Lecture 1: Introduction

Ernesto F. L. Amaral

August 22–24, 2023
Advanced Methods of Social Research (SOCI 420)

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
- General Social Survey (GSS)
- 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 General Social Survey (GSS) 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

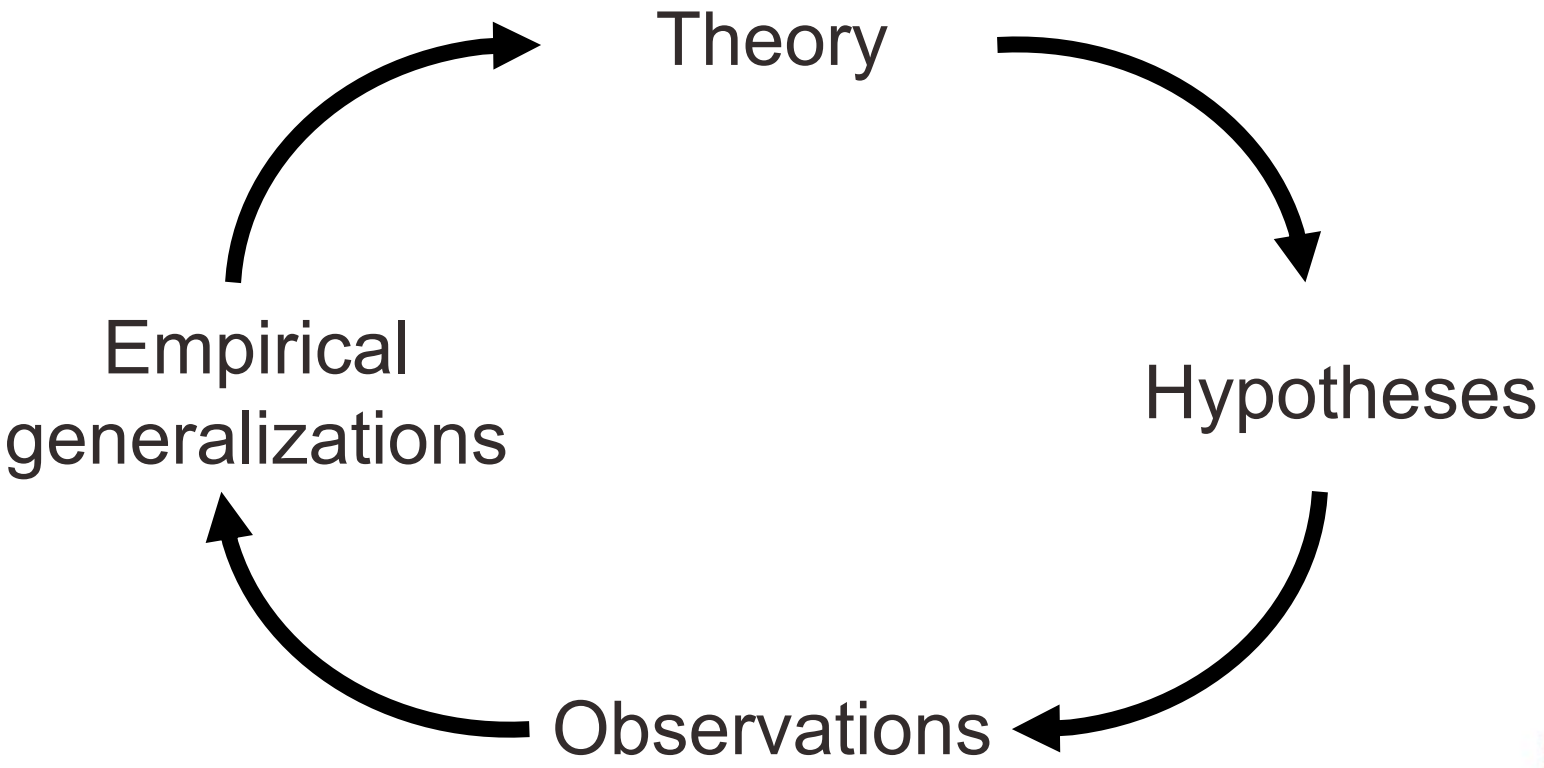
- **Studies without statistics**
  - Some of the most important works in the social sciences do not utilize statistics
  - There is nothing magical about data and statistics
  - Presence of numbers guarantees nothing about the quality of a scientific inquiry
- **Studies with statistics**
  - Data can 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

