IPUMS-INTERNATIONAL: DATA RESOURCE FOR DEMYSTIFYING COMPARATIVE STATISTICS ABOUT SOCIETY

Lara Cleveland, Patricia Kelly Hall, and Kristen Jeffers
Minnesota Population Center
University of Minnesota
clevelan@umn.edu

IPUMS-International, the world's largest collection of high-precision census data samples, contains individual-level information on 614 million people in 82 countries spanning five decades. The database, built in cooperation with national statistical offices, provides remarkable access to data for educators wishing to expose students to real-world governmental data. Distinct census responses for each person are coded consistently across time and place; documentation is thorough, harmonized and easily accessible; and the web delivery system enables customized data extracts. Individual-level responses mean data can be used in analyses from simple descriptive tables to advanced statistical modeling. Uniform coding means a statistical algorithm developed to answer a question with one sample (country and year), can readily be applied to other samples, inviting students to extend their exploration of social change. Access to the data is free of charge.

INTRODUCTION

Statistics educators from many countries have issued calls for the use of real-world data in teaching statistics (Engel 2014; Connor & Davies 2002; Garfield & Ben-Zvi 2009, Ramsey 2002). In a recent study of students in a first-year university statistics course, Neumann, Hood & Neumann (2013) found that students reported high levels of interest, motivation, and engagement in the course material in addition to finding that course material had real-life relevance. Joachim Engel (2014) argues that the future of sound evidence-based decision-making in democratic society "requires a certain level of statistical literacy that implies critical thinking and reflecting on metadata" (1). In a world of increasing technology and access to vast quantities of data, the need to educate all citizens to be good and critical consumers of statistics is paramount. Use of manufactured, clean test data for use in teaching isolated statistical concepts is inadequate in today's world. Statistics, both good and bad, both soundly analyzed and manipulated, abound on the internet. Statistics educators must equip students to be good arbiters of statistics they encounter in the real world.

IPUMS-INTERNATIONAL OVERCOMES BARRIERS TO ACCESS

The Integrated Public Use Microdata Series-International (IPUMS International) is a data infrastructure project which disseminates high-precision census microdata samples to researchers world-wide free of cost. The samples in IPUMS are drawn from the very data source used to create official statistics by each country for policy and planning purposes. In partnership with most of the world's national statistical agencies, as well as data archives, research centers, and international organizations, IPUMS International has assembled the most comprehensive collection of census microdata in the world (Figure 1). The microdata records describe more than one-half billion persons nested within families and households, spanning five continents and more than 80 percent of the world's population. IPUMS-International lowers barriers to cross-national and cross-temporal research and teaching by converting international census microdata to a uniform format, providing comprehensive documentation, and making custom downloadable data files available through a user-friendly Web-based access system.

For each person, data include detailed information about geographic location, demographic characteristics, and economic activities. Individuals are nested within families and households, thereby preserving information about inter-relationships within residential groups. The data cover a broad range of population characteristics, including education and literacy, fertility history, child mortality, migration and place of former residence, marital status and consensual unions, disabilities, characteristics of the building (floor, roof, etc.), and a host of other characteristics.

In: J. Engel (Ed.), *Promoting understanding of statistics about society.*
Proceedings of the Roundtable Conference of the International Association of Statistics Education (IASE), July 2016, Berlin, Germany. iase-web.org/Conference_Proceedings.php

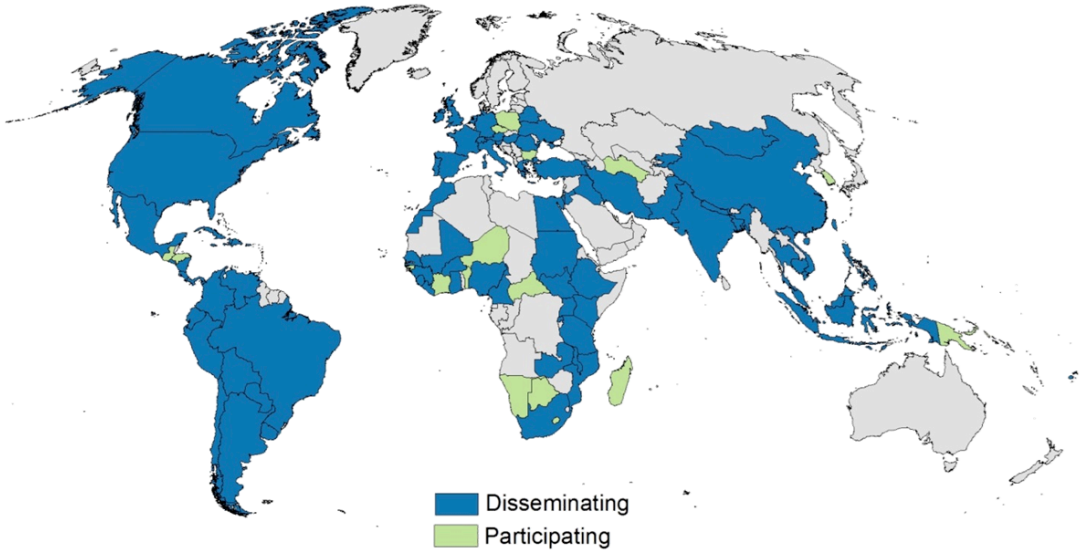


Figure 1. IPUMS-International disseminates population data
for 85 percent of the world's population from 82 countries and 277 census samples

