Lecture 1: Introduction

Ernesto F. L. Amaral

August 22–24, 2023
Introduction to Sociological Data Analysis (SOCI 600)

www.ernestoamaral.com

Source: Healey, Joseph F. 2015. "Statistics: A Tool for Social Research." Stamford: Cengage Learning. 10th edition. Chapter 1 (pp. 1–22).

TEXAS A&M UNIVERSITY

Outline

- Course objective
- Why study statistics?
  - Describe role of statistics in social research
- Types of variables
  - Causal relationships: independent, dependent
  - Unit of measurement: discrete, continuous
  - Level of measurement: nominal, ordinal, interval-ratio
- General classes of statistics
  - Univariate, bivariate, multivariate, inferential
- American Community Survey (ACS)
- Stata

Main objectives of this course

- **Statistics are tools** used to analyze data and answer research questions
- Our focus is on how these techniques are applied in the **social sciences**
- Be familiar with **advantages and limitations** of the more commonly used statistical techniques
- Know **which techniques are appropriate** for a given purpose
- Develop statistical and computational skills to carry out **elementary forms of data analysis**

Data, software, and techniques

- This course is an introduction to social statistics using data from the American Community Survey (ACS) and the statistical package Stata
  - Univariate analysis
    - Mode, median, mean, boxplot
  - Measure of association for nominal-level variables
    - Chi Square
  - Measure of association for ordinal-level variables
    - Spearman's Rho
  - Measures of association for interval-ratio-level variables
    - Scatterplots, Pearson's *r*, analysis of variance (ANOVA)
  - Multivariate analysis
    - Ordinary least square regression (linear regression)

Why study statistics?

- Scientists conduct research to answer questions, examine ideas, and test theories
- Statistics are relevant for **quantitative research projects**: numbers and data used as information
- Statistics are mathematical techniques used by social scientists to analyze data in order to **answer questions and test theories**

Importance of data manipulation

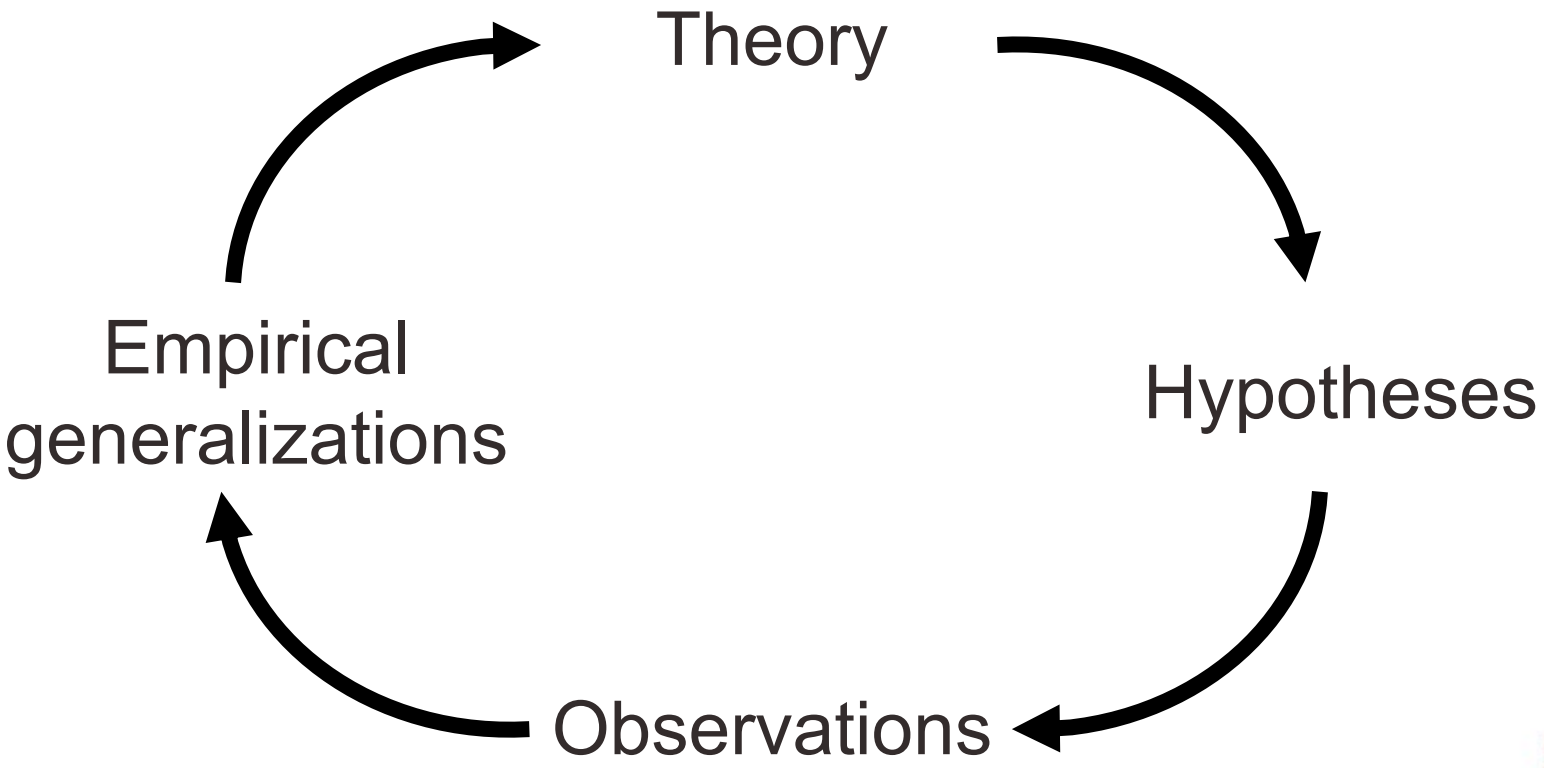
- **Studies without statistics**
  - Some of the most important works in the social sciences do not utilize statistics
  - There is nothing magical about data and statistics
  - Presence of numbers guarantees nothing about the quality of a scientific inquiry
- **Studies with statistics**
  - Data can be the most trustworthy information available to the researcher
  - Researchers must organize, evaluate, analyze data
  - Without understanding of statistical analysis, researcher will be unable to make sense of data

Statistics role in scientific inquiry

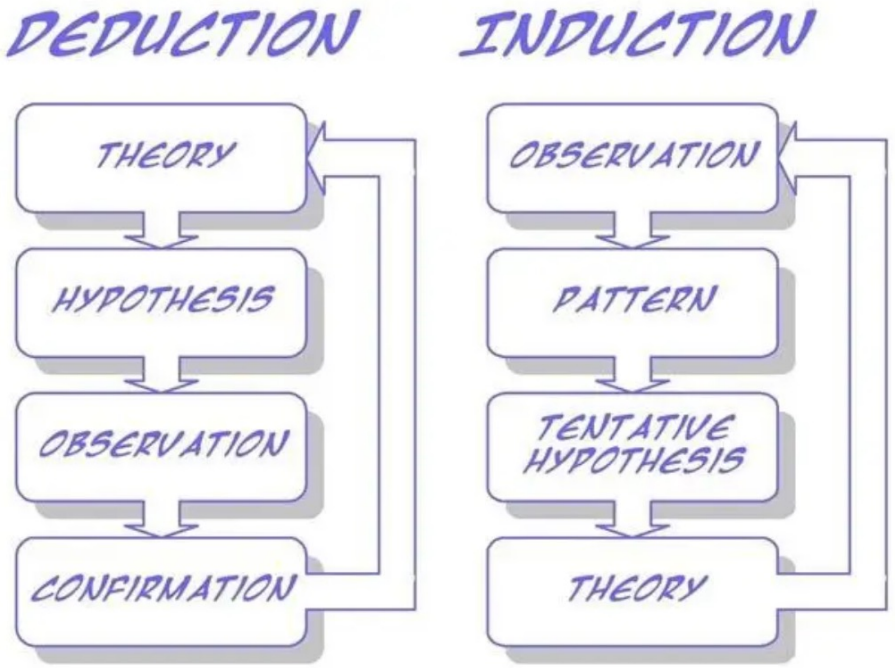
- **Research** is a disciplined inquiry to answer questions, examine ideas, and test theories
- **Statistics** are mathematical tools used to organize, summarize, and manipulate data
- **Quantitative research** collects and uses information in the form of numbers
- **Data** refers to information that is collected in the form of numbers

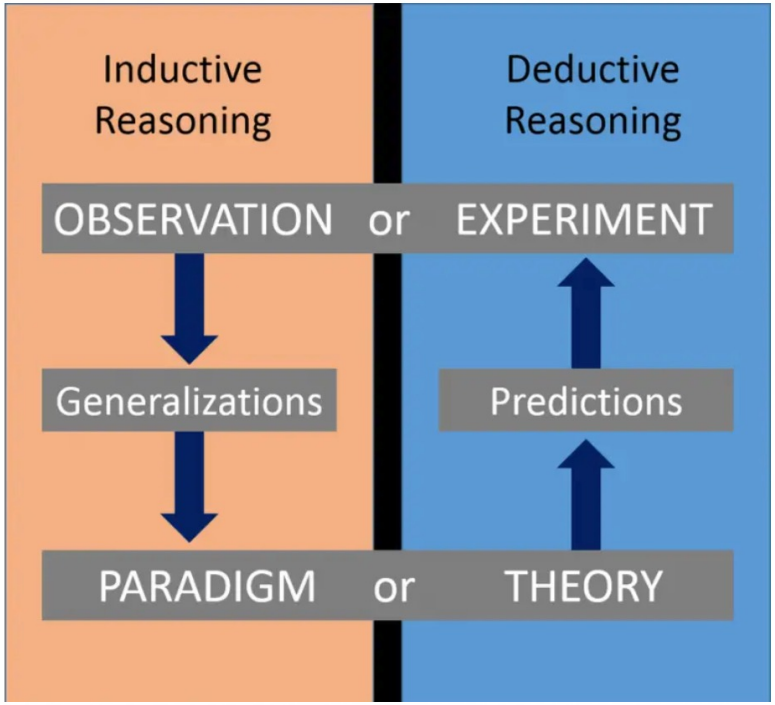
The wheel of science

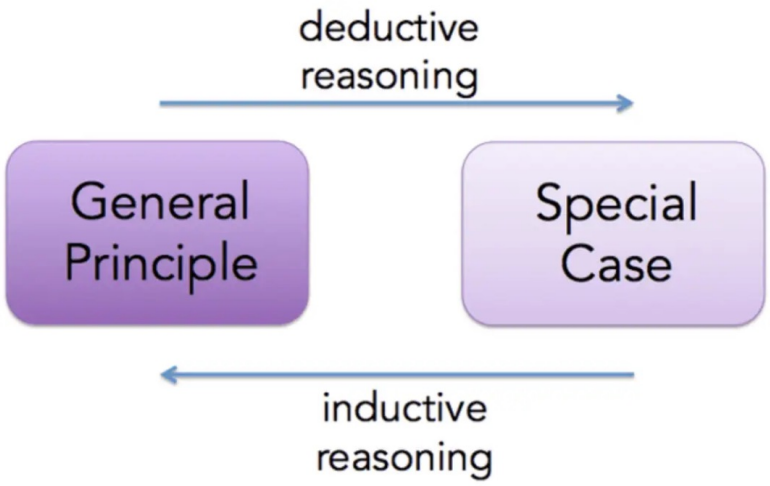
- Scientific theory and research continually shape each other

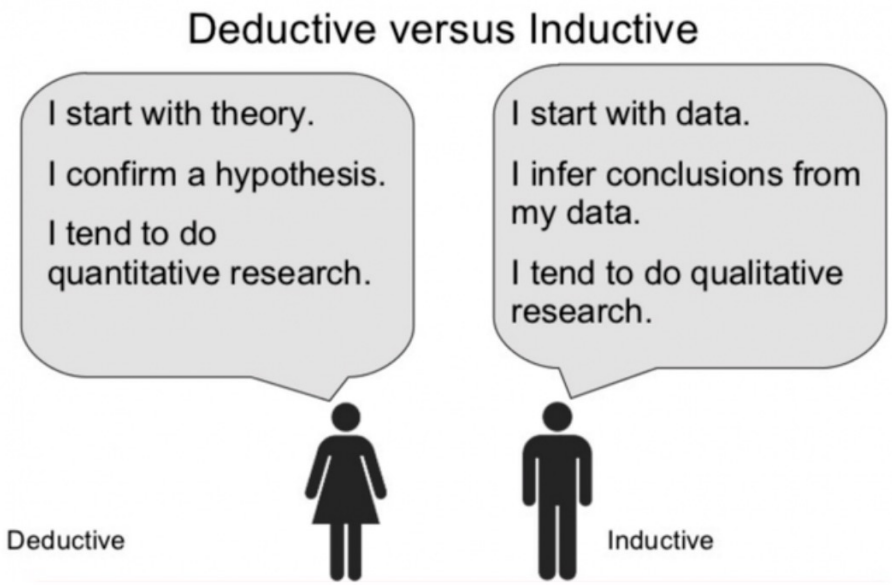


Source: Healey, 2015, p.2.