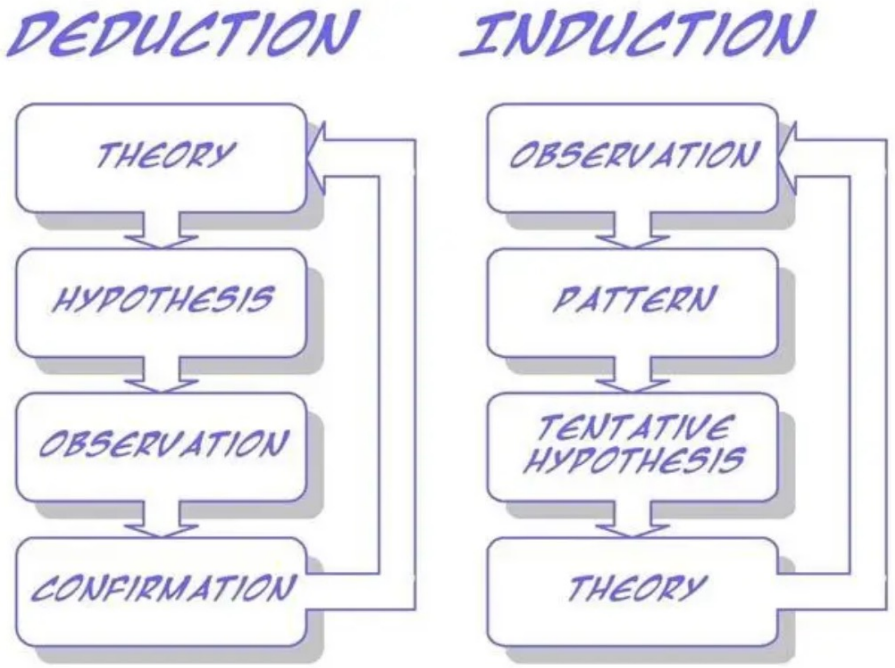
- **Research** is a disciplined inquiry to answer questions, examine ideas, and test theories
- **Statistics** are mathematical tools used to organize, summarize, and manipulate data
- **Quantitative research** collects and uses information in the form of numbers
- **Data** refers to information that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