IPUMS DELIVERS HARMONIZED MICRODATA

IPUMS-International converts census microdata for multiple countries into a consistent format, supplying comprehensive documentation and making the data available through a web-based data dissemination system. Although some countries do provide microdata through their country's statistical office, barriers to access remain high (Woolfrey, 2009). Along with supplying unique access to these nationally representative datasets, the principal advantage of IPUMS-International is its replacement of sample-specific variable codes with new integrated codes that are consistent across time and space. This "variable integration" ensures that identical concepts have identical codes, simplifying comparative analysis of multiple samples. More than 700 integrated variables are included in the IPUMS-International database, and the website displays at a glance which variables are included in each sample. Original or "source" variables are also available.

For uncomplicated variables, such as SEX, harmonization simply requires imposing the same codes across all samples (e.g., 1 for male and 2 for female). For more complicated variables, response categories may differ across censuses. Variable integration in IPUMS-International retains original detail by using composite coding. The first digit, called the “general code,” provides information available across all samples (the lowest common denominator data). The second digit provides information available in a substantial subset of the samples, while trailing digits supply additional detail only rarely available. As an example, marital status categories for single, married/in union, separated/divorced, and widowed are available in all samples and have common general codes at the first digit, additional concepts such as consensual union status, distinctions among civil and religious marriages, or polygamous marriage status are captured in the trailing digits for those countries or regions to which they apply.

Guidelines from organizations like the United Nations and the International Labor Organization encourage consistency in census question wording and coding (ILO 2012, United Nations 2007). However, each country's statistical office ultimately decides the subjects covered, the question wording, who was asked a question (i.e., the question universe), and the response categories included in their national census. Inevitably, then, other issues of comparability not covered by IPUMS-International's composite coding schemes arise for researchers doing comparative analysis of census data. Sample descriptions and variable-specific documentation on the IPUMS-International website are designed to highlight possible comparability problems, so users can make informed judgments or adjustments and avoid inadvertent errors. The online documentation for every variable shows, with a few clicks, the codes and unweighted frequencies, the universe, question wording, instructions to enumerators, and a discussion of comparability

issues for each country/sample. Because users generally care about a subset of countries and/or years, the documentation can be easily limited to show only the sample(s) of interest.

Data are useful only when researchers understand what they mean. Most of the challenges Gal (2003) cites in his analysis of products from statistical agencies are related to metadata discovery and availability. IPUMS-International provides harmonized English-language documentation on each sample. This documentation covers enumeration procedures and instructions; definitions of households, dwellings, group quarters, and other enumeration units; and scanned images of original-language versions of the questionnaires. IPUMS also provides descriptions of the sources for each variable, including question wording and instructions (in the original and translated into English), universe definitions, frequency distributions, and variable codes. Comparability discussions describe any deviations of particular censuses from the standard variable definition and address differences over time and across countries.

INTERACTIVE, USER-DRIVEN TOOLS AND FEATURES

The metadata browser limits the information displayed to only those elements relevant to a given research project, as defined by the user. This filtering capacity creates individually-tailored documentation that highlights subtle problems of comparability without overwhelming users with unnecessary information.

The IPUMS data access system allows users to merge datasets, select variables, define population subsets, construct customized family interrelationship variables, and draw subsamples tailored to their specific needs. The IPUMS disseminates pooled extracts in a single dataset, custom-tailored to the precise research needs of the user. Each user-generated extract contains only the requested microdata accompanied by the corresponding set of DDI (Document Data Initiative) compatible metadata and a codebook suitable for constructing a system file in SPSS, SAS or STATA. Interactive metadata and original source material in the official language are also available on the website.

IPUMS-International adds value to official data through the cross-temporal and geospatial integration of population data and metadata. Additional value-added features of interest to statistics educators include extensive integrated metadata such as documentation and guidance on the variability in sample design (Cleveland, Davern & Ruggles, 2011); within-household relationship pointer variables (Sobek and Kennedy, 2009), an online data tabulator, GIS boundary files, and several tools to help instructors manage classroom data usage.

Researcher virtual data carrel

Each registered user of IPUMS-International has a private, password-protected extract history page. This page contains the statistical package syntax files and data for download of recent extract requests, as well as the extract syntax file and description (if user provided) for each data order ever requested by that user. With one click on the "resubmit" button, a researcher can regenerate the same extract. The "Revise" button opens the syntax file so the researcher can modify the data request. This is particularly useful in the classroom where a complex classroom exercise or exam can be re-used for a new class by modifying the data request with a different country or year. It is also useful for novice statistical researchers.

Classroom features

The IPUMS international user register system also includes a classroom feature. Course instructors can apply to register a class for a specified duration of time. Upon approval, the instructor receives a code for their students. Students get facilitated registration and are automatically assigned to membership in the class. Instructors can push common data extracts to students enrolled in the class. Students also have full IPUMS user capabilities for the duration of the quarter or semester and can also make their own customized data extracts.