Source: https://danielmiessler.com/blog/the-difference-between-deductive-and-inductive-reasoning/

Theory

- **Theory** is an explanation of the relationships among social phenomena
- Scientific theory is subject to a rigorous testing process
- Social theories are complex and abstract explanations about problems in society
  - They develop explanations about these issues

Hypotheses

- Since theories are often complex and abstract, we need to be specific to conduct a valid test
- Hypotheses are preliminary answers to research questions, based on theories
- Hypothesis is a specific and exact statement about the relationship between variables...

Variables and observations

- **Variables**
  - Characteristics that can change values from case to case
  - E.g. gender, age, race/ethnicity, number of children, place of residence, income...

- **Observations (cases)**
  - Refer to the entity from which data are collected
  - Also known as ”unit of analysis”
  - E.g. individuals, households, states, countries...

Variables

- **Variable:** a characteristic/phenomenon whose value varies (changes) from case to case, and is empirically quantifiable

- **Dependent variable:** a variable whose variation depends on another variable

- **Independent variable:** a variable whose variation produces ("causes") variation in another variable

Observations

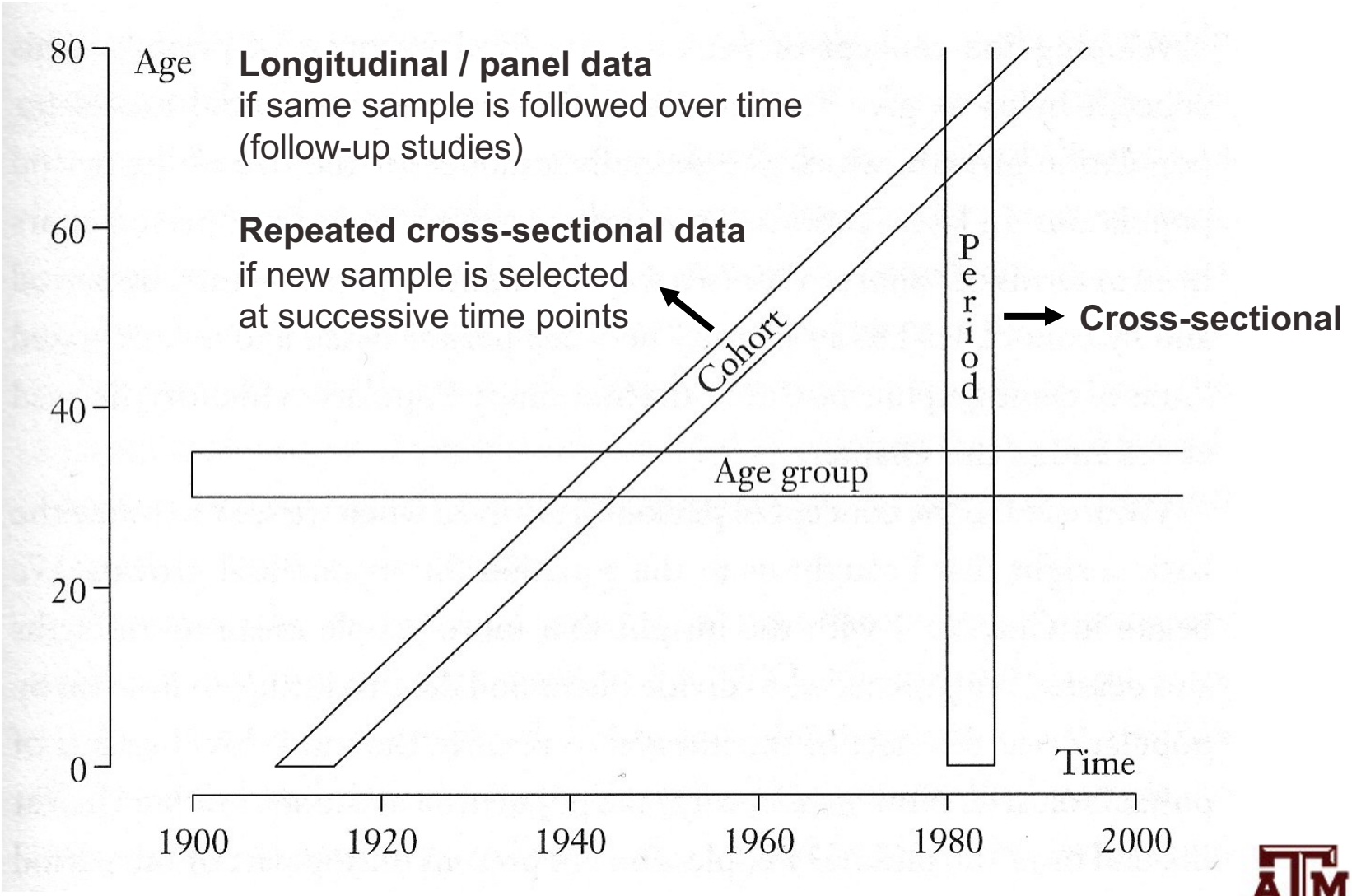
- **Observations** (cases) are collected information used to test hypotheses
- Decide how variables will be measured and how cases will be selected and tested
- Measure social reality: collect numerical data
- Information can be organized in databases
  - Variables as columns
  - Observations as rows

Example of a database

Observation	Salary per hour	Years of schooling	Years of experience in the labor market	Female	Marital status (married)
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
...	...	...	...	...	...
525	11.56	16	5	0	1
526	3.50	14	5	1	0

Source: Wooldridge, 2008.

Lexis diagram



Source: Wachter 2014, p.33.

Empirical generalizations

- **Empirical generalizations** are conclusions based on the analysis of collected observations that evaluate hypotheses and assess theory

- As we developed tentative explanations, we would begin to revise or elaborate the theory that guides the research project
  - If we changed our theory because of our empirical generalizations, a new research project would be needed to test the revised theory
  - The **wheel of science** would begin to turn again

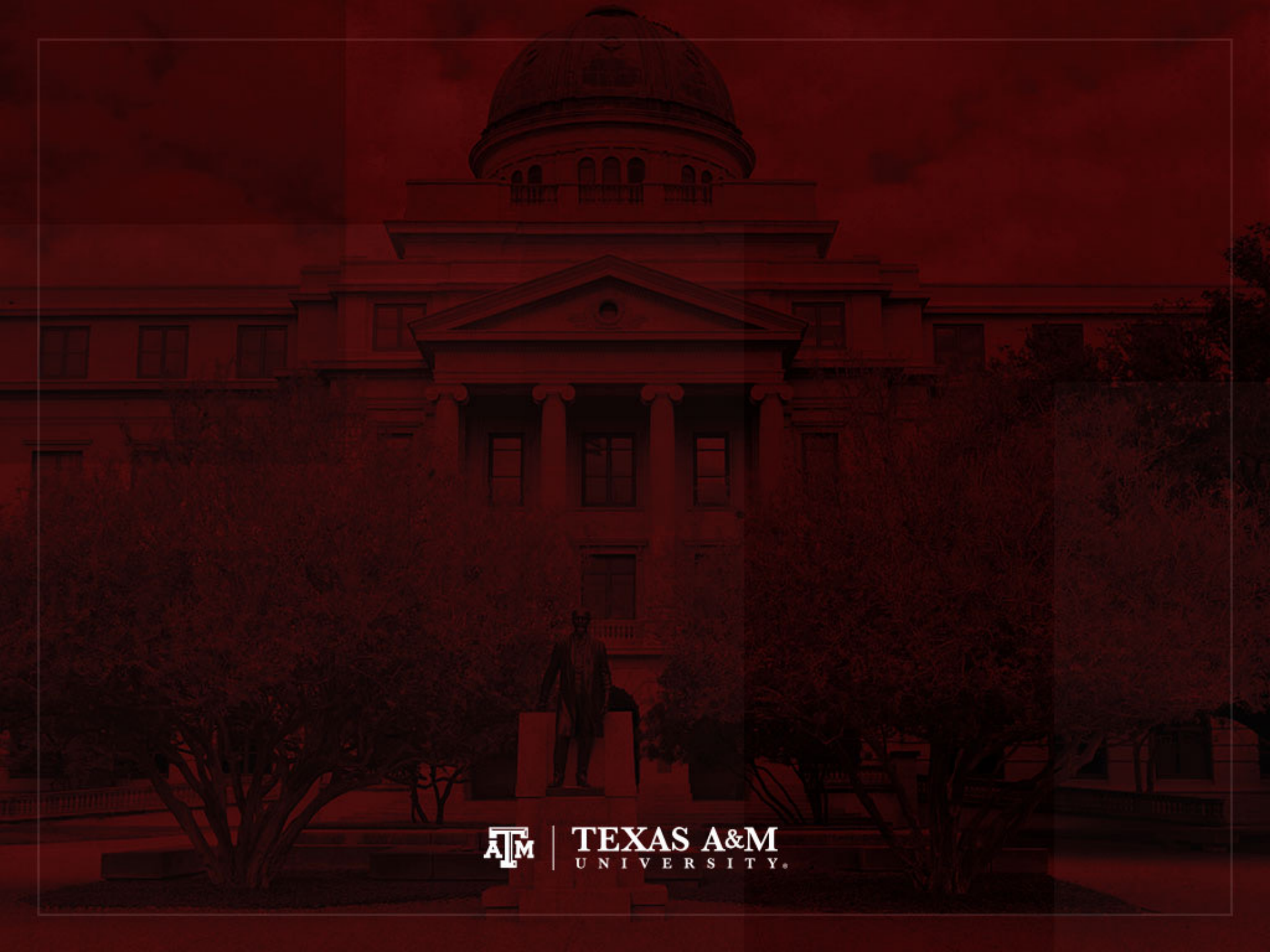
Statistical analysis

- Statistical analysis of data should be applied after successfully completing earlier phases
  - Rigorous conceptualization and use of theory
  - Well-defined research design and methods
  - Well-conceived research questions

- Review research literature to learn how to
  - Develop and clarify definitions
  - Understand social concepts
  - Develop questions and indicators to measure concepts

Theory and research

- In the normal course of science, we rarely are in a position to declare a **theory true or false**
  - Evidence will gradually accumulate over time
  - Ultimate judgments of truth will be the result of many years of research and debate
- **Theory stimulates research and research shapes theory**
  - This is the key to enhance our understanding of the social world
- Statistics is one of the most important links between theory and research


TEXAS A&M
UNIVERSITY

Types of variables

- **Variables** may be classified in different forms
- **Causal relationships**
  - Independent or dependent
- **Unit of measurement**
  - Discrete or continuous
- **Level of measurement**
  - Nominal, ordinal, or interval-ratio

Causation

- Theories and hypotheses are often stated in terms of the **relationships between variables**
  - Causes: independent variables
  - Effects or results: dependent variables

y	x	Use
Dependent variable	Independent variable	Econometrics
Explained variable	Explanatory variable	
Response variable	Control variable	Experimental science
Predicted variable	Predictor variable	
Outcome variable	Covariate	
Regressand	Regressor	

Correlation vs. causation

- Correlation and causation are different
  - Strong associations (correlation) may be used as evidence of causal relationships (causation)
  - Associations do not prove variables are causally related
- We might have problems of reverse causality
  - e.g., immigration increases competition in the labor market and affects earnings
  - Availability of jobs and income levels influence migration

Discrete or continuous

- **Discrete** variables
  - Have a basic unit of measurement that cannot be subdivided (whole numbers)
  - Count number of units (e.g. people, cars, siblings) for each case (e.g. household, person)
- **Continuous** variables
  - Have scores that can be subdivided infinitely (fractional numbers)
  - Report values as if continuous variables were discrete
- Statistics and graphs vary depending on whether variable is discrete or continuous

Level of measurement

- Level of measurement
  - Mathematical nature of the scores of a variable
  - It is crucial because statistical analysis must match the mathematical characteristics of variables
- Three levels of measurement
  - **Nominal:** scores are labels only, not numbers
  - **Ordinal:** scores have some numerical quality and can be ranked
  - **Interval-ratio:** scores are numbers

Nominal-level variables

- Have non-numerical scores or categories
  - Scores are different from each other, but cannot be treated as numbers (they are just labels)
  - Statistical analysis is limited to comparing relative sizes of categories