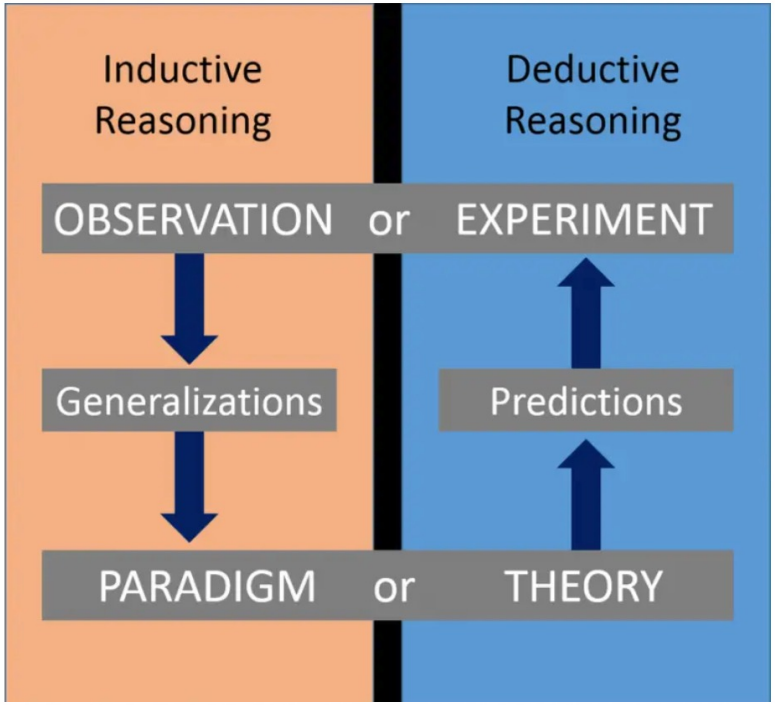
The wheel of science

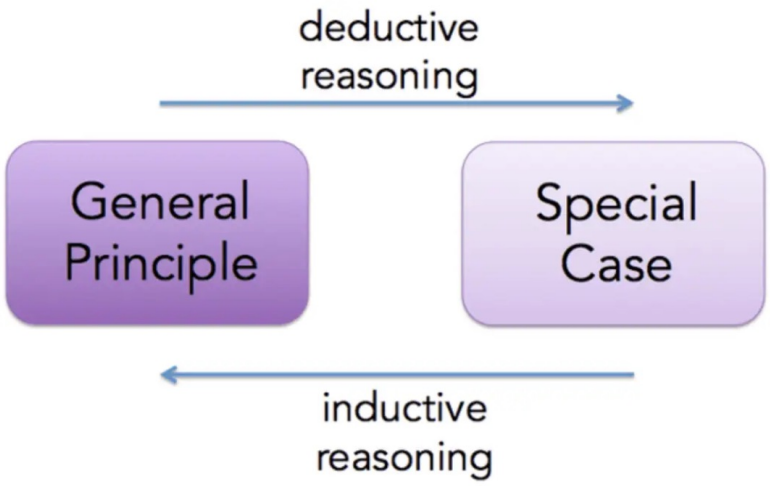
- Scientific theory and research continually shape each other

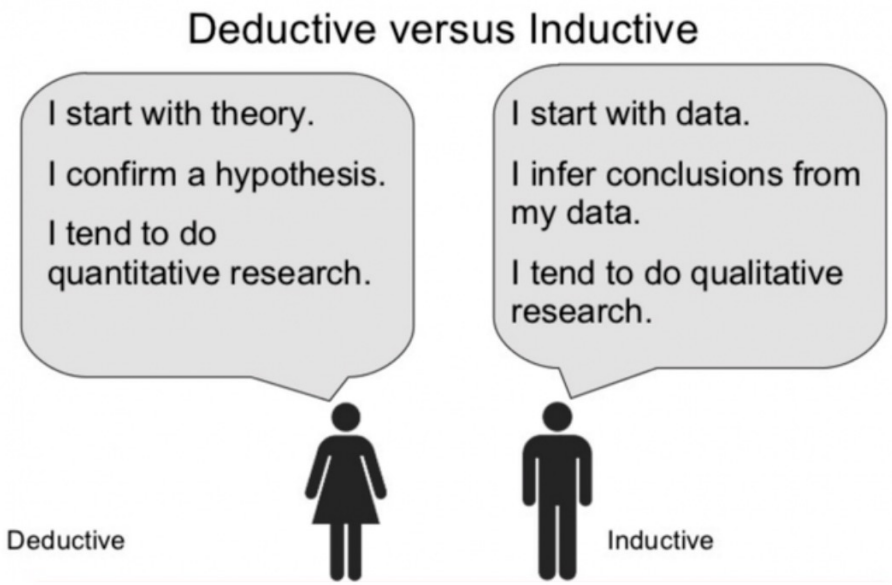


Source: Healey, 2015, p.2.









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.