Educators also find the uniform variables and coding schemes in IPUMS useful in teaching. Uniform coding means that a statistical algorithm developed to answer a question with one sample (country and year), can be readily applied to other samples. This feature is useful for facilitating student exploration of new contexts. It can also come in handy for developing exam

materials, since a problem set used for an exam one semester can be re-run to output a different set of answers (using a different sample) for the same exam in a different semester.

Online tabulator

The IPUMS International website also features a robust online tabulation system available to registered IPUMS International data users. Researchers can quickly analyze data files for individual samples, pooled samples from multiple census years within a country, or all pooled samples from a world region. From the home page, select "Analyze Data Online" and chose a single year or group of data files for analysis. An example below provides illustrates of analytic capabilities and sample output from the tabulation system.

GETTING TO KNOW REAL-WORLD DATA

Real world data are messy. Some people are hard to reach, especially those who also tend to be most vulnerable and most in need of services of some kind. Surveys and registries will miss people who do not live in traditional housing, do not have formal or permanent addresses, do not have phones, or who live in hard to reach geographical areas. Data about people are tricky; misreporting, misinterpretation and simple transcription errors can happen at every stage of data collection, transformation, and analysis.

Basic descriptives, expected patterns and missing or unknown values

The first step in data analysis is review of descriptive information about the data. Checking descriptive output means more than simply checking variables for the right names, codes and labels. The researcher must check that values, summaries and distributions "make sense." For example, should the mean age of the population in a developing country be higher or lower than the mean age in a developed country? The answer should almost always be "lower." Age distributions of a country's population should bulge at lower ages and gradually trail off at higher ones. Male-to-female ratios in a population should be relatively even, but we know that some countries will have higher male-to-female ratios due to sex selective social practices. Knowing something about the expected population helps a researcher make an initial pass through descriptive information "at a glance." The workshop will provide a guide for statistical educators detailing some of the expected patterns in variables based on known demographic and social trends.

Missing values are common in census data. Statistical offices have different means of coding missing cases. Some impute values for a select set of variables. Others assign non-numeric codes, still others assign extreme high or low values to cases. IPUMS International recodes values consistently across countries, but often assigns high 998-style codes to missing or unknown response categories. Some survey and census questions are asked only of subsets of the surveyed population. Employment status is often asked only of people older than 10, 12 or most often 15 years of age. Fertility questions are often asked only of women of childbearing age. The IPUMS International project devotes a special section of variable documentation to universes and assigns zero or high 999-style codes to "not in universe" categories. Failure to consider universe implications in analysis can severely distort results and results in very bad research conclusions. Examples of universe issues will be covered in the IPUMS International workshop.

Cross-year and cross-cohort assessments of data quality

IPUMS International uses a few assessment strategies in our data processing activity that could be useful to share with statistics students confronting real world data for the first time. Countries partnering with IPUMS to provide census data samples to the public typically contribute more than one census year. Social change often occurs because younger cohorts of people adopt new behaviors, which means that population trends change slowly and gradually over time. When we review our output data for errors, one technique we use is to compare distributions within a single variable across time--i.e. a crosstab with the variable of interest as rows and census years in the columns. Barring a shock to the country's social or physical environment or significant changes in the categorization of the variable of interest, we expect to see similar distributional patterns

across census years. This data review aids us in identifying changing universes across census years, differences in sample characteristics, and errors in data coding.

To evaluate IPUMS data, we also construct birth cohorts for the population of each census in order to compare rates of relatively time-invariant characteristics within common cohorts across census years. For example, we expect sex-ratios at each age in the population to remain relatively stable over time. International migration and mortality usually contribute only minimally to changes from one census to the next. Likewise, we expect that educational attainment has been reached for most individuals by age 25 or 30. If we look at the percent of 35 year-olds completing secondary education in 1970, we would expect to find the same percent of 45 year-olds having completed secondary education in 1980, and of 55 year-olds in 1990. This technique for reviewing data has helped us identify unknown sample characteristics, find subtle differences in a country's variable coding approach from one census to the next. We also gain new insight into the social context of the country. Is the population old or young, fertile or less so, highly educated or employed or not, etc. For example, we have found that significant deviations in age-sex ratios signal periods of hardship or heavy out-migration in a country. Similar investigations by students will help them better understand their variables of interest and can potentially help them better understand the social environment in the country or sub-region of interest.

ENGAGE STUDENTS IN REAL WORLD ISSUES

One advantage of having access to microdata is the ability to conduct a wide range of customized analytical approaches. One very powerful feature of microdata that does not have to involve complicated modeling is the ability to disaggregate effects by age, by sex, by geographic area, or by any number of other characteristics. Once a student runs an analysis at the country level, they can also use a number of techniques to compare results for different groups or for different geographic areas within the country. Here we walk through a rather basic investigation of an age distribution in Colombia followed by two additional suggestions for examining group differences and inequalities. Piquing student interest by uncovering slight data anomalies, time trends, and group differences help raise myriad questions for further investigation.