Variables	Gender	Political party preference	Religious preference
Categories	1 Male	1 Democrat	1 Protestant
	2 Female	2 Republican	2 Catholic
		3 Other	3 Jew
		4 Independent	4 None
			5 Other

Criteria to measure variables

- Be **mutually exclusive**
  - Each case must fit into one and only one category
- Be **exhaustive**
  - There must be a category for every case
- Include elements that are **homogenous**
  - The cases in each category must be similar to each other

Measuring religious affiliation

- Scale A (not mutually exclusive)
  - Protestant and Episcopalian overlap
- Scale B (not exhaustive)
  - Lacks no religion and other
- Scale C (not homogeneous)
  - Non-Protestant seems too broad

Scale A	Scale B	Scale C	Scale D
Protestant	Protestant	Protestant	Protestant
Episcopalian	Catholic	Non-Protestant	Catholic
Catholic	Jew		Jew
Jew			None
None			Other
Other			

Ordinal-level variables

- Categories can be ranked from high to low
  - We can say that one case is higher or lower, more or less than another
- Scores have no absolute or objective meaning
  - Only represent position with respect to other scores
  - We can distinguish between high and low scores
  - But distance between scores cannot be described
  - Average is not permitted with ordinal-level variables

Examples: ordinal-level variables

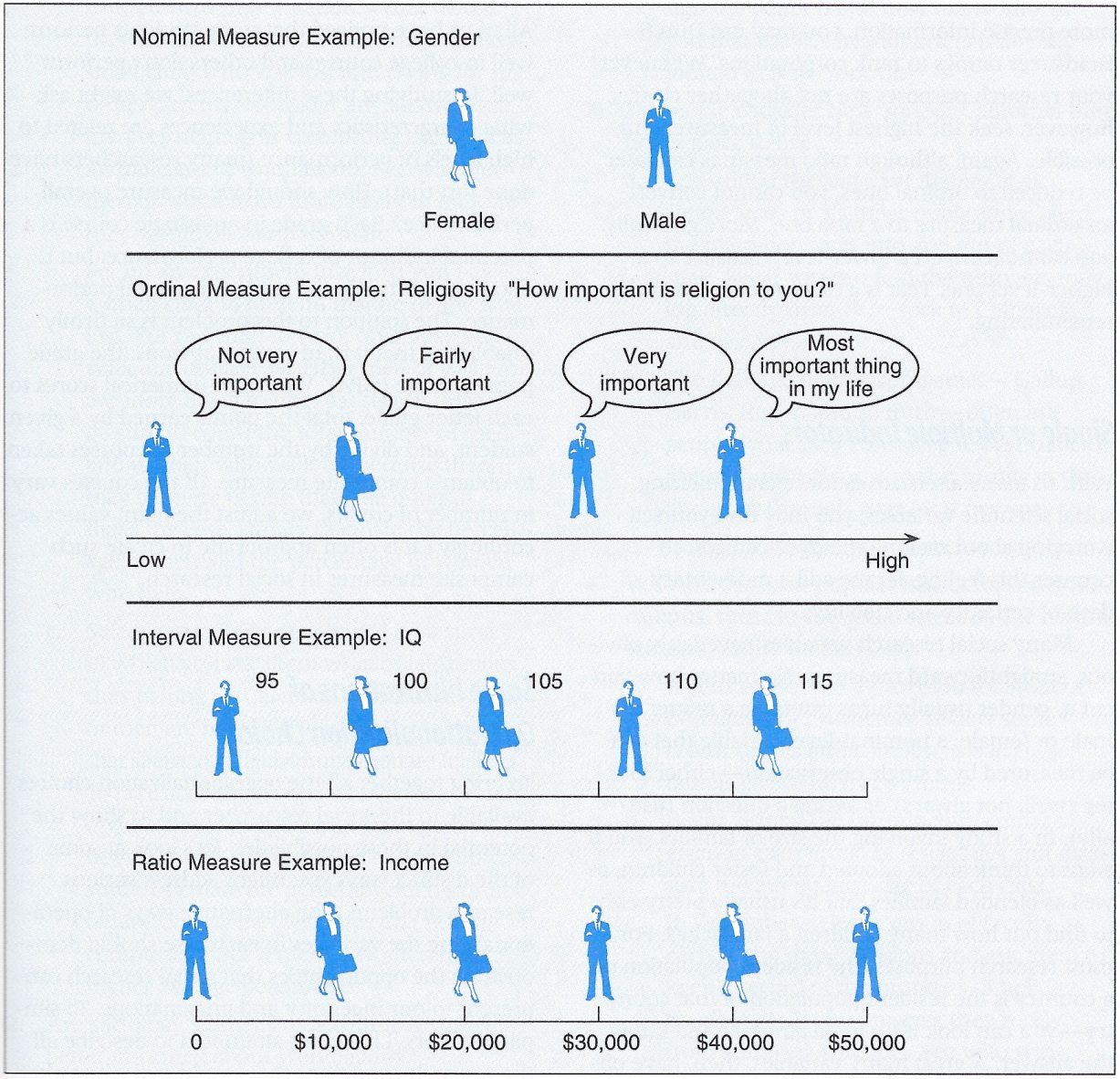
- Attitude and opinion scales
  - Prejudice, alienation, political conservatism...
- Likert scale:
  - (1) strongly disagree; (2) disagree; (3) neither agree nor disagree; (4) agree; (5) strongly agree
- Into which of the following classes would you say you belong?

Score	Class
1	Lower class
2	Working class
3	Middle class
4	Upper class

Interval-ratio-level variables

- Scores are actual numbers that can be analyzed with all possible statistical techniques
- Have equal intervals between scores
- Have true zero points
  - Score of zero is not arbitrary
  - It indicates absence of whatever is being measured
- Examples:
  - Age (in years)
  - Income (in dollars)
  - Year of education
  - Number of children

Examples



Source: Babbie 2001, p.137.

Importance

- Level of measurement of a variable is crucial
  - It tells us which statistics are appropriate and useful
- Different statistics require different mathematical operations
  - Ranking, addition, square root...
- The first step in dealing with a variable and selecting appropriate statistics is to determine its level of measurement

Determine level of measurement

- Change the order of the scores. Do they still make sense?
  - If yes: the variable is **nominal**
  - If no: proceed to the next step

- Is the distance between the scores unequal?
  - If yes: the variable is **ordinal**
  - If no: the variable is **interval-ratio**

Nominal- and ordinal-level

- Nominal-level (e.g. marital status) and ordinal-level (e.g. capital punishment support) variables are almost always **discrete**

What is your marital status? Are you presently:		**Do you support the death penalty for persons convicted of homicide?**	
Score	**Category**	**Score**	**Category**
1	Married	1	Strongly support
2	Divorced	2	Somewhat support
3	Separated	3	Neither support nor oppose
4	Widowed	4	Somewhat oppose
5	Single	5	Strongly oppose

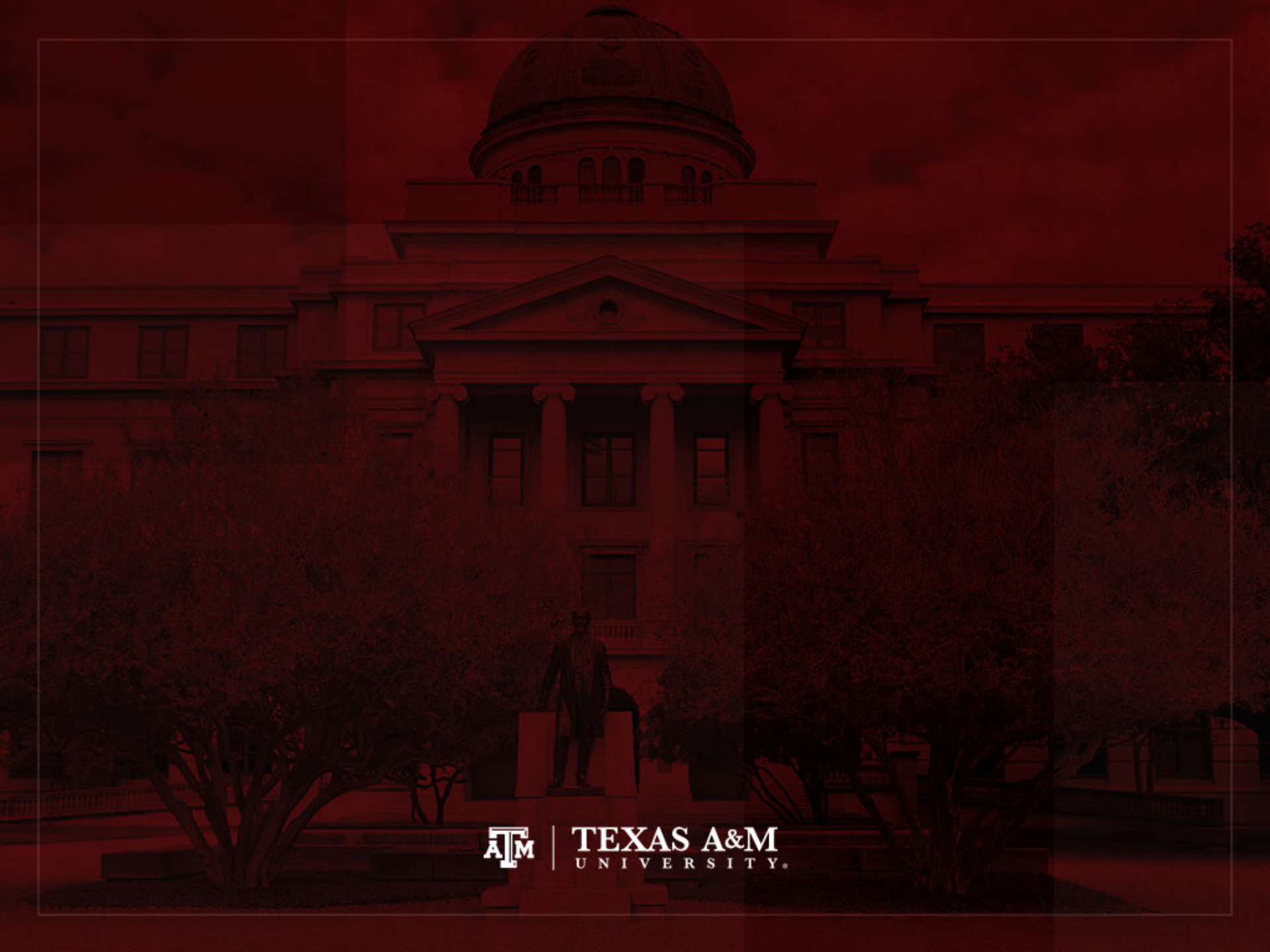
Income at the ordinal level

- Always examine the way in which the scores of the variable are actually stated
  - Be careful to look at the way in which the variable is measured before defining its level of measurement
- This is a problem with interval-ratio variables that have been measured at the ordinal level

Score	Income range
1	Less than $24,999
2	$25,000 to $49,999
3	$50,000 to $99,999
4	$100,000 or more

Variables' level of measurement

Variables' level of measurement	Examples of variables	Measurement procedures	Mathematical operations permitted	Examples of available techniques
Nominal	– Gender – Race/ethnicity – Religion – Marital status	– Classification into categories – Mode	– Counting number in each category (tabulation) – Comparing sizes of categories	– Chi Square – Logistic regression – Multinomial logistic regression
Ordinal	– Social class – Attitude scales – Opinion scales	– All of the above – Plus ranking of categories with respect to each other (scale) – Mode, median	– All of the above – Plus judgments of "greater than" and "less than"	– Spearman's Rho – Ordered logistic regression
Interval-ratio	– Age – Number of children – Income	– All of the above – Plus description of scores in terms of equal units – Mode, median, mean	– All of the above – Plus mathematical operations (addition, subtraction, multiplication, division, square roots...)	– Scatterplots – Pearson's r – Analysis of variance (ANOVA) – Ordinary least square regression (linear regression)


TEXAS A&M
UNIVERSITY

General classes of statistics

- Two main types of statistical techniques are available to analyze data and answer questions
- Descriptive statistics
- Inferential statistics

Descriptive statistics

- **Univariate** descriptive statistics
  - Summarize or describe the distribution of a single variable
- **Bivariate** descriptive statistics
  - Describe the relationship between two variables
- **Multivariate** descriptive statistics
  - Describe the relationship among three or more variables

Univariate descriptive statistics

- **Univariate descriptive** statistics
  - Include percentages, averages, and graphs
  - Data reduction: few numbers summarize many
- **U.S. population by age groups, 2010**

Age group	Percent
Under 18 years	24.0
18 to 44 years	36.6
45 to 64 years	26.4
65+ years	13.0
Total (N)	**308,745,538**

  - The median age was 37.2 years in 2010

Source: Census Bureau (https://www.census.gov/newsroom/releases/archives/2010_census/cb11-cn147.html).

