IPUMS-International census data and 9 exercises based on IPUMS-DHS health survey data. Each exercise noted the statistics topics covered, specified the variables used, recommended country- and year-specific samples to use, supplied R programming code, and included answers and interpretive questions to test students' understanding of the topic. Specifically, these exercises covered:

- Exploring Data (through frequency distributions and graphs)
- Probability
- Probability Distributions
- Confidence Intervals
- Hypothesis Testing
- Comparing Two Groups
- Association between Categorical Variables
- Regression Analysis
- Multiple Regression

These exercises are posted online in the "Help" section of the IPUMS-International and IPUMS-DHS websites, for direct use or modification by statistics educators. Exercises on using IPUMS-PMA to teach basic statistics will be posted on that database's website by the time of the ICOTS 10 conference.

Why might a statistics educator choose to modify this material? For example, a statistics educator from a South Asian country might choose to include samples from that region, as the most meaningful to her students. A statistics teacher who uses group projects might assign different countries and/or different variables to measure, so groups can share and compare their different findings. An educator who needs to vary exercises or test material every semester to prevent cheating via sharing answers from earlier classes can modify the source material (e.g., which countries' data are included), rather than starting from scratch with every class.

And, of course, statistics educators can use the online exercises using IPUMS-DHS and IPUMS-International as just a starting point--as concrete examples to spur his own creativity on how to bring these real-world global data into the classroom.

ADVANTAGES OF USING IPUMS DATA IN THE CLASSROOM

Teachers of statistics classes are ideally preparing their students to use statistical methods after they finish with the class, in jobs and in further studies requiring analysis of real data and "big data." Our conversations with both students and teachers of statistics suggest that teaching materials that accompany statistics textbooks are all too often unlike real-world data. For example, the number of cases may be very small, and there may be no missing or inconsistent data. In real-world data, from censuses, surveys, corporate or government records, or other sources, the number of cases may be very large, and missing and inconsistent data are a given. We contend that it is better to train students with real-world, imperfect data, to prepare them to answer such questions as:

- Who was asked this question or included in this dataset? Does the subject population change over time or vary across samples? How can we deal with such inconsistencies?

- What was the wording of the question that people responded to? Did this change over time or vary across samples? Are these changes or differences likely to influence results?
- What proportion of people refused to answer or didn't know the answer to a question? What are the options for dealing with these missing data? What are we assuming when we just exclude the missing cases?

The variable-specific online documentation for IPUMS is designed to help students and researchers grapple with these questions. For every variable, the variable universe (who was asked the question) is empirically checked and reported for every sample. For example, some countries ask about labor force participation for everyone over age 5, while the United States asks about labor force participation for people age 16 and over. This clear statement of variable universes makes students aware of possible inconsistencies across samples and encourages them to adjust the dataset to compare similar entities (e.g., restrict the dataset to people age 16+ for labor force questions). This lesson learned using IPUMS data can be carried forward to other datasets in future.

Similarly, IPUMS datasets show, for every variable, the question wording for every sample, translated into English. Here, too, students learn that they must make judgment calls and adjustments when using real world data that may come from different sources or change over time. The online "Comparability" text for every variable highlights major comparability issues and differences and serves as a model for thinking critically about data for statistics students.

Finally, censuses and national surveys collect information on topics that countries consider most important for evaluating well-being and planning for the future. They also collect basic demographic information on such topics as age, sex, race, ethnicity, religion, and urban-rural residence. This material supports analysis of disparities and change within countries over time, as well as differences across countries, on important topics affecting people's lives. How has female labor force participation changed over time? Do boys have greater access than girls to primary or secondary schooling? How does the acceptability of wife-beating vary across countries? What factors are associated with stunting of child growth due to sustained malnutrition? Such questions can be answered using IPUMS global census and survey data, and can engage students' interest while they are learning statistical methods.

For IPUMS-International census data, teachers of statistics can create classroom accounts that allow them to share with students exactly the data they want analyzed, while minimizing the work that students must do to access the material.

CONCLUSION

The free, consistently coded, and well-documented online census and survey data available from ipums.org provides a wealth of material for researchers and for teachers of statistics. Based on feedback from teachers of statistics, IPUMS-International and IPUMS-DHS staff created exercises, posted online, to teach basic statistical concepts and methods using global census and survey data. Such exercises can be modified to fit the needs of teachers or serve as examples that spur the instructor's own creativity in creating new material using census and survey data from nearly 100 countries. Using IPUMS real-world data also prepares students to think critically about comparability issues when comparing data from different sources or across multiple years, and to cope with challenges such as missing and inconsistent data. Using tiny and implausibly clean datasets from a statistics textbook is poor preparation for a world increasing inundated with imperfect "big data."

REFERENCES

Boyle, E. H., King, M., & Sobek, M. (2017). *IPUMS-Demographic and Health Surveys: Version 4.1*. [dataset] Minnesota Population Center and ICF International. www.idhsdata.org.

Boyle, E. H., Kristiansen, D., & Sobek, M. (2018). *IPUMS-PMA: Version 1.0* [dataset] Minneapolis, MN: IPUMS. www.pma.ipums.org.

Minnesota Population Center. (2017). *Integrated Public Use Microdata Series, International Version 6.5* [dataset]. Minneapolis: University of Minnesota. www.international.ipums.org.