The Simple, or Not So Simple, Case of Age Distributions in Colombia

A very simple, or seemingly simple, example of cross-temporal age distributions in Colombia show the range of possibilities and flexibility made possible through access to census microdata. Table 1 show a screen shot of output from the IPUMS International tabulator. It includes only a subsection of the age distribution for persons ages 0 to 12 years across 5 censuses from Columbia (1964, 1973, 1985, 1993, and 2005). The actual web output shows the full age distribution from age 0 to 100+ and includes a small number of unknown ages, coded 999. By default, the output cells are color-coded to show deviations from expected values, with smaller than expected values in blue and larger than expected in red. Corresponding Z values are shown in the lower inset. Evidently, the percentage of the population at low ages is declining over time. As always, with real world data, it is useful to ask students whether such a pattern "makes sense" and how one could follow up on whether their explanation is correct. Decreasing fertility rates seem evident from the distribution and provide one potential explanation of the evident change over time. Alternately, or in addition, people might be living longer.

In addition to reviewing change over time, an instructor could focus on anomalies in the distribution with any given year. On average, we expect age patterns and age-sex ratios to follow relatively smooth trends. One exception occurs with peaks based on digit preferences, such as spikes in age distributions at ages ending with zero or 5. Such age heaping tends to be most prevalent among older individuals and in developing countries with low numeracy skills among subsets of the population. In young-age portion of the age distribution shown in Table 1, dips in population at certain ages (seen in the difference between 0 and 1 year olds in 1964) or larger than expected jumps (such as that between 1 and 2 year olds in 1985 or 1993) are slightly larger than we see across the rest of the distribution. Most percentage changes from one year of age to the next are within a couple tenths of a percent.

Table 1. Age of population (percent and weighted totals) for 5 census samples in Columbia (1964, 1973, 1985, 1993, 2005)

Cells contain: -**Column percent** -Weighted N	sample					
	170196401 Colombia 1964	170197301 Colombia 1973	170198501 Colombia 1985	170199301 Colombia 1993	170200501 Colombia 2005	***ROW TOTAL***
0: Less than 1 year	**3.6** 630,700.0	**2.5** 504,070.0	**2.2** 610,983.0	**2.0** 626,730.0	**1.9** 785,366.1	***2.3*** *3,157,849.1*
1: 1 year	**3.0** 520,000.0	**2.6** 509,430.0	**2.2** 607,247.0	**2.0** 642,540.0	**1.9** 776,050.0	***2.2*** *3,055,267.0*
2: 2 years	**3.7** 655,400.0	**3.0** 595,830.0	**2.5** 669,956.0	**2.4** 785,680.0	**1.9** 771,925.6	***2.5*** *3,478,791.6*
3: 3	**3.7** 651,000.0	**3.2** 633,320.0	**2.6** 711,333.0	**2.5** 793,550.0	**1.9** 774,043.9	***2.6*** *3,563,246.9*
4: 4	**3.6** 626,450.0	**3.2** 627,110.0	**2.6** 722,360.0	**2.4** 778,560.0	**2.0** 793,461.2	***2.6*** *3,547,941.2*
5: 5	**3.4** 595,150.0	**3.2** 631,090.0	**2.7** 750,827.0	**2.4** 757,350.0	**2.1** 848,433.1	***2.6*** *3,582,850.1*
6: 6	**3.3** 569,200.0	**3.1** 619,550.0	**2.6** 702,695.0	**2.3** 725,820.0	**2.1** 834,011.2	***2.5*** *3,451,276.2*
7: 7	**3.4** 592,900.0	**3.3** 650,250.0	**2.5** 686,214.0	**2.3** 736,670.0	**2.0** 812,865.2	***2.5*** *3,478,899.2*
8: 8	**3.2** 562,650.0	**3.2** 642,580.0	**2.4** 656,438.0	**2.4** 762,920.0	**2.1** 836,809.7	***2.5*** *3,461,397.7*
9: 9	**2.7** 480,650.0	**2.9** 575,670.0	**2.2** 600,643.0	**2.2** 711,370.0	**2.1** 851,508.5	***2.3*** *3,219,841.5*
10: 10	**2.9** 514,000.0	**3.2** 641,390.0	**2.4** 659,797.0	**2.3** 751,150.0	**2.1** 866,233.2	***2.5*** *3,432,570.2*
11: 11	**2.4** 427,900.0	**2.8** 547,140.0	**2.2** 591,774.0	**2.2** 716,250.0	**2.1** 861,325.2	***2.3*** *3,144,389.2*
12: 12	**2.9** 502,700.0	**3.1** 623,880.0	**2.4** 658,290.0	**2.4** 773,890.0	**2.1** 837,219.9	***2.5*** *3,395,979.9*

Color coding:	<-2.0	<-1.0	<0.0	>0.0	>1.0	>2.0	Z
N in each cell:	Smaller than expected			Larger than expected			

If we consider the dip in the age distribution between 0 and 1 year olds in 1964, we can set up a whole range of questions relevant to working with real world data. Does understanding the dip require subject-matter or country contextual expertise? Does it require advanced data quality assessment techniques? What role does metadata play in understanding real world data? That dip may be attributable to bad weather, bad agricultural output or a difficult political situation in the country. It could be attributable to a data quality issue age reporting or in data coding. The anomaly gives an instructor a way of introducing a number of concepts and considerations at play in analyzing real-world data.