Bivariate descriptive statistics

- **Bivariate descriptive** statistics
  - Describe the strength and direction of the relationship between two variables
  - **Measures of association:** quantify the strength and direction of a relationship
  - Allow us to investigate causation and prediction
- E.g. relationship between **study time and grade**
  - Strength: closely related
  - Direction: as one increases, the other also increases
  - Prediction: the longer the study time, the higher the grade

Multivariate descriptive statistics

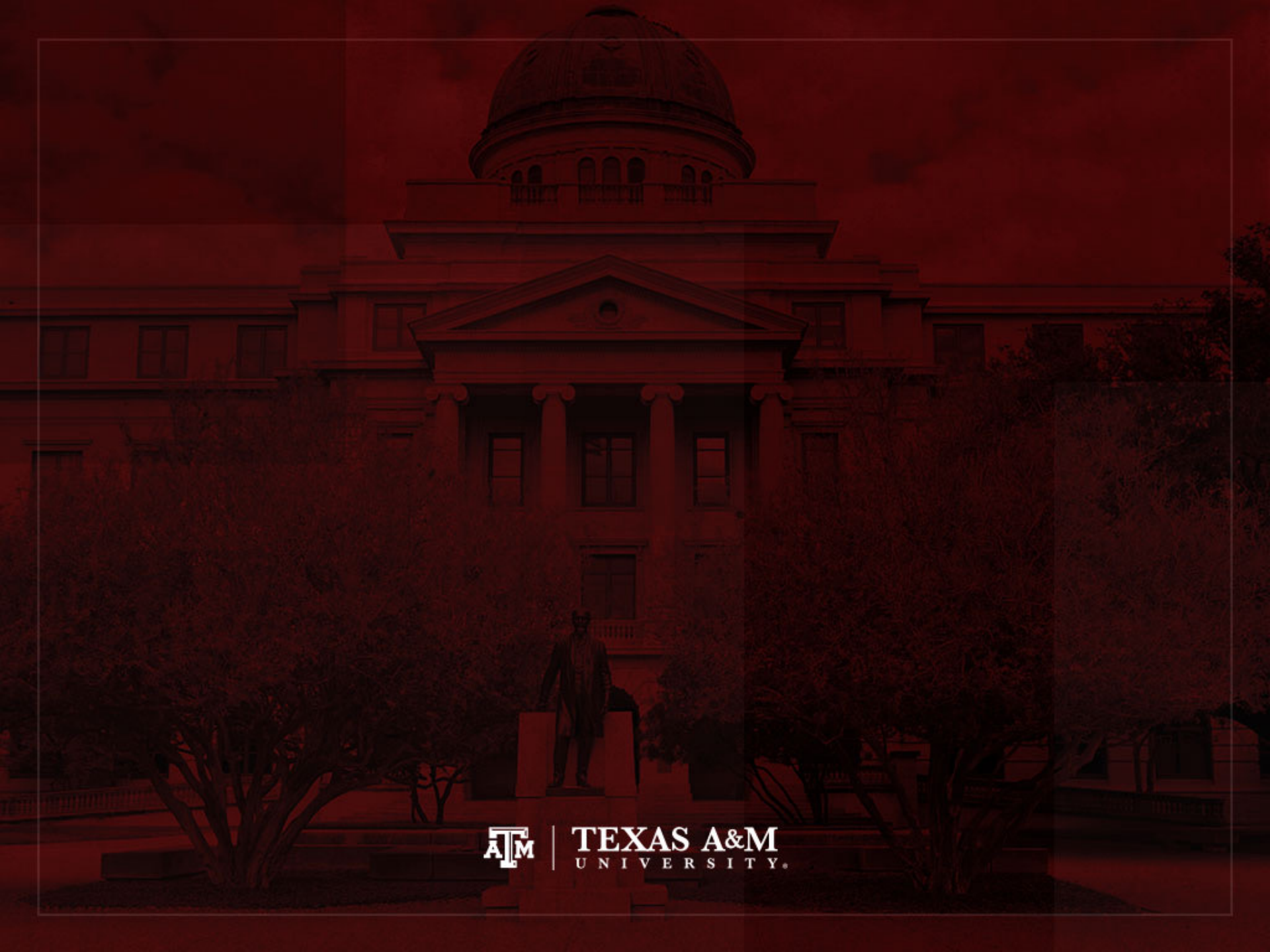
- **Multivariate descriptive** statistics
  - Describe the relationships between three or more variables
  - **Measures of association:** quantify the strength and direction of a multivariate relationship
- E.g. **grade, age, gender**
  - Strength: relationship between age and grade is strong for women, but weak for men
  - Direction: grades increase with age only for females
  - Prediction: older females will experience higher grades than younger females. Older males will have similar grades to younger males.

Inferential statistics

- Social scientists need inferential statistics
  - They almost never have the resources or time to collect data from every case in a population
- Inferential statistics uses data from samples to make generalizations about populations
  - **Population** is the total collection of all cases in which the researcher is interested
  - **Samples** are carefully chosen subsets of the population
- With proper techniques, generalizations based on samples can represent populations

Public-opinion polls

- **Public-opinion polls** and election projections are a familiar application of inferential statistics
  - Several thousand carefully selected voters are interviewed about their voting intentions
  - This information is used to estimate the intentions of all voters (millions of people)
- E.g. public-opinion poll reports that 42% of voters plans to vote for a certain candidate
  - 2,000 respondents are used to generalize to the American electorate population (130 million people)


TEXAS A&M
UNIVERSITY

IPUMS

- Integrated Public Use Microdata Series (https://ipums.org)
  - Provides census and survey data from around the world integrated across time and space
  - Minnesota Population Center (https://www.pop.umn.edu)
  - Steven Ruggles (http://users.hist.umn.edu/~ruggles)
- IPUMS USA provides access to over 60 integrated, high-precision samples of the American population
  - Federal censuses
  - American Community Survey (ACS): 2000-present
  - Puerto Rican Community Survey (PRCS): 2005-present
  - Assigns uniform codes across all the samples and brings relevant documentation into a coherent form to facilitate analysis of social and economic change

Source: https://usa.ipums.org/usa/intro.shtml

2010 Decennial Census

- The 2010 Decennial Census consisted of a single short-form questionnaire
  - The short form asked age, sex, race, ethnicity, relationship to household head, and whether the housing unit was rented or owned by a member of the household

- The annual ACS survey was designed to replace the Census long-form questionnaire
  - The ACS/PRCS sample design approximates the Census 2000 long-form sample design and oversamples areas with smaller populations

Source: https://usa.ipums.org/usa/chapter2/chapter2.shtml

American Community Survey

- ACS and PRCS samples include about 3 million households nationwide
  - The sampling unit is the household and all persons residing in the household
- IPUMS samples of ACS and PRCS come from the Census Bureau's larger internal data files
  - They are subject to additional sampling error and further data processing (e.g., imputation, allocation)
  - Estimates from ACS IPUMS may not be consistent with ACS summary tables

Source: https://usa.ipums.org/usa/chapter2/chapter2.shtml

Confidentiality measures

- Measures to protect individual confidentiality in ACS public available data
  - Individual variables, such as income and housing values are top coded
  - Geographic identifiers are currently restricted to the state and PUMA levels
- Public use microdata area (PUMA)
  - Consist of 100,000+ residents
  - Do not cross state lines
  - Codes must be combined with state codes
  - 2,101 PUMAs in the 2005–2011 ACS
  - 2,378 PUMAs in the 2012–2019 ACS

Source: https://usa.ipums.org/usa/chapter2/chapter2.shtml

13197017

U.S. DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. CENSUS BUREAU

THE American Community Survey

This booklet shows the content of the American Community Survey questionnaire.

Start Here

Respond online today at:
https://respond.census.gov/acs
OR
Complete this form and mail it back as soon as possible.

This form asks for information about the people who are living or staying at the address on the mailing label and about the house, apartment, or mobile home located at the address on the mailing label.

If you need help or have questions about completing this form, please call **1-800-354-7271.** The telephone call is free.

Telephone Device for the Deaf (TDD):
Call 1–800–582–8330. The telephone call is free.

¿NECESITA AYUDA? Si usted habla español y necesita ayuda para completar su cuestionario, llame sin cargo alguno al **1-877-833-5625.** Usted también puede completar su entrevista por teléfono con un entrevistador que habla español. O puede responder por Internet en: https://respond.census.gov/acs

For more information about the American Community Survey, visit our web site at: http://www.census.gov/acs

Please print today's date.
Month Day Year

Please print the name and telephone number of the person who is filling out this form. We will only contact you if needed for official Census Bureau business.
Last Name

First Name MI

Area Code + Number

How many people are living or staying at this address?
- **INCLUDE** everyone who is living or staying here for more than 2 months.
- **INCLUDE** yourself if you are living here for more than 2 months.
- **INCLUDE** anyone else staying here who does not have another place to stay, even if they are here for 2 months or less.
- **DO NOT INCLUDE** anyone who is living somewhere else for more than 2 months, such as a college student living away or someone in the Armed Forces on deployment.

Number of people

Fill out pages 2, 3, and 4 for everyone, including yourself, who is living or staying at this address for more than 2 months. Then complete the rest of the form.

FORM **ACS-1(INFO)(2017)**
(03-14-2016)

OMB No. 0607-0810
OMB No. 0607-0936

Person 1

(Person 1 is the person living or staying here in whose name this house or apartment is owned, being bought, or rented. If there is no such person, start with the name of any adult living or staying here.)

1 What is Person 1's name?
Last Name *(Please print)* | First Name | MI

2 How is this person related to Person 1?
[X] Person 1

3 What is Person 1's sex? *Mark (X) ONE box.*
[ ] Male [ ] Female

4 What is Person 1's age and what is Person 1's date of birth?
Please report babies as age 0 when the child is less than 1 year old.
Print numbers in boxes.
Age (in years) | Month | Day | Year of birth

➔ **NOTE: Please answer BOTH Question 5 about Hispanic origin and Question 6 about race. For this survey, Hispanic origins are not races.**

5 Is Person 1 of Hispanic, Latino, or Spanish origin?
[ ] **No,** not of Hispanic, Latino, or Spanish origin
[ ] Yes, Mexican, Mexican Am., Chicano
[ ] Yes, Puerto Rican
[ ] Yes, Cuban
[ ] Yes, another Hispanic, Latino, or Spanish origin – *Print origin, for example, Argentinean, Colombian, Dominican, Nicaraguan, Salvadoran, Spaniard, and so on.*

6 What is Person 1's race? *Mark (X) one or more boxes.*
[ ] White
[ ] Black or African Am.
[ ] American Indian or Alaska Native — *Print name of enrolled or principal tribe.*
[ ] Asian Indian
[ ] Chinese
[ ] Filipino
[ ] Other Asian – *Print race, for example, Hmong, Laotian, Thai, Pakistani, Cambodian, and so on.*
[ ] Japanese
[ ] Korean
[ ] Vietnamese
[ ] Native Hawaiian
[ ] Guamanian or Chamorro
[ ] Samoan
[ ] Other Pacific Islander – *Print race, for example, Fijian, Tongan, and so on.*
[ ] Some other race – *Print race.*

2 [barcode]

Person 1

➔ **Please copy the name of Person 1 from page 2, then continue answering questions below.**
Last Name
First Name | MI

7 Where was this person born?
[ ] In the United States – *Print name of state.*
[ ] Outside the United States – *Print name of foreign country, or Puerto Rico, Guam, etc.*

8 Is this person a citizen of the United States?
[ ] Yes, born in the United States ➔ *SKIP to question 10a*
[ ] Yes, born in Puerto Rico, Guam, the U.S. Virgin Islands, or Northern Marianas
[ ] Yes, born abroad of U.S. citizen parent or parents
[ ] Yes, U.S. citizen by naturalization – *Print year of naturalization*
[ ] No, not a U.S. citizen

9 When did this person come to live in the United States? *If this person came to live in the United States more than once, print latest year.*
Year

10 a. At any time IN THE LAST 3 MONTHS, has this person attended school or college? *Include only nursery or preschool, kindergarten, elementary school, home school, and schooling which leads to a high school diploma or a college degree.*
[ ] No, has not attended in the last 3 months ➔ *SKIP to question 11*
[ ] Yes, public school, public college
[ ] Yes, private school, private college, home school

b. What grade or level was this person attending? *Mark (X) ONE box.*
[ ] Nursery school, preschool
[ ] Kindergarten
[ ] Grade 1 through 12 – *Specify grade 1 – 12*
[ ] College undergraduate years (freshman to senior)
[ ] Graduate or professional school beyond a bachelor's degree *(for example: MA or PhD program, or medical or law school)*