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

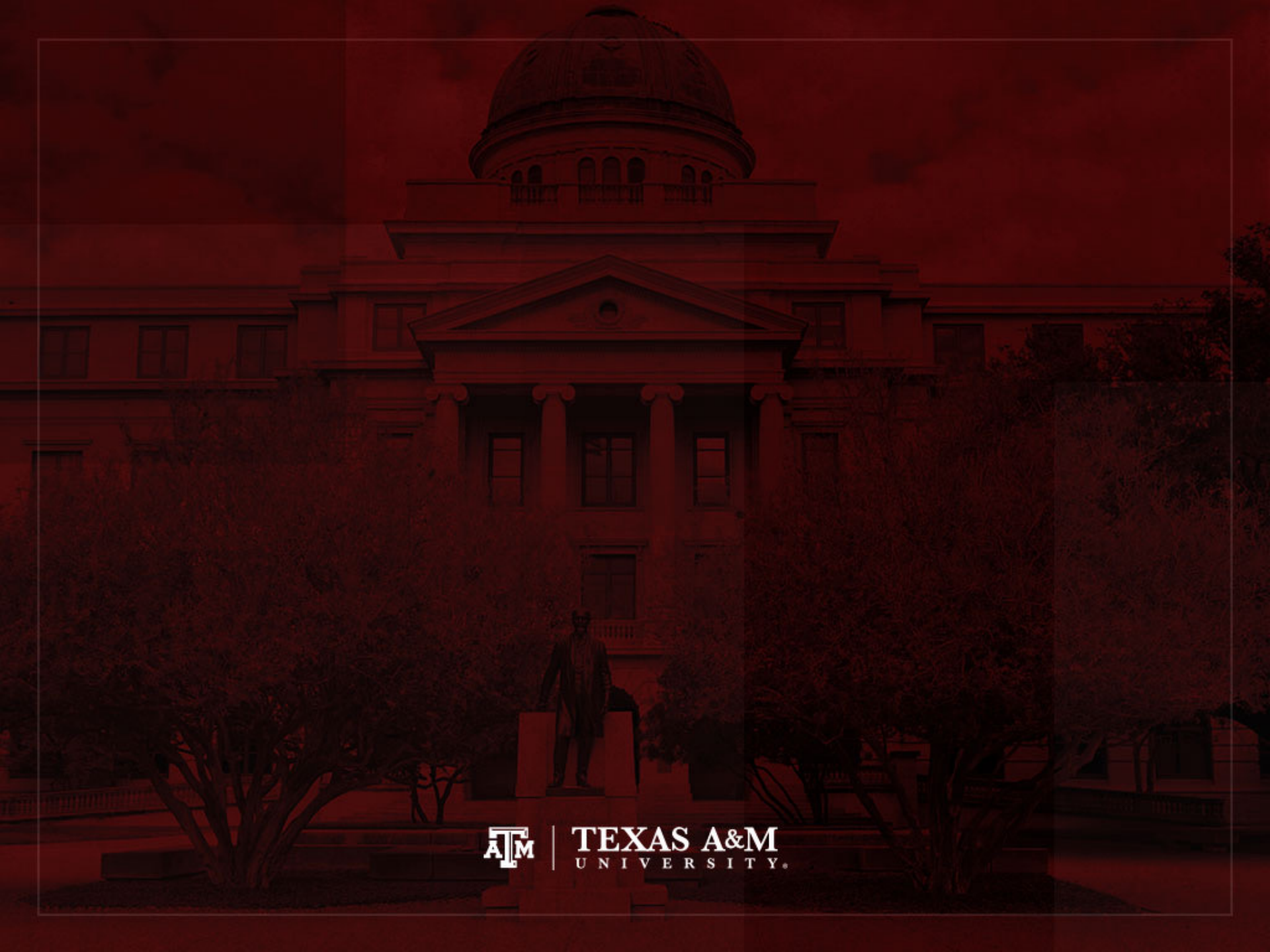
Statistical analysis

- 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

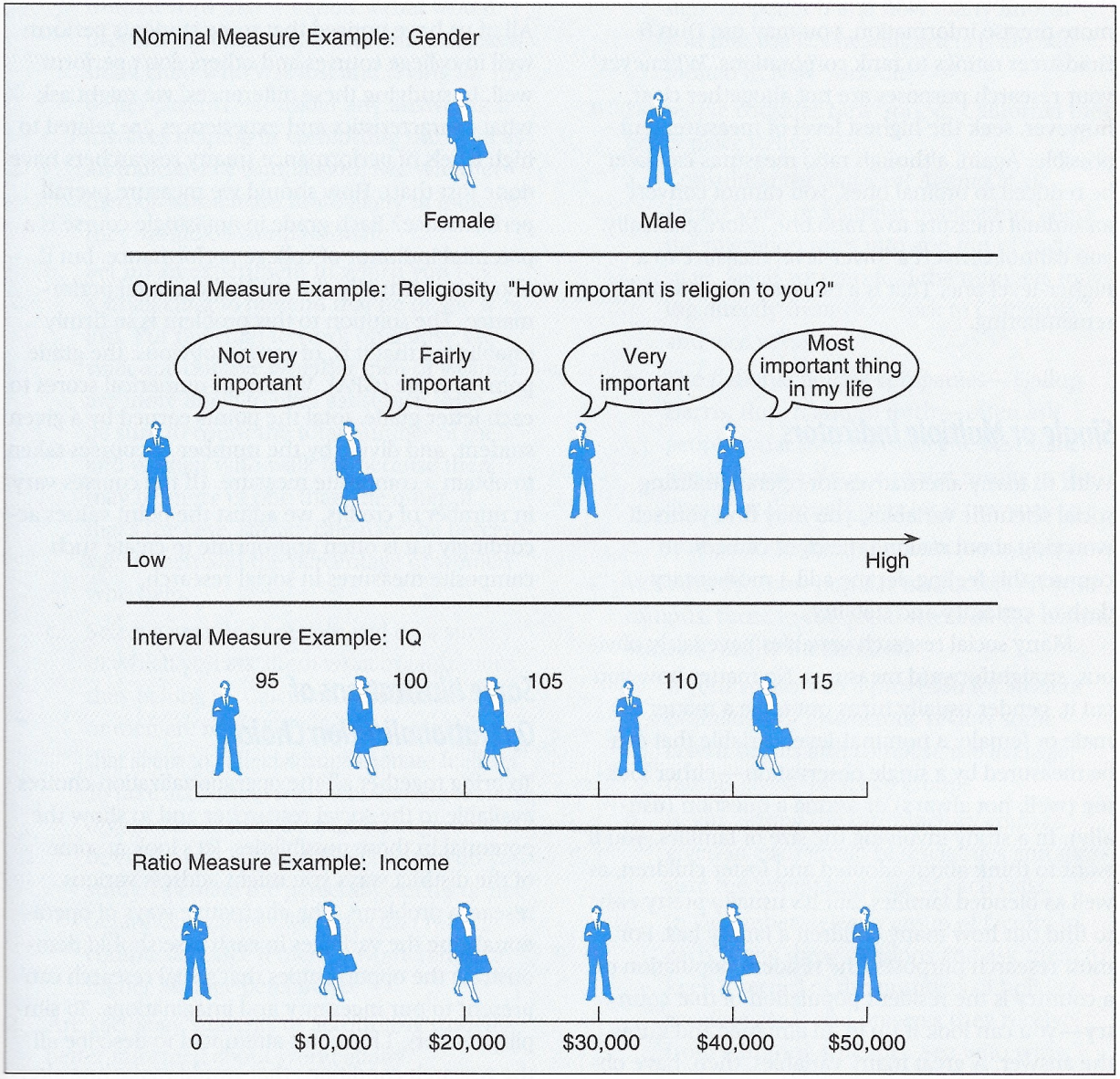
- Attitude and opinion scales
  - Prejudice, alienation, political conservatism...
- Likert scale:
  - (1) strongly disagree; (2) disagree; (3) neither agree nor disagree; (4) agree; (5) strongly agree
- Into which of the following classes would you say you belong?