Subject matter knowledge. It can be helpful to have some subject matter expertise to understand an age distribution anomaly. A dip could be evidence of lower birth rates in 1963 in Colombia or it could indicate low infant survival rates from birth to age one, or some combination of the two. It would be useful to know something about the conditions in Columbia during the relevant time period. Were weather or political conditions uncharacteristically harsh during 1963 in ways that they were not in the years before and the year after 1963? Instructors teaching with real world data may sometimes have to provide a bit of subject matter context to help students interpret potential factors contributing to observed patterns.

Data quality issues. To consider the potential explanatory factors listed above, one must assume that the data reflect the actual population totals with a good degree of accuracy. Students and instructors will find that real world data is fraught with errors. Consider the act of fielding a census. In a typical traditional census, thousands of census enumerators walk from door-to-door, asking a representative from a household a range of census questions about the inhabitants of that household. Although enumerators are trained on how to interpret questions and about the range of

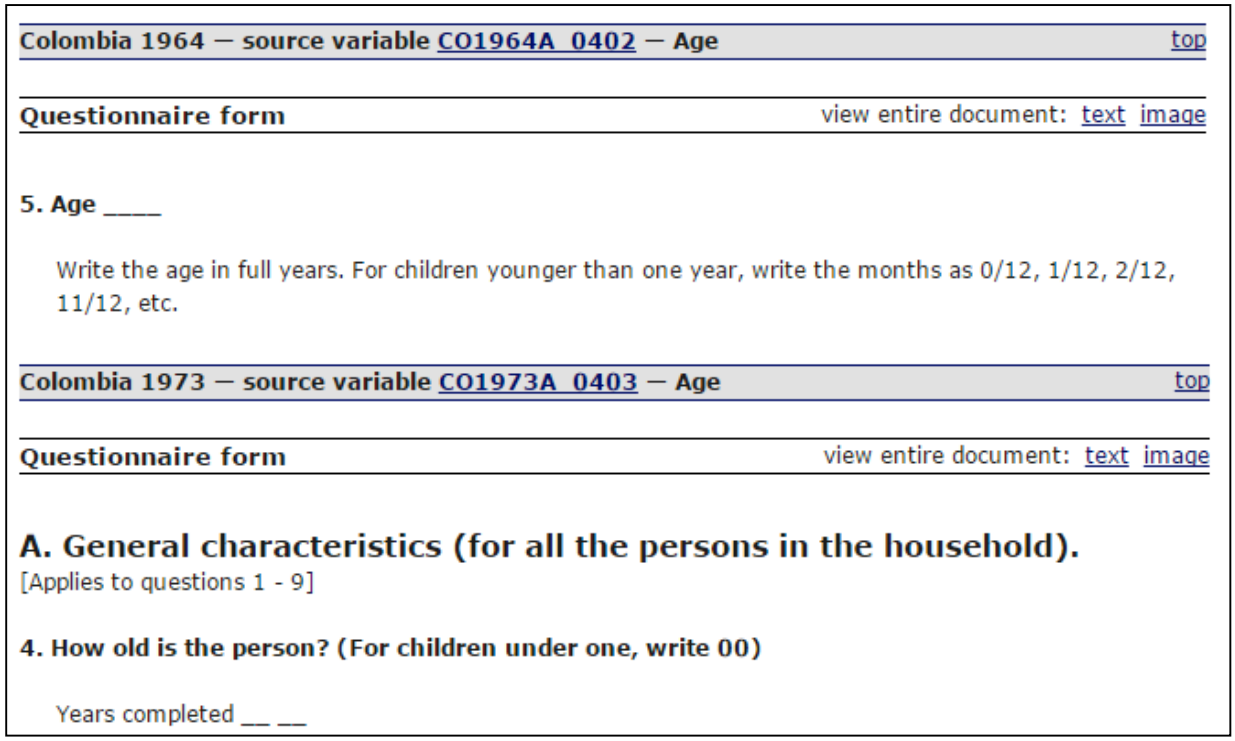
acceptable response options, responses are ultimately doubly interpreted by the respondent and the enumerator. While a question about sex or age may seem relatively straightforward, a question about employment status requires that both parties understand how work is defined, when the "work" reference period starts and ends, etc. Add to that the fact that responses must be entered or captured in digital format. Data often undergo editing and adjustment by statistical officer personnel, whom we must rely on for accuracy.

Is our small age distribution anomaly in Colombia 1963 a result of the process of data collection or does it reflect real numbers in the population? Data quality assessment involves knowing something about the context in which the data were collected, such as those issues defined above. We can also learn something by examining the data directly. Students can look at distributions of other variables to see whether they make sense or have similar anomalies. They can look at rates of missing values in the data. They might look at the distribution among subgroups of the population or within different geographic areas.

At IPUMS International, we have implemented a data quality evaluation process to check for cohort consistencies over time in characteristics we expect to remain relatively stable in the population. Cohort comparisons are probably best done with knowledge of age by birth month as well as information about mortality and migration rates. However, even a simple comparison of characteristics among cohorts reporting the same birth year can help highlight egregious inconsistencies in the data from one census to the next. In this example, we can look at the relative age distribution of birth cohorts in subsequent censuses. Is the cohort of people born in 1963 stable? Even though the table is truncated, we can take a cursory look at the group of people who are age 10 in the census taken 9 years later (in 1973). Rather than 10-year-olds, it seems to be the 9-year old group where we see the dip in age distribution. There might be a consistent population trend here, but we need more information about when the census was fielded. Metadata is the key, and it is an essential component of working with real world data.