11 What is the highest degree or level of school this person has COMPLETED? *Mark (X) ONE box. If currently enrolled, mark the previous grade or highest degree received.*

NO SCHOOLING COMPLETED
[ ] No schooling completed

NURSERY OR PRESCHOOL THROUGH GRADE 12
[ ] Nursery school
[ ] Kindergarten
[ ] Grade 1 through 11 – *Specify grade 1 – 11*
[ ] 12th grade – **NO DIPLOMA**

HIGH SCHOOL GRADUATE
[ ] Regular high school diploma
[ ] GED or alternative credential

COLLEGE OR SOME COLLEGE
[ ] Some college credit, but less than 1 year of college credit
[ ] 1 or more years of college credit, no degree
[ ] Associate's degree *(for example: AA, AS)*
[ ] Bachelor's degree *(for example: BA, BS)*

AFTER BACHELOR'S DEGREE
[ ] Master's degree *(for example: MA, MS, MEng, MEd, MSW, MBA)*
[ ] Professional degree beyond a bachelor's degree *(for example: MD, DDS, DVM, LLB, JD)*
[ ] Doctorate degree *(for example: PhD, EdD)*

F *Answer question 12 if this person has a bachelor's degree or higher. Otherwise, SKIP to question 13.*

12 This question focuses on this person's BACHELOR'S DEGREE. Please print below the specific major(s) of any BACHELOR'S DEGREES this person has received. *(For example: chemical engineering, elementary teacher education, organizational psychology)*

13 What is this person's ancestry or ethnic origin?
(For example: Italian, Jamaican, African Am., Cambodian, Cape Verdean, Norwegian, Dominican, French Canadian, Haitian, Korean, Lebanese, Polish, Nigerian, Mexican, Taiwanese, Ukrainian, and so on.)

14 a. Does this person speak a language other than English at home?
[ ] Yes
[ ] No ➔ *SKIP to question 15a*

b. What is this language?
For example: Korean, Italian, Spanish, Vietnamese

c. How well does this person speak English?
[ ] Very well
[ ] Well
[ ] Not well
[ ] Not at all

15 a. Did this person live in this house or apartment 1 year ago?
[ ] Person is under 1 year old ➔ *SKIP to question 16*
[ ] Yes, this house ➔ *SKIP to question 16*
[ ] No, outside the United States and Puerto Rico – *Print name of foreign country, or U.S. Virgin Islands, Guam, etc., below; then SKIP to question 16*
[ ] No, different house in the United States or Puerto Rico

b. Where did this person live 1 year ago?
Address (Number and street name)
Name of city, town, or post office
Name of U.S. county or municipio in Puerto Rico
Name of U.S. state or Puerto Rico | **ZIP Code**

8 [barcode]

13197058

Housing

Please answer the following questions about the house, apartment, or mobile home at the address on the mailing label.

1 Which best describes this building? *Include all apartments, flats, etc., even if vacant.*

- ☐ A mobile home
- ☐ A one-family house detached from any other house
- ☐ A one-family house attached to one or more houses
- ☐ A building with 2 apartments
- ☐ A building with 3 or 4 apartments
- ☐ A building with 5 to 9 apartments
- ☐ A building with 10 to 19 apartments
- ☐ A building with 20 to 49 apartments
- ☐ A building with 50 or more apartments
- ☐ Boat, RV, van, etc.

2 About when was this building first built?

- ☐ 2000 or later – *Specify year* ↙ ☐☐☐☐
- ☐ 1990 to 1999
- ☐ 1980 to 1989
- ☐ 1970 to 1979
- ☐ 1960 to 1969
- ☐ 1950 to 1959
- ☐ 1940 to 1949
- ☐ 1939 or earlier

3 When did PERSON 1 (listed on page 2) move into this house, apartment, or mobile home?

Month ☐☐ Year ☐☐☐☐

A *Answer questions 4 – 5 if this is a HOUSE OR A MOBILE HOME; otherwise, SKIP to question 6a.*

4 How many acres is this house or mobile home on?

- ☐ Less than 1 acre ➔ *SKIP to question 6a*
- ☐ 1 to 9.9 acres
- ☐ 10 or more acres

5 IN THE PAST 12 MONTHS, what were the actual sales of all agricultural products from this property?

- ☐ None
- ☐ $1 to $999
- ☐ $1,000 to $2,499
- ☐ $2,500 to $4,999
- ☐ $5,000 to $9,999
- ☐ $10,000 or more

6 a. How many separate rooms are in this house, apartment, or mobile home? *Rooms must be separated by built-in archways or walls that extend out at least 6 inches and go from floor to ceiling.*

- *INCLUDE bedrooms, kitchens, etc.*
- *EXCLUDE bathrooms, porches, balconies, foyers, halls, or unfinished basements.*

Number of rooms ☐☐

b. How many of these rooms are bedrooms? *Count as bedrooms those rooms you would list if this house, apartment, or mobile home were for sale or rent. If this is an efficiency/studio apartment, print "0".*

Number of bedrooms ☐☐

7 Does this house, apartment, or mobile home have –

	Yes	No
a. hot and cold running water?	☐	☐
b. a bathtub or shower?	☐	☐
c. a sink with a faucet?	☐	☐
d. a stove or range?	☐	☐
e. a refrigerator?	☐	☐
f. telephone service from which you can both make and receive calls? *Include cell phones.*	☐	☐

8 At this house, apartment, or mobile home – do you or any member of this household own or use any of the following types of computer?

	Yes	No
a. Desktop or laptop	☐	☐
b. Smartphone	☐	☐
c. Tablet or other portable wireless computer	☐	☐
d. Some other type of computer *Specify* ↙	☐	☐

9 At this house, apartment, or mobile home – do you or any member of this household have access to the Internet?

- ☐ Yes, by paying a cell phone company or Internet service provider
- ☐ Yes, without paying a cell phone company or Internet service provider ➔ *SKIP to question 11*
- ☐ No access to the Internet at this house, apartment, or mobile home ➔ *SKIP to question 11*

10 Do you or any member of this household have access to the Internet using a –

	Yes	No
a. cellular data plan for a smartphone or other mobile device?	☐	☐
b. broadband (high speed) Internet service such as cable, fiber optic, or DSL service installed in this household?	☐	☐
c. satellite Internet service installed in this household?	☐	☐
d. dial-up Internet service installed in this household?	☐	☐
e. some other service? *Specify service* ↙	☐	☐

Housing (continued)

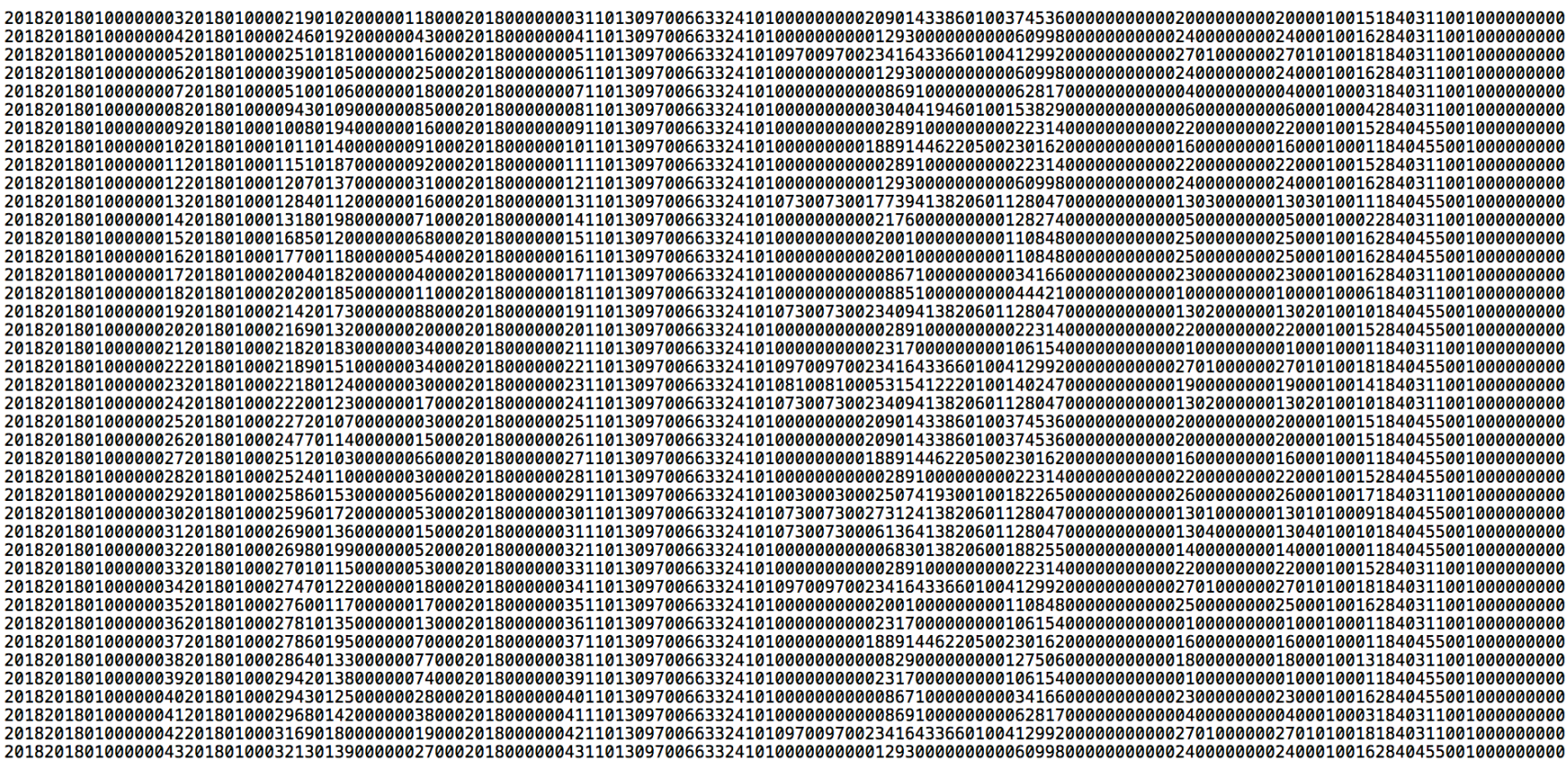
11 How many automobiles, vans, and trucks of one-ton capacity or less are kept at home for use by members of this household?

- ☐ None
- ☐ 1
- ☐ 2
- ☐ 3
- ☐ 4
- ☐ 5
- ☐ 6 or more

12 Which FUEL is used MOST for heating this house, apartment, or mobile home?

- ☐ Gas: from underground pipes serving the neighborhood
- ☐ Gas: bottled, tank, or LP
- ☐ Electricity
- ☐ Fuel oil, kerosene, etc.
- ☐ Coal or coke
- ☐ Wood
- ☐ Solar energy
- ☐ Other fuel
- ☐ No fuel used

ACS raw microdata



ACS codebook

Variable: "YEAR"

Name:	YEAR
Label:	Census year
Variable Text:	YEAR reports the four-digit year when the household was enumerated or included in the census, the ACS, and the PRCS. For the multi-year ACS/PRCS samples, YEAR indicates the last year of data included (e.g., 2007 for the 2005-2007 3-year ACS/PRCS; 2008 for the 2006-2008 3-year ACS/PRCS; and so on). For the actual year of survey in these multi-year data, see MULTYEAR.
Concept:	Technical Variables -- HOUSEHOLD
Start Position:	1
End Position:	4
Width:	4
Variable Format:	numeric
Implied Decimal Places:	0

Variable: "SAMPLE"

Name:	SAMPLE
Label:	IPUMS sample identifier
Variable Text:	SAMPLE identifies the IPUMS sample from which the case is drawn. Each sample receives a unique 6-digit code. The codes are structured as follows: The first four digits are the year of the census/survey. The next two digits identify the sample within the year. For most censuses, IPUMS has multiple datasets which were constructed using different sampling techniques (i.e. size/demographic of the sample population, geographic coverage level or location, or duration of the sampling period for the ACS/PRCS samples). The availability table for each variable indicates whether that variable is available in only certain samples for a given year. For further discussion of sample differences, see "Sample Designs." [URL omitted from DDI.]. Note: SAMPLE replaces DATANUM. Though the last two digits in SAMPLE do not correlate exactly with the now-deprecated DATANUM, the variable serves the same purpose of assigning a unique id to all cases that belong to the same dataset.
Concept:	Technical Variables -- HOUSEHOLD
Start Position:	5
End Position:	10
Width:	6
Variable Format:	numeric
Implied Decimal Places:	0