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

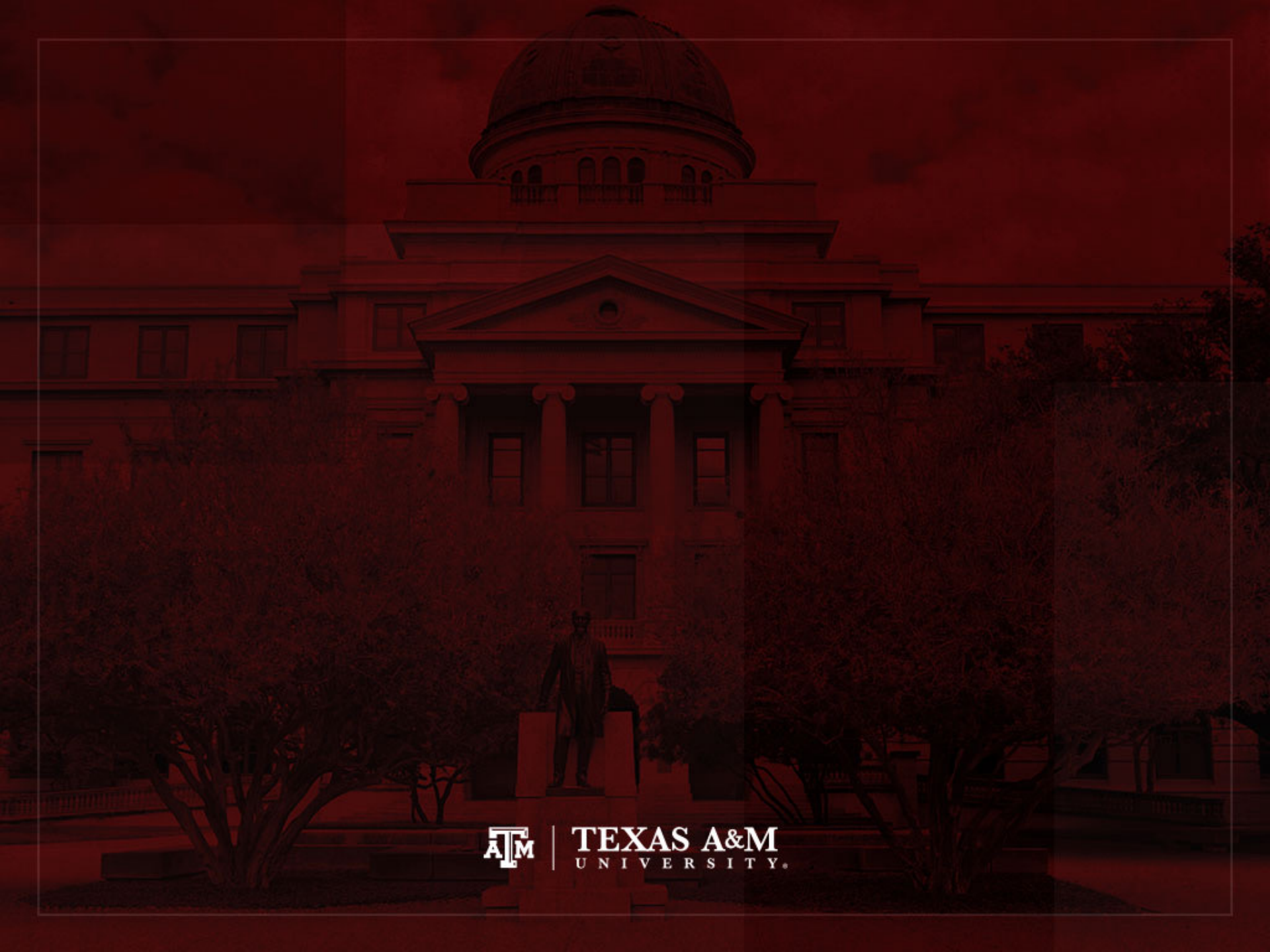
Income at the ordinal level

- 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

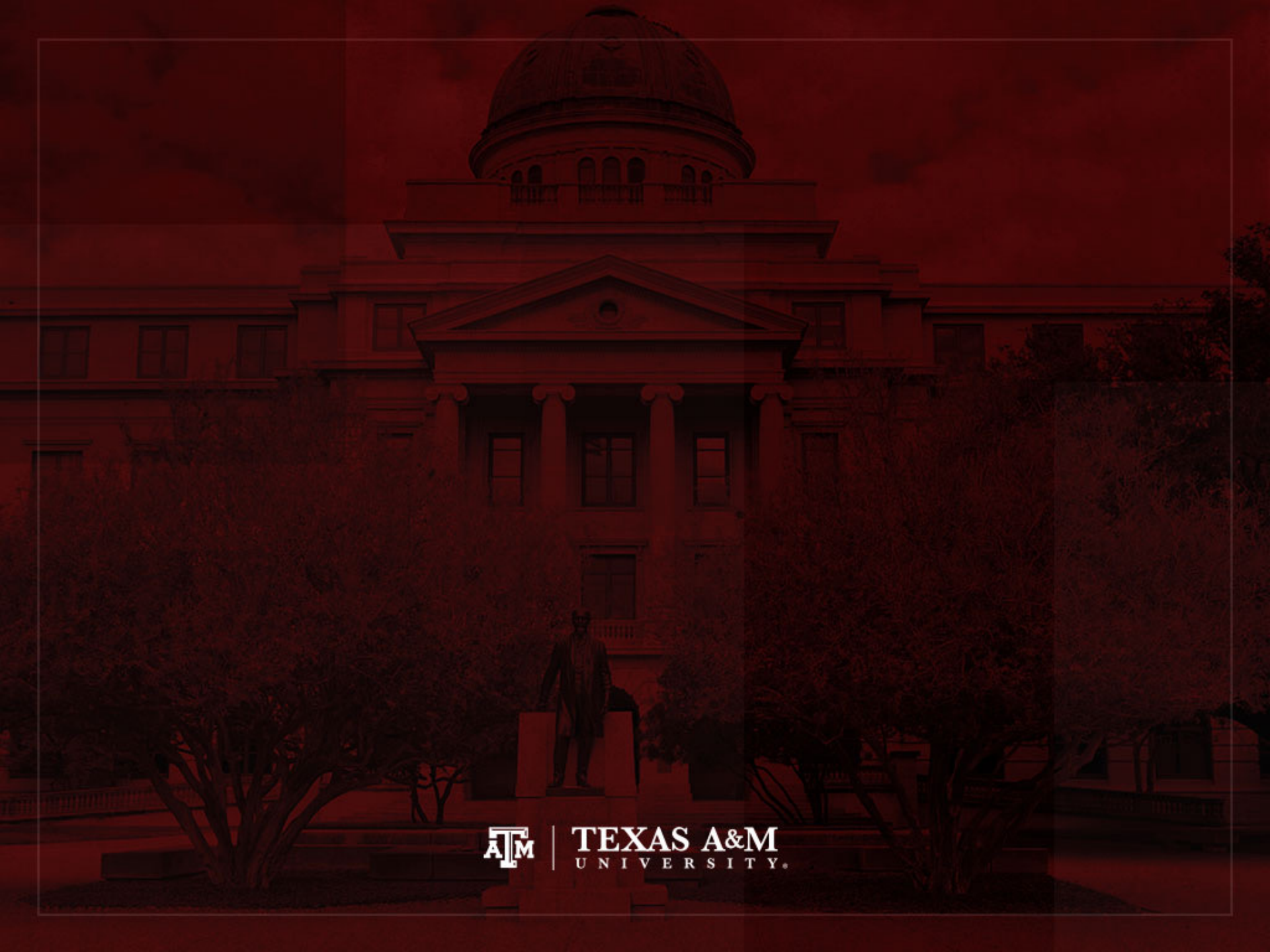
- **Multivariate descriptive** statistics
  - Describe the relationships between three or more variables
  - **Measures of association:** quantify the strength and direction of a multivariate relationship
- E.g. **grade, age, gender**
  - Strength: relationship between age and grade is strong for women, 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