Importance of metadata. As mentioned above, real world data is collected, recorded and recoded by humans and is subject to error. Real world data about the social world is subject to interpretation by respondents and by data collectors. IPUMS International provides a wealth of information to aid researchers in understanding detailed nuances in the data. Sample level information about field enumeration methods and census structure are provided in the "Sample

Figure 2. Census questionnaire text for Colombia from the IPUMS International metadata system



Descriptions" section of the website. We learn from the page on Colombia that the census dates were July 15, 1964 and October 24, 1973, thus partially explaining our cohort comparison age difference in the example. We also learn that the 1964 sample is a 2% sample of individuals from the census while the 1973 sample is a 10% sample drawn at the household level and containing the individuals within the household. Knowing about the sample design may have implications for how we approach the data and how we account for estimation error rates.

At the variable level, IPUMS provides basic descriptions of the variables, detailed comparability discussions, full wording of the census questionnaire and instructions to census enumerators, and information about the input variable structure. Figure 2 shows the wording of the age question in the Colombia censuses of 1963 and 1974. For children younger than one year, the text in 1963 asked enumerators to write the number of months of age for persons under age 1. In 1974, the questionnaire asked about age in number of years completed. Students could be asked a series of questions about how the differences in age reporting might contribute to data error or accuracy. They might be asked to consider when and why one would need to know about number of months under age 1 and about how reporting different units of measurement in a single response item can be confusing. This is also an opportunity to teach students about how historical data often used composite recording techniques to save computer storage space in an era when storage was more than 10,000 times as expensive as it is today.

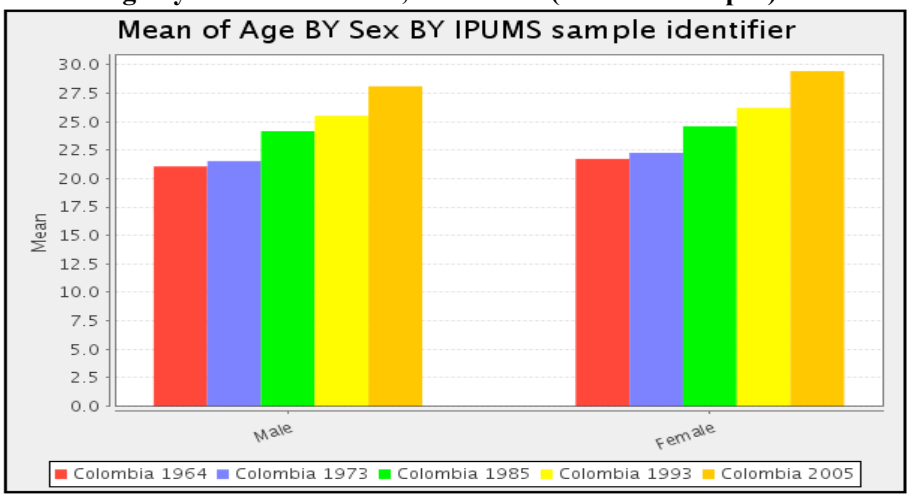
The tabulator available on the IPUMS website is quite powerful. Users can recode or subset variables, compare means, create correlation matrices, and even do regression analysis. For example Table 2 shows standard tabulator output for mean age of the population by sex and by

Table 2. Mean population age by sex in Colombia, 1964-2005 (tabulator output)

Main Statistics							
Cells contain: -Means -SRS Std Errs -Weighted N		sample					
		170196401 Colombia 1964	170197301 Colombia 1973	170198501 Colombia 1985	170199301 Colombia 1993	170200501 Colombia 2005	***ROW TOTAL***
sex	1: Male	**21.07** .044 8,610,150.0	**21.54** .018 9,687,950.0	**24.17** .016 13,459,058.0	**25.52** .015 15,743,970.0	**28.11** .014 19,812,706.0	***24.87*** *.008* *67,313,834.0*
	2: Female	**21.72** .044 8,872,400.0	**22.26** .018 10,200,360.0	**24.59** .016 13,847,377.0	**26.22** .015 16,392,600.0	**29.44** .014 20,751,346.9	***25.70*** *.008* *70,064,083.9*
	COL TOTAL	***21.40*** *.031* *17,482,550.0*	***21.91*** *.013* *19,888,310.0*	***24.38*** *.011* *27,306,435.0*	***25.88*** *.011* *32,136,570.0*	***28.79*** *.010* *40,564,052.9*	***25.30*** *.006* *137,377,917.9*

Color coding:	<-1.0	<-0.5	<0.0	>0.0	>0.5	>1.0	Z
Mean in each cell:	Smaller than average			Larger than average			

Figure 3. Mean Age by Sex in Colombia, 1964-2005 (tabulator output)



sample in Colombia. Figure 3 shows a default bar graph of the data. A number of optional graph formats are available to output to the screen. We would expect an aging of the population in countries undergoing development. We also typically expect to see slightly higher average ages for women, reflecting their slightly longer life expectancies.