ACS codebook

Variable: "SEX"

Name:	SEX
Label:	Sex
Variable Text:	SEX reports whether the person was male or female.
Concept:	Demographic Variables -- PERSON
Start Position:	340
End Position:	340
Width:	1
Variable Format:	numeric
Implied Decimal Places:	0
Categories	

Value	Label
1	Male
2	Female

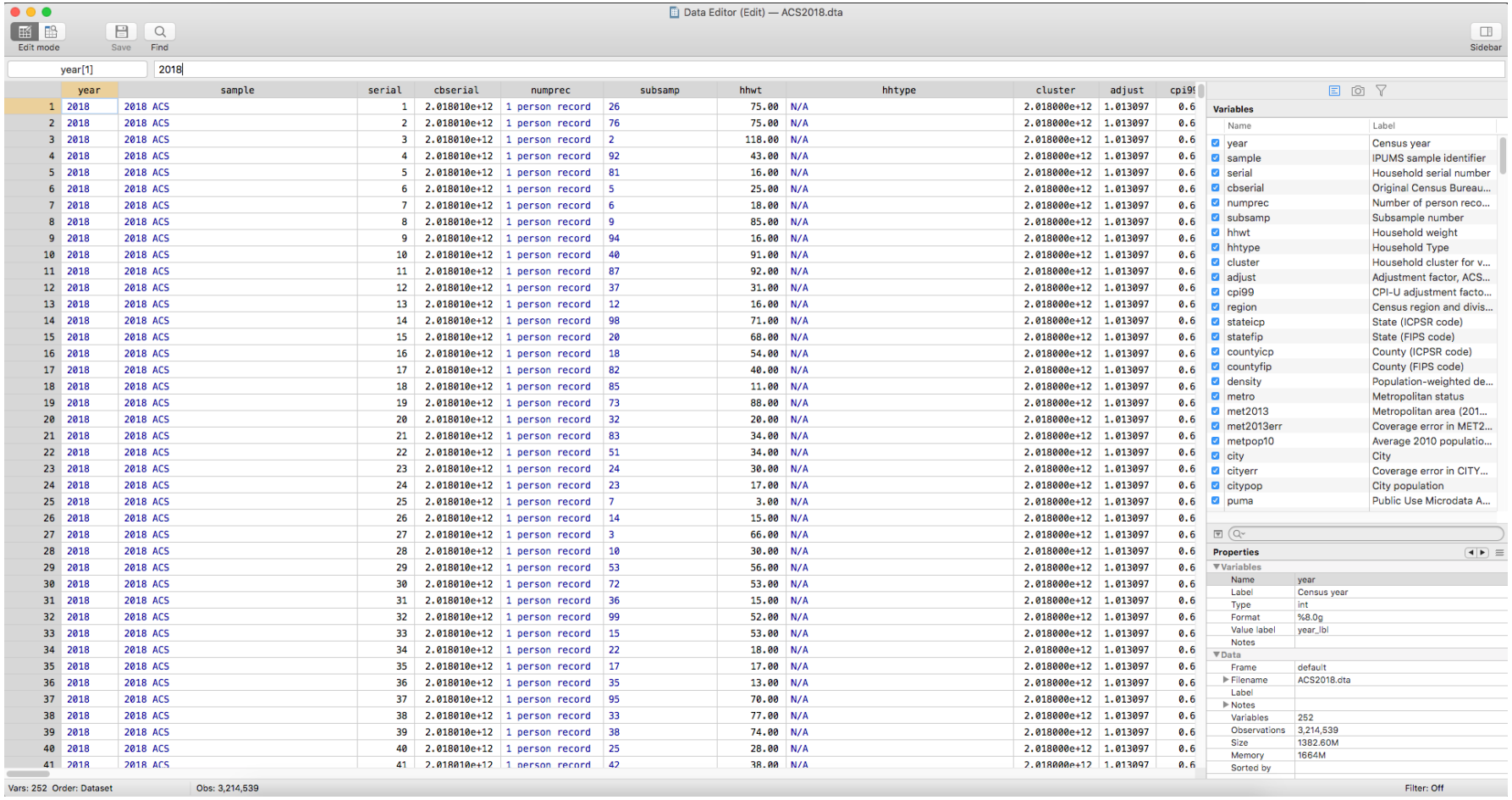
Variable: "AGE"

Name:	AGE
Label:	Age
Variable Text:	AGE reports the person's age in years as of the last birthday. Please see the Comparability section regarding a known Universe issue with AGE and AGEORIG which effects EMPSTAT and LABFORCE for the 2004 ACS Sample.
Concept:	Demographic Variables -- PERSON
Start Position:	341
End Position:	343
Width:	3
Variable Format:	numeric
Implied Decimal Places:	0

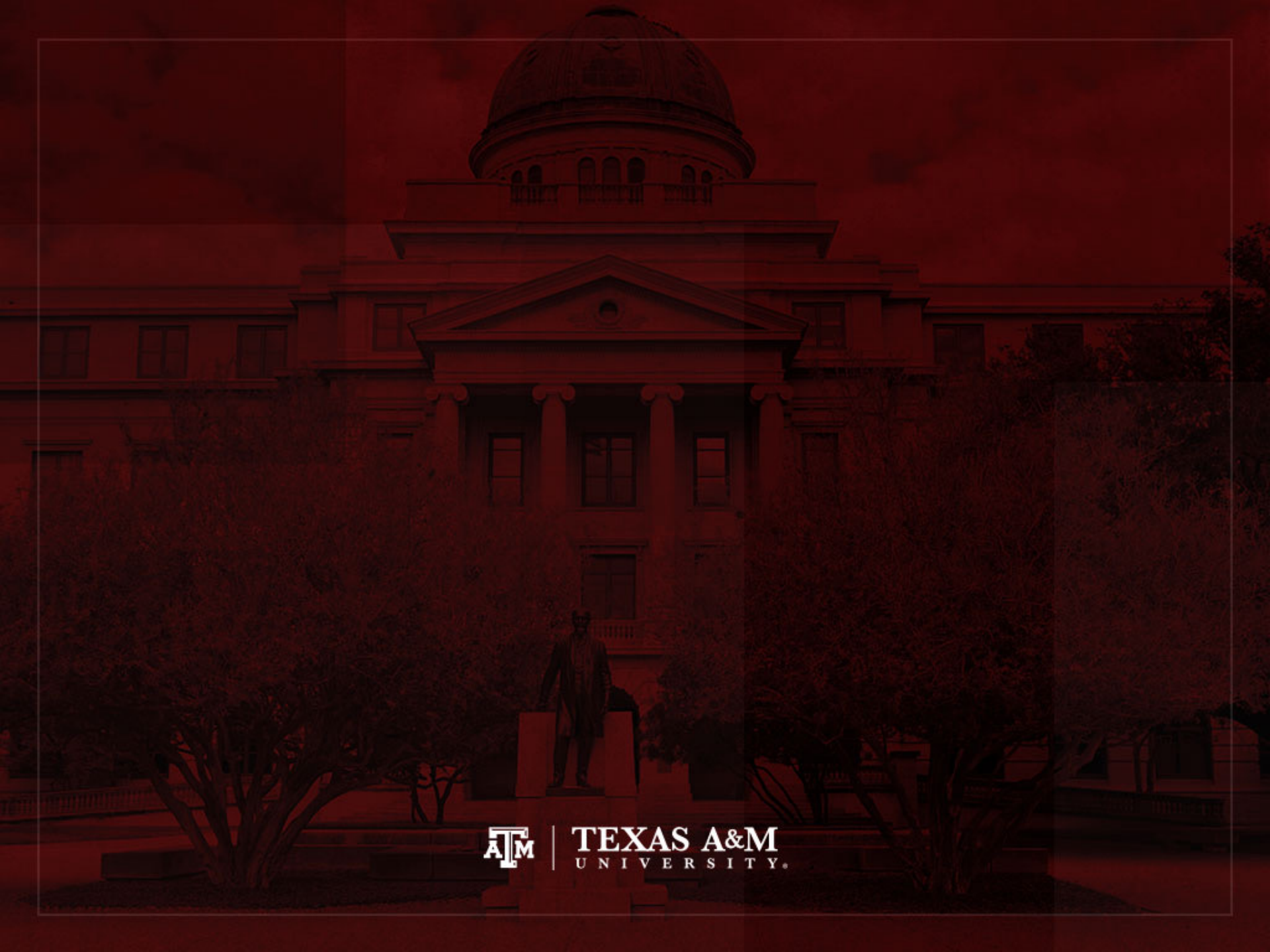
Stata command file from IPUMS

```
* NOTE: You need to set the Stata working directory to the path
* where the data file is located.

set more off

clear
quietly infix                   ///
  int     year          1-4     ///
  long    sample        5-10    ///
  double  serial        11-18   ///
  double  cbserial      19-31   ///
  byte    numprec       32-33   ///
  byte    subsamp       34-35   ///
  double  hhwt          36-45   ///
  byte    hhtype        46-46   ///
  double  cluster       47-59   ///
  double  adjust        60-66   ///
  double  cpi99         67-71   ///
  byte    region        72-73   ///
  byte    stateicp      74-75   ///
  byte    statefip      76-77   ///
  int     countyicp     78-81   ///
  int     countyfip     82-84   ///
  double  density       85-91   ///
  byte    metro         92-92   ///
  long    met2013       93-97   ///
  byte    met2013err    98-98   ///
  double  metpop10      99-106  ///
  int     city          107-110 ///
  byte    cityerr       111-111 ///
  long    citypop       112-116 ///
  long    puma          117-121 ///
  double  strata        122-133 ///
  int     cpuma0010     134-137 ///
  byte    homeland      138-138 ///
  int     cntry         139-141 ///
  byte    gq            142-142 ///
  byte    gqtype        143-143 ///
  int     gqtyped       144-146 ///
  byte    farm          147-147 ///
  byte    ownershp      148-148 ///
  byte    ownershpd     149-150 ///
  byte    mortgage      151-151 ///
  byte    mortgag2      152-152 ///
  byte    farmprod      153-153 ///
  byte    acrehous      154-154 ///
  long    mortamt1      155-159 ///
  int     mortamt2      160-163 ///
  byte    taxincl       164-164 ///
  byte    insincl       165-165 ///
  int     propinsr      166-169 ///
  byte    proptx99      170-171 ///
  long    owncost       172-176 ///
  int     rent          177-180 ///
  int     rentgrs       181-184 ///
  byte    rentmeal      185-185 ///
  int     condofee      186-189 ///
  long    moblhome      190-194 ///
  int     costelec      195-198 ///
  int     costgas       199-202 ///
  int     costwatr      203-206 ///
  int     costfuel      207-210 ///
  long    hhincome      211-217 ///
  byte    foodstmp      218-218 ///
  long    valueh        219-225 ///
  byte    gcmonths      624-624 ///
  byte    gcrespon      625-625 ///
  using `"usa_00070.dat"'

replace hhwt        = hhwt        / 100
replace adjust      = adjust      / 1000000
replace cpi99       = cpi99       / 1000
replace density     = density     / 10
replace perwt       = perwt       / 100
replace slwt        = slwt        / 100

format serial      %8.0g
format cbserial    %13.0g
format hhwt        %10.2f
format cluster     %13.0g
format adjust      %7.6f
format cpi99       %5.3f
format density     %7.1f
format metpop10    %8.0g
format strata      %12.0g
format perwt       %10.2f
format slwt        %10.2f

label var year        `"Census year"'
label var sample      `"IPUMS sample identifier"'
label var serial      `"Household serial number"'
label var cbserial    `"Original Census Bureau household serial number"'
label var numprec     `"Number of person records following"'
label var subsamp     `"Subsample number"'
label var hhwt        `"Household weight"'
label var hhtype      `"Household Type"'
label var cluster     `"Household cluster for variance estimation"'
label var adjust      `"Adjustment factor, ACS/PRCS"'
label var cpi99       `"CPI-U adjustment factor to 1999 dollars"'
label var region      `"Census region and division"'
label var stateicp    `"State (ICPSR code)"'
label var statefip    `"State (FIPS code)"'
label var countyicp   `"County (ICPSR code)"'
label var countyfip   `"County (FIPS code)"'
label var density     `"Population-weighted density of PUMA"'
label var metro       `"Metropolitan status"'
label var met2013     `"Metropolitan area (2013 OMB delineations)"'
label var met2013err  `"Coverage error in MET2013 variable"'
label var metpop10    `"Average 2010 population of 2013 metro/micro areas in PUMA"'
label var city        `"City"'
label var cityerr     `"Coverage error in CITY variable"'
label var citypop     `"City population"'
label var puma        `"Public Use Microdata Area"'
label var strata      `"Household strata for variance estimation"'
label var cpuma0010   `"Consistent PUMA, 2000-2010"'
label var homeland    `"American Indian, Alaska Native, or Native Hawaiian homeland area"'
label var cntry       `"Country"'
label var gq          `"Group quarters status"'
label var gqtype      `"Group quarters type [general version]"'
label var gqtyped     `"Group quarters type [detailed version]"'
label var farm        `"Farm status"'
label var ownershp    `"Ownership of dwelling (tenure) [general version]"'
label var ownershpd   `"Ownership of dwelling (tenure) [detailed version]"'
label var mortgage    `"Mortgage status"'
label var mortgag2    `"Second mortgage status"'
label var farmprod    `"Sales of farm products"'
label var acrehous    `"House acreage"'
label var mortamt1    `"First mortgage monthly payment"'
label var mortamt2    `"Second mortgage monthly payment"'
label var taxincl     `"Mortgage payment includes property taxes"'
```