Microdata flexibility for disaggregation. With microdata, any number of approaches are available to researchers, instructors, and students to further investigate the changes in mean age. One might review full age distributions in more detail noting percentages in old age categories. Alternately, one might use the children ever born variable to calculate fertility rates among particular age groups of women, a truer measure of fertility. Since microdata samples from censuses are quite large, slicing and dicing across subgroups is entirely possible. For example, many social phenomena vary significantly between urban and rural areas. Using the same set of samples and the same mean age totals, Tables 3 and 4 show mean age for males and females separated by rural and urban areas respectively. Interestingly, not only is average age higher in cities, but the gender with higher average age opposite in the two areas. Females have much higher average ages than males in urban areas, but slightly lower average age in rural areas. Many potential explanations for these differences, and the extent to which the differences should be considered significant, could be investigated in a series of guided exercises. The examples

Table 3. Mean age of the population by sex for rural areas in Colombia, 1964-2005

Statistics for urban = 1(Rural)							
Cells contain: -Means -SRS Std Errs -Weighted N		sample					
		170196401 Colombia 1964	170197301 Colombia 1973	170198501 Colombia 1985	170199301 Colombia 1993	170200501 Colombia 2005	***ROW TOTAL***
sex	1: Male	**21.10** .062 4,354,550.0	**21.34** .030 3,946,920.0	**24.00** .030 4,551,817.0	**24.84** .028 4,684,920.0	**27.52** .022 5,146,476.7	***23.95*** *.013* *22,684,683.7*
	2: Female	**20.77** .065 4,046,450.0	**20.84** .030 3,632,800.0	**23.30** .031 4,049,661.0	**24.29** .030 4,194,180.0	**27.01** .023 4,654,451.2	***23.41*** *.014* *20,577,542.2*
	COL TOTAL	***20.94*** *.045* *8,401,000.0*	***21.10*** *.021* *7,579,720.0*	***23.67*** *.021* *8,601,478.0*	***24.58*** *.021* *8,879,100.0*	***27.27*** *.016* *9,800,927.8*	***23.69*** *.009* *43,262,225.8*

Color coding:	<-1.0	<-0.5	<0.0	>0.0	>0.5	>1.0	Z
Mean in each cell:	Smaller than average			Larger than average			

Table 4. Mean age for the population by sex for urban Areas in Colombia, 1964-2005

Statistics for urban = 2(Urban)							
Cells contain: -Means -SRS Std Errs -Weighted N		sample					
		170196401 Colombia 1964	170197301 Colombia 1973	170198501 Colombia 1985	170199301 Colombia 1993	170200501 Colombia 2005	***ROW TOTAL***
sex	1: Male	**21.04** .062 4,255,600.0	**21.68** .024 5,741,030.0	**24.25** .020 8,907,241.0	**25.80** .018 11,059,050.0	**28.32** .019 14,666,229.3	***25.34*** *.010* *44,629,150.3*
	2: Female	**22.52** .060 4,825,950.0	**23.05** .022 6,567,560.0	**25.12** .019 9,797,716.0	**26.89** .017 12,198,420.0	**30.14** .018 16,096,895.8	***26.66*** *.009* *49,486,541.8*
	COL TOTAL	***21.82*** *.043* *9,081,550.0*	***22.41*** *.016* *12,308,590.0*	***24.71*** *.013* *18,704,957.0*	***26.37*** *.012* *23,257,470.0*	***29.27*** *.013* *30,763,125.1*	***26.03*** *.007* *94,115,692.1*

Color coding:	<-1.0	<-0.5	<0.0	>0.0	>0.5	>1.0	Z
Mean in each cell:	Smaller than average			Larger than average			

presented here are simple to carry out using the IPUMS International online tabulator. Of course, in a course that is designed to teach or utilize student statistical software skills, instructors have the flexibility to guide students in producing similar output using statistical software packages.

Subnational Sustainable Development Goal (SDG) measures

The United Nations 2030 Agenda for Sustainable Development proposes 17 goals and 169 targets that aim to complete what the Millennium Development Goals did not achieve. The goals and targets concentrate on the eradication of poverty, hunger and inequality; access to education and healthcare; gender equality; environmental sustainability; economic, social, and technological progress, and the establishment of new partnerships for the achievement of these goals. The proposed framework for monitoring the Sustainable Development Goals (SDGs) emphasizes the need for disaggregated indicators that measure progress among different demographic and social groups at various levels of sub-national geography. The Sustainable Development Solutions Network recommends spatial disaggregation and stratification by sex, gender, age, income, disability, ethnicity, indigenous status, economic activity, and migrant status for nearly half of the proposed monitoring indicators. While enhanced data collection will almost certainly be necessary to monitor several SDGs, high-precision census microdata samples, like those disseminated by IPUMS-International, represent useful data that are part of the existing statistical infrastructure of most developing countries. These data are highly representative of national populations, are collected at regular intervals, and include measures of the population characteristics required for SDG indicator disaggregation.