ACS microdata in Stata



	year	sample	serial	cbserial	numprec	subsamp	hhwt	hhtype	cluster	adjust	cpi9
1	2018	2018 ACS	1	2.018010e+12	1 person record	26	75.00	N/A	2.018000e+12	1.013097	0.6
2	2018	2018 ACS	2	2.018010e+12	1 person record	76	75.00	N/A	2.018000e+12	1.013097	0.6
3	2018	2018 ACS	3	2.018010e+12	1 person record	2	118.00	N/A	2.018000e+12	1.013097	0.6
4	2018	2018 ACS	4	2.018010e+12	1 person record	92	43.00	N/A	2.018000e+12	1.013097	0.6
5	2018	2018 ACS	5	2.018010e+12	1 person record	81	16.00	N/A	2.018000e+12	1.013097	0.6
6	2018	2018 ACS	6	2.018010e+12	1 person record	5	25.00	N/A	2.018000e+12	1.013097	0.6
7	2018	2018 ACS	7	2.018010e+12	1 person record	6	18.00	N/A	2.018000e+12	1.013097	0.6
8	2018	2018 ACS	8	2.018010e+12	1 person record	9	85.00	N/A	2.018000e+12	1.013097	0.6
9	2018	2018 ACS	9	2.018010e+12	1 person record	94	16.00	N/A	2.018000e+12	1.013097	0.6
10	2018	2018 ACS	10	2.018010e+12	1 person record	40	91.00	N/A	2.018000e+12	1.013097	0.6
11	2018	2018 ACS	11	2.018010e+12	1 person record	87	92.00	N/A	2.018000e+12	1.013097	0.6
12	2018	2018 ACS	12	2.018010e+12	1 person record	37	31.00	N/A	2.018000e+12	1.013097	0.6
13	2018	2018 ACS	13	2.018010e+12	1 person record	12	16.00	N/A	2.018000e+12	1.013097	0.6
14	2018	2018 ACS	14	2.018010e+12	1 person record	98	71.00	N/A	2.018000e+12	1.013097	0.6
15	2018	2018 ACS	15	2.018010e+12	1 person record	20	68.00	N/A	2.018000e+12	1.013097	0.6
16	2018	2018 ACS	16	2.018010e+12	1 person record	18	54.00	N/A	2.018000e+12	1.013097	0.6
17	2018	2018 ACS	17	2.018010e+12	1 person record	82	40.00	N/A	2.018000e+12	1.013097	0.6
18	2018	2018 ACS	18	2.018010e+12	1 person record	85	11.00	N/A	2.018000e+12	1.013097	0.6
19	2018	2018 ACS	19	2.018010e+12	1 person record	73	88.00	N/A	2.018000e+12	1.013097	0.6
20	2018	2018 ACS	20	2.018010e+12	1 person record	32	20.00	N/A	2.018000e+12	1.013097	0.6
21	2018	2018 ACS	21	2.018010e+12	1 person record	83	34.00	N/A	2.018000e+12	1.013097	0.6
22	2018	2018 ACS	22	2.018010e+12	1 person record	51	34.00	N/A	2.018000e+12	1.013097	0.6
23	2018	2018 ACS	23	2.018010e+12	1 person record	24	30.00	N/A	2.018000e+12	1.013097	0.6
24	2018	2018 ACS	24	2.018010e+12	1 person record	23	17.00	N/A	2.018000e+12	1.013097	0.6
25	2018	2018 ACS	25	2.018010e+12	1 person record	7	3.00	N/A	2.018000e+12	1.013097	0.6
26	2018	2018 ACS	26	2.018010e+12	1 person record	14	15.00	N/A	2.018000e+12	1.013097	0.6
27	2018	2018 ACS	27	2.018010e+12	1 person record	3	66.00	N/A	2.018000e+12	1.013097	0.6
28	2018	2018 ACS	28	2.018010e+12	1 person record	10	30.00	N/A	2.018000e+12	1.013097	0.6
29	2018	2018 ACS	29	2.018010e+12	1 person record	53	56.00	N/A	2.018000e+12	1.013097	0.6
30	2018	2018 ACS	30	2.018010e+12	1 person record	72	53.00	N/A	2.018000e+12	1.013097	0.6
31	2018	2018 ACS	31	2.018010e+12	1 person record	36	15.00	N/A	2.018000e+12	1.013097	0.6
32	2018	2018 ACS	32	2.018010e+12	1 person record	99	52.00	N/A	2.018000e+12	1.013097	0.6
33	2018	2018 ACS	33	2.018010e+12	1 person record	15	53.00	N/A	2.018000e+12	1.013097	0.6
34	2018	2018 ACS	34	2.018010e+12	1 person record	22	18.00	N/A	2.018000e+12	1.013097	0.6
35	2018	2018 ACS	35	2.018010e+12	1 person record	17	17.00	N/A	2.018000e+12	1.013097	0.6
36	2018	2018 ACS	36	2.018010e+12	1 person record	35	13.00	N/A	2.018000e+12	1.013097	0.6
37	2018	2018 ACS	37	2.018010e+12	1 person record	95	70.00	N/A	2.018000e+12	1.013097	0.6
38	2018	2018 ACS	38	2.018010e+12	1 person record	33	77.00	N/A	2.018000e+12	1.013097	0.6
39	2018	2018 ACS	39	2.018010e+12	1 person record	38	74.00	N/A	2.018000e+12	1.013097	0.6
40	2018	2018 ACS	40	2.018010e+12	1 person record	25	28.00	N/A	2.018000e+12	1.013097	0.6
41	2018	2018 ACS	41	2.018010e+12	1 person record	42	38.00	N/A	2.018000e+12	1.013097	0.6


TEXAS A&M
UNIVERSITY

Stata

- Stata is a software package that provides tools for data manipulation, visualization, and estimation of various statistics
- Stata programming language is easier to understand than other statistical software packages (SPSS, SAS, R)
- Stata is popular across various social sciences, such as sociology, demography, and economics
- See more information on

https://www.stata.com/why-use-stata/

Popularity of statistical software

- Bob Muenchen has been tracking popularity of data science software using a variety of different approaches
  - E.g., he uses Google Scholar to count the number of scholarly articles found each year for each software

https://r4stats.com/articles/popularity/

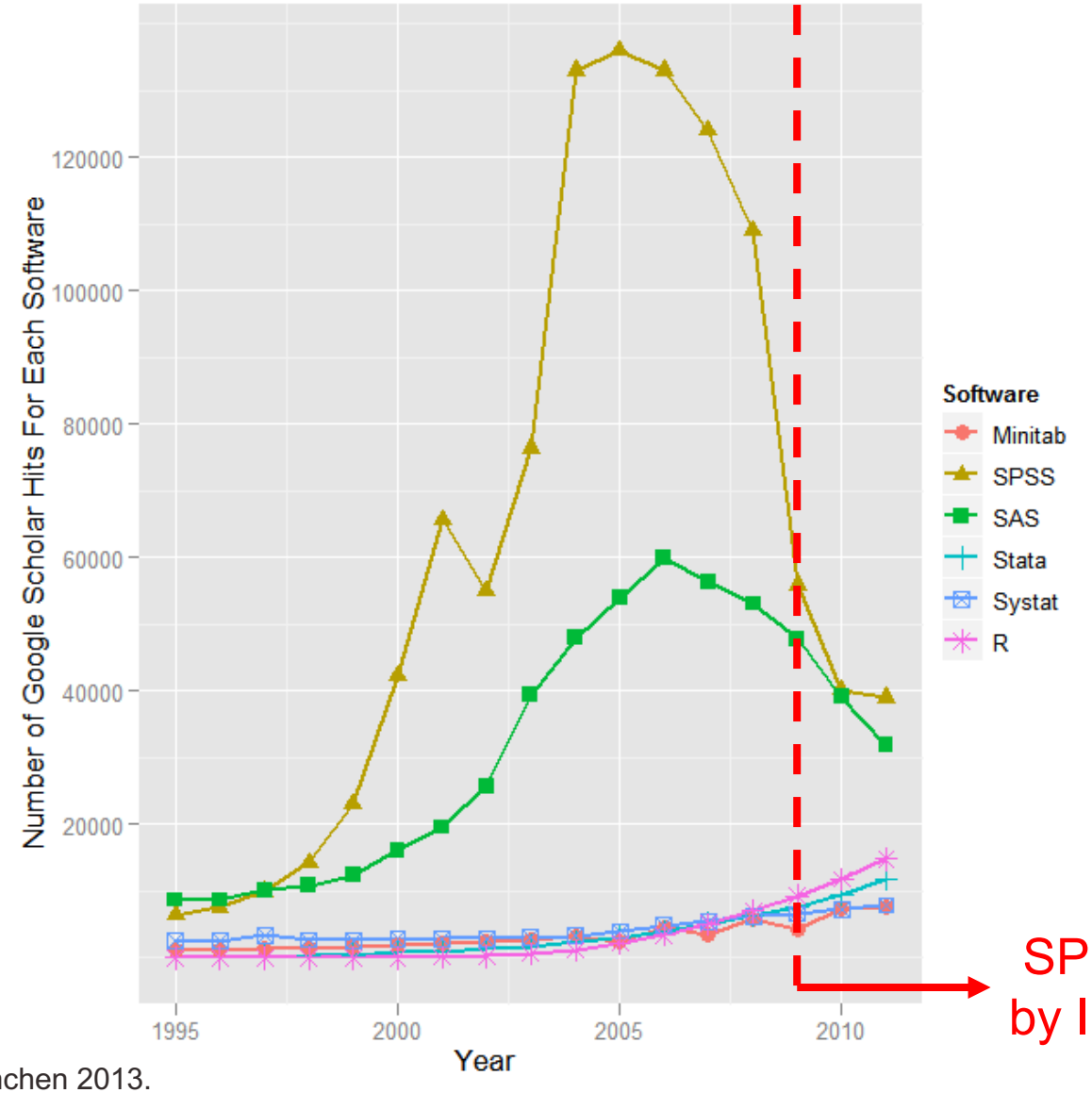
- Forecast Update: Will 2014 be the Beginning of the End for SAS and SPSS?
  - May 14, 2013, by Bob Muenchen

https://www.r-bloggers.com/forecast-update-will-2014-be-the-beginning-of-the-end-for-sas-and-spss/