Goal 8.8 pertains to the "percentage of young people not in education, employment, or training (NEET)." For analytical purposes, we favor % PEET (100%-NEET), recasting the indicator in a positive form and can construct the indicator using two IPUMS integrated variables: SCHOOL (school attendance) and EMPSTAT (Economic activity status) to construct the proportion of youths aged 15-24 who are engaged in some form of employment, education or training. Calculating this indicator at the national level meets the basic reporting requirement. However, calculating the indicator by subnational geographic unit (also available in the IPUMS with corresponding shapefiles for mapping), provides a much richer picture of which areas of the country best meet the goal or require the most attention.

Binder Oaxaca decomposition of group differences

The Binder-Oaxaca-Duncan decomposition technique is a statistical tool that enables researchers to delve even deeper into social processes using microdata. The method is named after three quantitative social scientists: Otis Dudley Duncan, Alan S. Binder and Ronald L. Oaxaca. It is one of the few statistical tests of discrimination allowed into evidence by the legal system of the United States (Committee on National Statistics, 2004). The method decomposes measured differentials on a variable (such as wages, health outcomes or homeownership) into two components. One component measures the "explained difference"—that is, amount of difference between the two groups based on their individual measurable characteristics (age, education, years of employment, etc.) *when the characteristics of the two groups are treated equally*. The second component–-the "unexplained difference"—measures the differential when society treats the characteristics of each group differently, because of discrimination, structural change over time or changes in public policies.

CONCLUSION

It is usually quite easy to find differences across sub-regions within a country, or among different groups based on gender, race, ethnic group, or even age group. Clear evidence of such differences, even in fairly simple comparative approaches easily carried out by students at many levels also provides a foundation for a critical discussion about the presentation of analytic results. Some levels of analysis or types of categorizations can mask important differences (intentionally or inadvertently). With a few tools to aid statistical educators in understanding the data, IPUMS can be used to effectively engage students in accessing real-world data to understand social problems. Students will be interested in how to critically evaluate news reports involving statistics and,

importantly, in how to make sure their own analyses are rigorous enough to qualify as accurate. Students and instructors may have to be prepared to consult references and conduct investigations outside the mathematical world of traditional statistical instruction in order to fully explore the questions raised by examining statistics about the social world, even those as simple as the age distribution comparisons shown in this paper. That alone makes teaching with real world data both exciting and daunting. Partnerships between statistically trained subject matter scholars and statistics educators are fertile ground for engaging students and training well-rounded social statisticians.

ACKNOWLEDGEMENT

Data disseminated through IPUMS-International are generously provided by the national statistical offices which collect census data. The IPUMS-International data infrastructure project is housed at the Minnesota Population Center, University of Minnesota, and is supported by grants from the National Institutes of Health and the National Science Foundation: 5 R01 HD043392, 5 R01 HD047283, and SES-085141.

REFERENCES

Cleveland, L., Davern, M. & Ruggles, S. (2011). Drawing Statistical Inferences from International Census Data. *IPUMS International Working Paper.*

Committee on National Statistics, The National Research Council. (2004) *Measuring Racial Discrimination,* Chapter 7: Measuring Racial Discrimination, R.M. Blank, M. Dabady and C.F. Citro, eds., Washington, DC: The National Academics Press, pp. 118-161.

Connor, D. & Davies, N. (2002). An International Resource for Learning and Teaching. *Teaching Statistics*, *24*(2), 59-61.

Engel, J. (2014). Open Data, Civil Society and Monitoring Progress: Challenges for Statistics Education. In Makar, K., de Sousa, B, & Gould, R. ed. *Proceedings of the Ninth International Conference on Teaching Statistics.* Voorburg, The Neterlands: International Statistical Institute. *iase-web.org*

Gal, I. (2003). Expanding Conceptions of Statistical Literacy: An Analysis of Products from Statistics Agencies. *Statistics Education Research Journal*, *2*(1), 3-21.

Garfield, J. & Ben-Zvi, D. (2009). Helping Students Develop Statistical Reasoning: Implementing a Statistical Reasoning Learning Environment. *Teaching Statistics*, *31*(3), 72-77.

International Labour Organization-ILO. (2012) International Standard Classification of Occupations, ISCO-08, Geneva: ILO 2012.

Neumann, D., Hood, M. & Neumann, M. (2013). Using Real-Life Data When Teaching Statistics: Student Perceptions of this Strategy in an Introductory Statistics Course. *Statistics Education Research Journal*, *12*(2), 59-70.

Rumsey, D.J. (2002). Statistical Literacy as a Goal for Introductory Statistics Courses. *Journal of Statistics Education*, *10*(3), http://www.amstat.org/publications/jse/v10n3/rumsey2.html.

Sobek, M. & Kenned, S. (2009). The Development of Family Interrelationship Variables for International Census Data. *MPC Working Paper Series 2009-02.* (http://www.pop.umn.edu/sites/www.pop.umn.edu/files/Working%20Paper%202009-02.pdf)

United Nations (2007). *Principles and Recommendations for Population and Housing Censuses, Revision 2*. New York: United Nations Publications.

Woolfrey, L. (2009). African microdata access survey: an investigation into current data sharing policies and practices at African National Statistics Offices. Presentation at the 57th Session of the International Statistical Institute: statistics: our past, present and future, Durban, South Africa, 16-22 August.