- Is Scholarly Use of R Use Beating SPSS Already?
  - July 15, 2019, by Bob Muenchen

https://www.r-bloggers.com/is-scholarly-use-of-r-use-beating-spss-already/

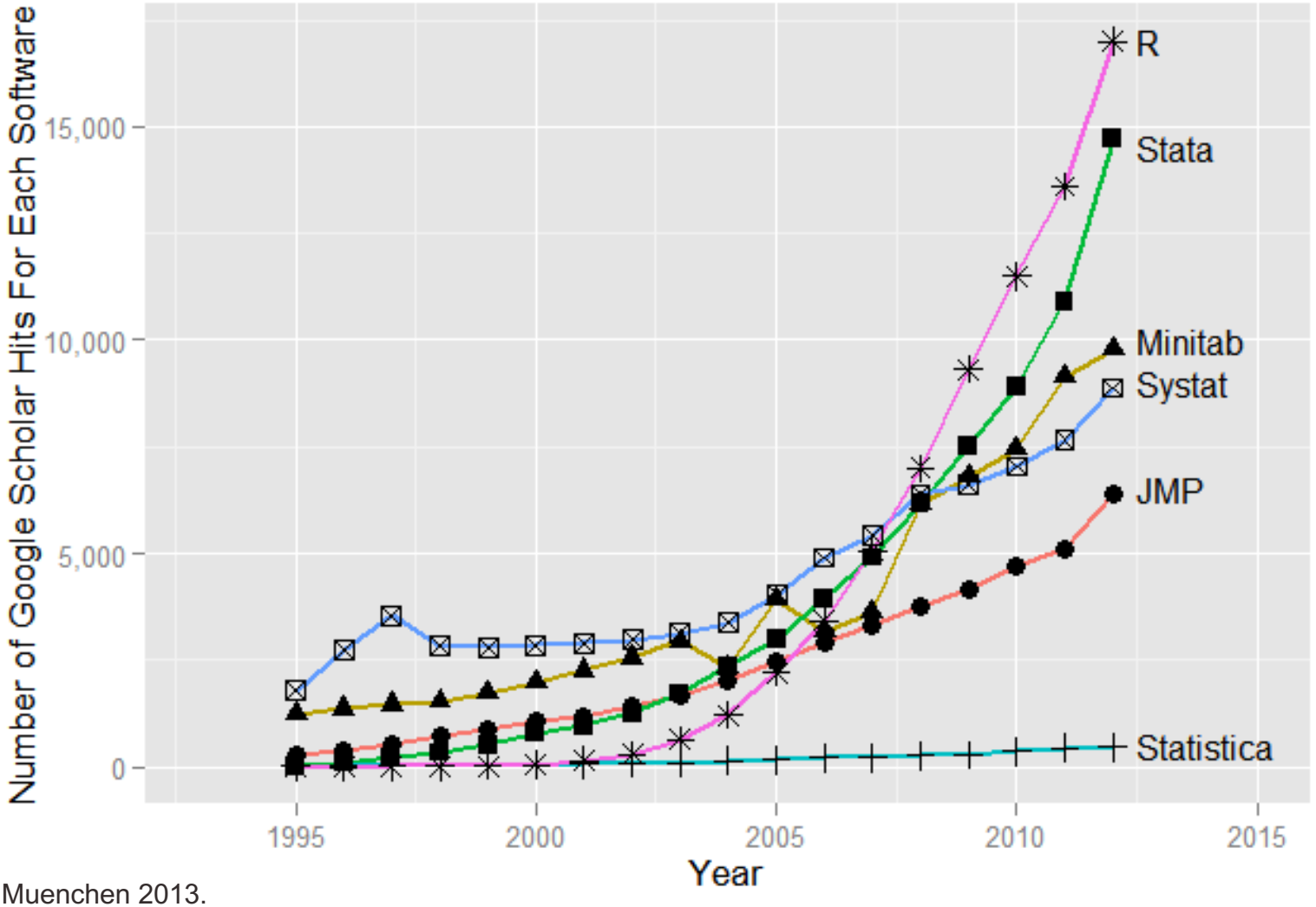
Scholarly use of data analysis software



Source: Muenchen 2013.

Site: https://www.r-bloggers.com/forecast-update-will-2014-be-the-beginning-of-the-end-for-sas-and-spss/

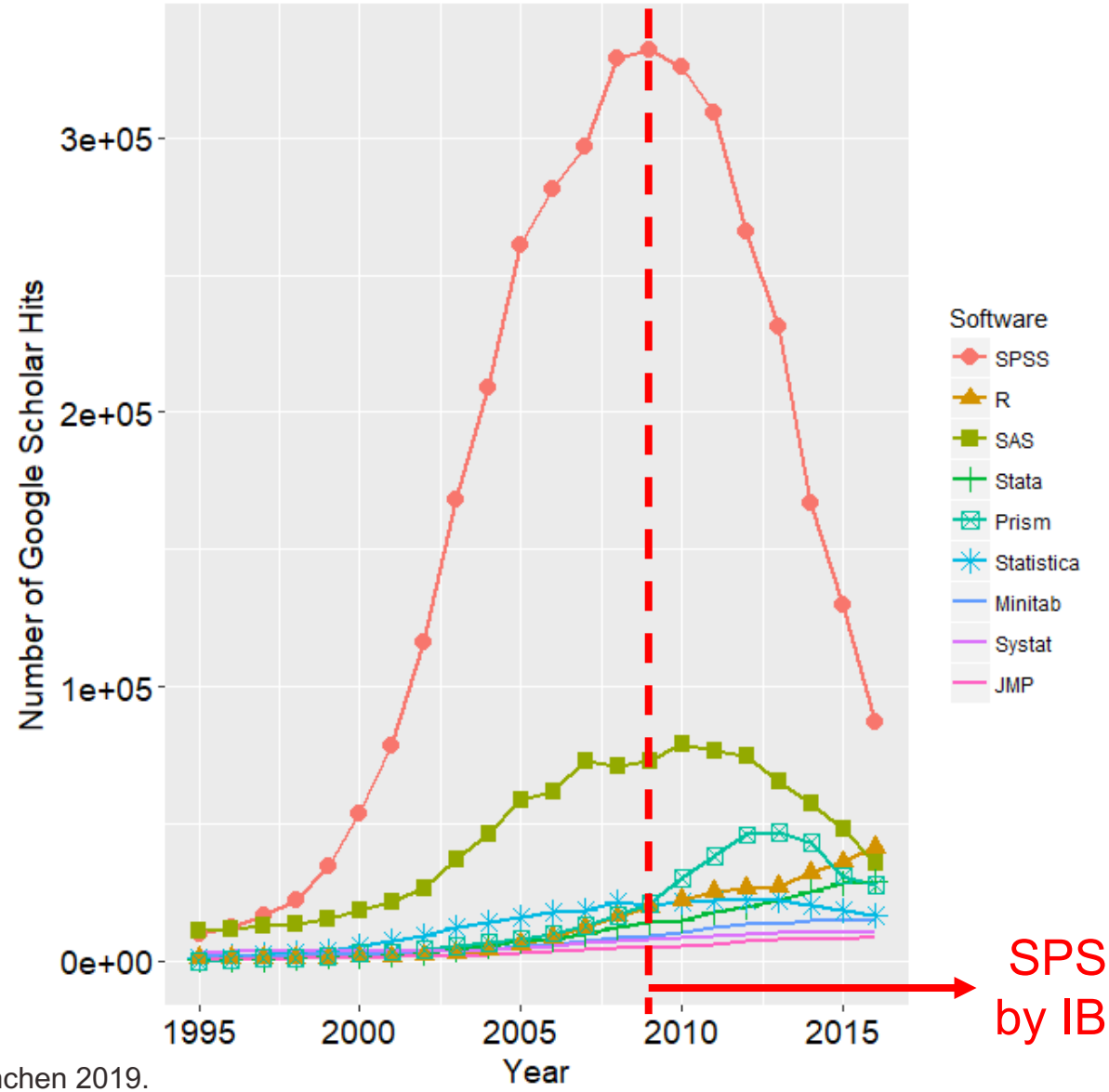
Scholarly use of data analysis software, SAS and SPSS removed



Source: Muenchen 2013.

Site: https://www.r-bloggers.com/forecast-update-will-2014-be-the-beginning-of-the-end-for-sas-and-spss/

Citations per year for each software



Source: Muenchen 2019.

Site: https://www.r-bloggers.com/is-scholarly-use-of-r-use-beating-spss-already/

Age-period-cohort effects

- Why most young demographers use R?

- Age effect
  - "You know, young people love free stuff and visualizations, they will grow up soon and will pay for Stata or SAS"

- Period effect
  - "I think it is because it is trendy nowadays, before everybody used Stata, later everybody will use Python"

- Cohort effect
  - "Maybe is because they learned R at the beginning of their carrier, and they will continue to use it for a long time"

Source: Acosta, Enrique. 2020. "Age-period-cohort analysis: Limitations and possibilities." Presentation at the 11th Demographic Conference of Young Demographers. February, 6.

R vs. Stata

- R is a free software package
  - The most advanced statistical models and techniques are made available quickly in R
  - Researchers, professors, and other professionals create extra commands for R with new methodological advances
  - The same happens for Stata, but not in the same pace
- Among our faculty, Stata is more popular

Stata licenses

- Instructions for accessing Stata through the Texas A&M Virtual Open Access Lab (VOAL)

  http://www.ernestoamaral.com/docs/soci420-23fall/Stata_VOAL_instructions.pdf

- Student short-term Stata license (free for a maximum of one week)

  https://www.stata.com/customer-service/short-term-license

- Student Single-User Stata License (lower prices)

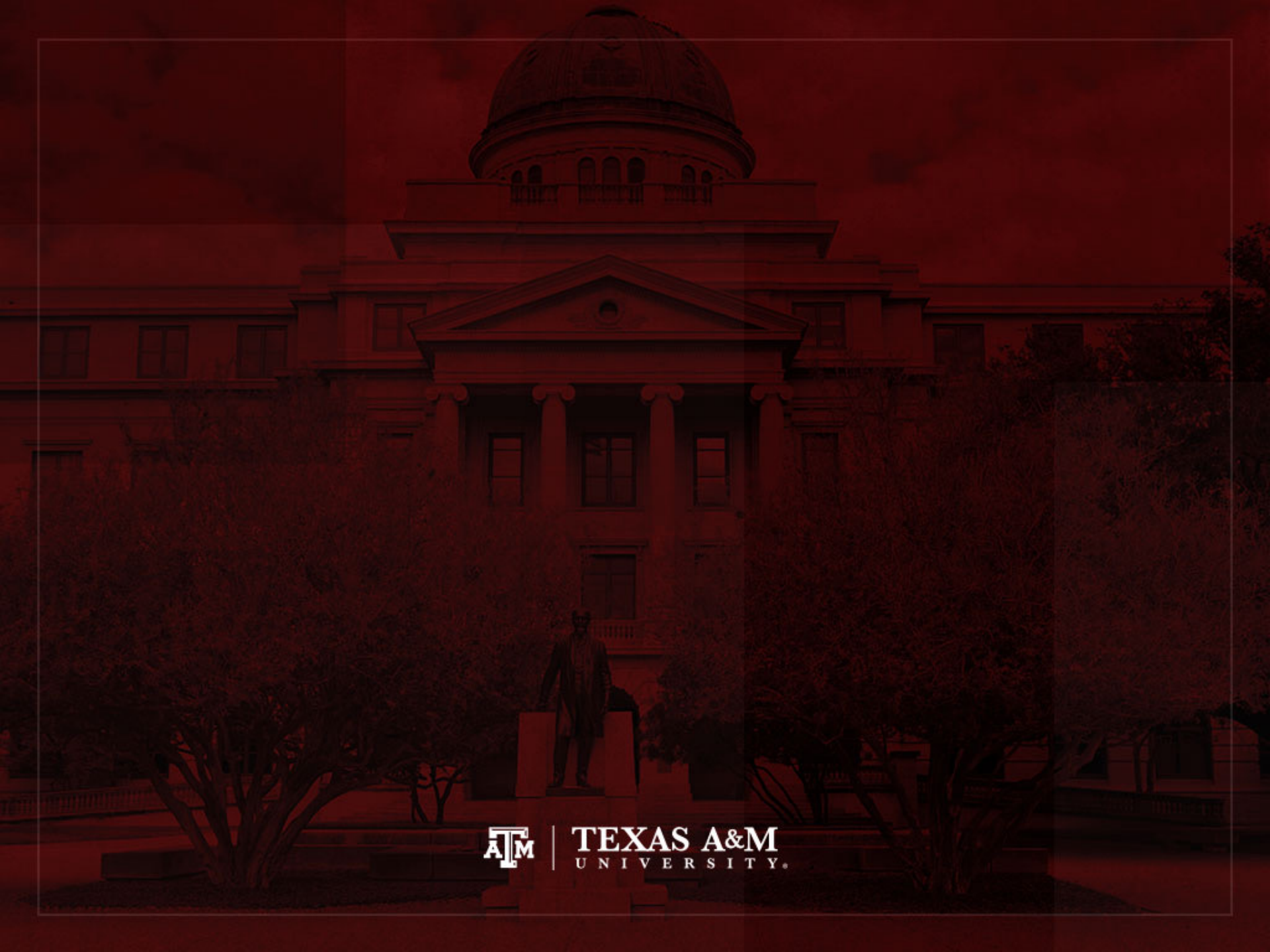
  https://www.stata.com/order/new/edu/gradplans/student-pricing

Stata help resources

- Stata: Data Analysis and Statistical Software
  http://www.stata.com/links

- Institute for Digital Research and Education (IDRE)
  - University of California, Los Angeles (UCLA)
    https://stats.idre.ucla.edu/stata/

- Carolina Population Center (CPC)
  - The University of North Carolina at Chapel Hill (UNC)
    http://www.cpc.unc.edu/research/tools/data_analysis/statatutorial


TEXAS A&M
UNIVERSITY