General Social Survey (GSS)

https://gss.norc.org/About-The-GSS

- Nationally representative survey of adults in the United States conducted since 1972
- Data on contemporary American society in order to monitor and explain trends in opinions, attitudes and behaviors
- The GSS has adapted questions from earlier surveys, thereby allowing researchers to conduct comparisons for up to 80 years
- GSS questionnaires:

  https://gss.norc.org/get-documentation/questionnaires

GSS microdata

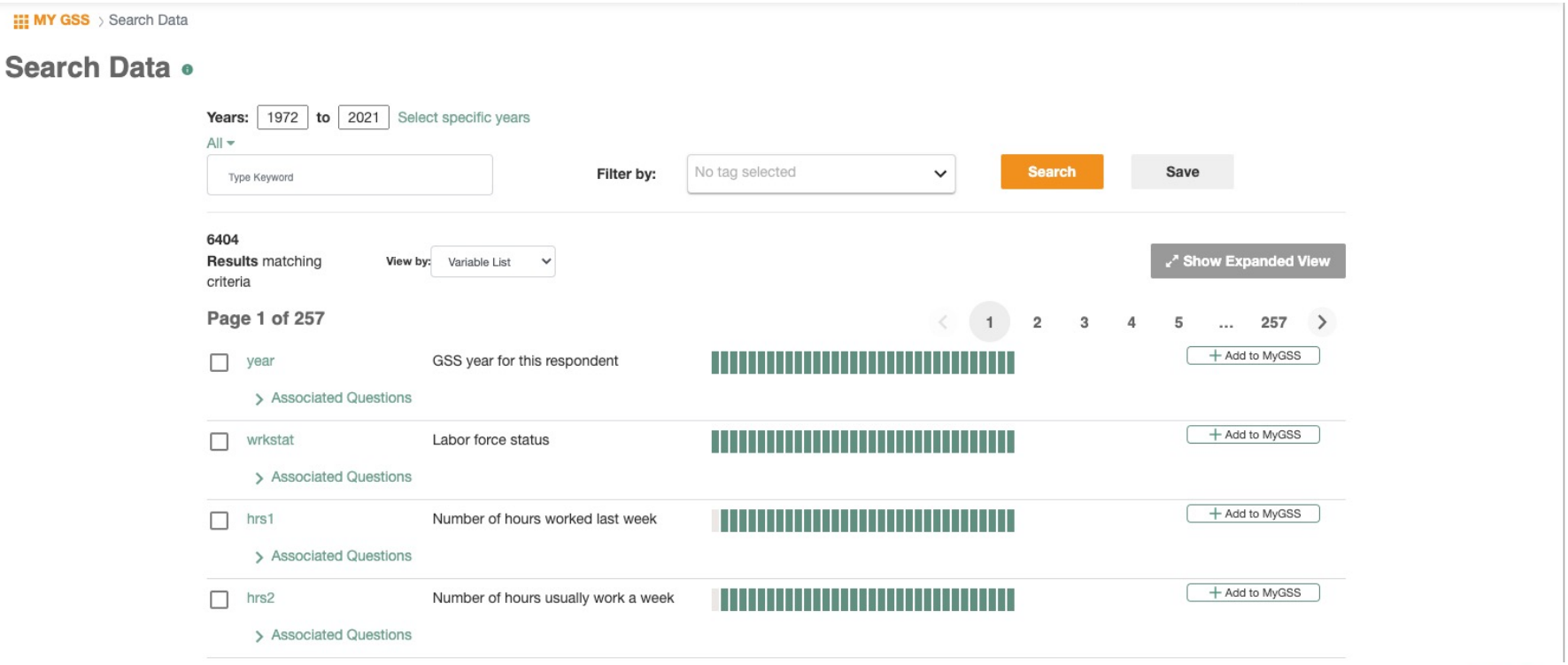
https://gss.norc.org/Get-The-Data

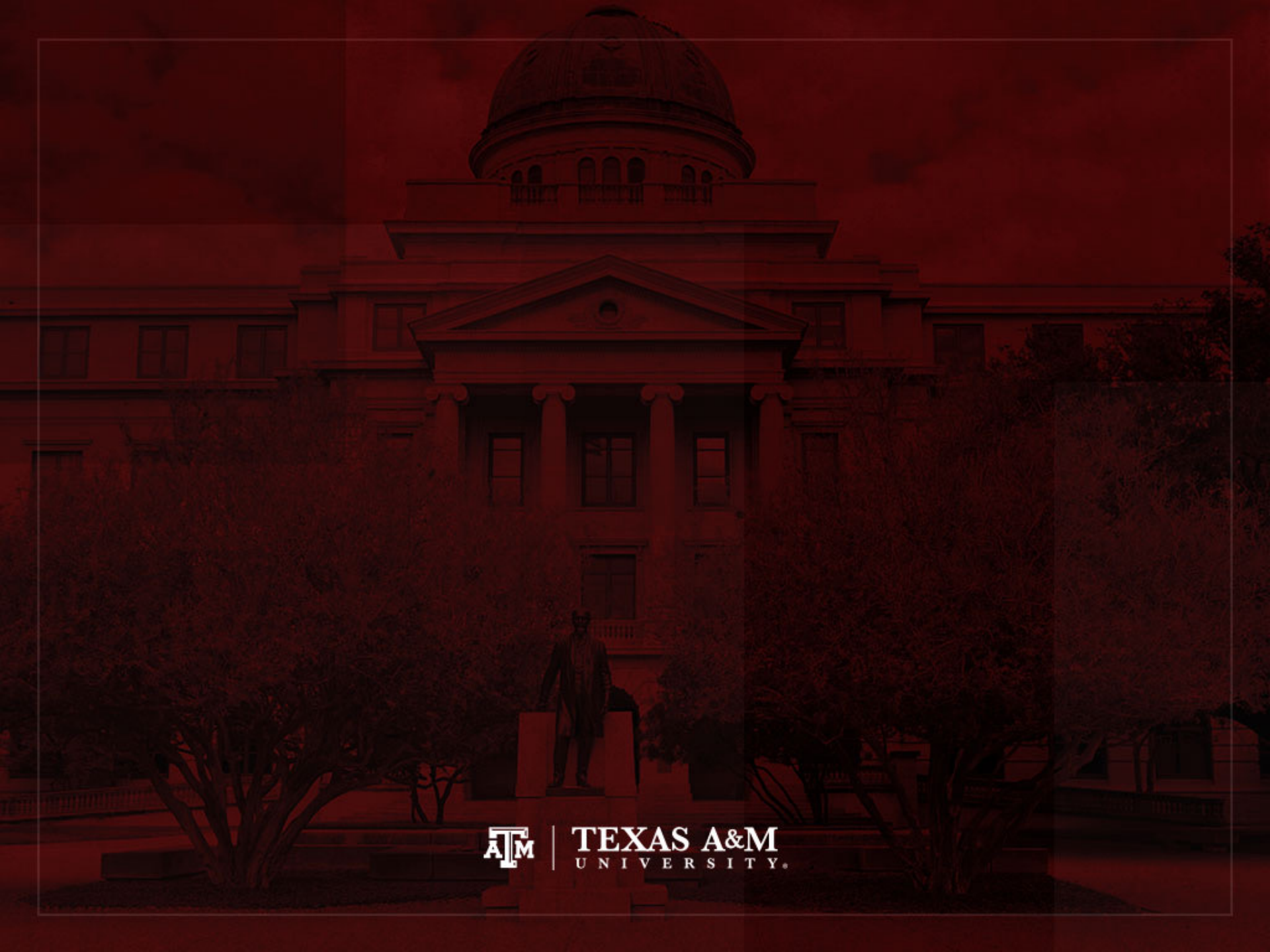
	year	id	wrkstat	wrkslf	wrkgovt	occ10	prestg10	indus10	marital	martype	divorce	widowed	pawrkslf	paocc10	papres10	paind10	mawrkslf	maocc10	mapres10	maind10
1	2021	1	1	2	.i	5400	38	7980	1	.i	2	2	1	9520	39	770	2	3255	64	8190
2	2021	2	1	2	.i	40	57	7470	3	.i	.i	2	2	10	72	2070	.i	.i	.i	.i
3	2021	3	2	2	.i	7750	35	4770	5	.i	.i	.i	2	2630	46	7380	2	4650	18	9090
4	2021	4	2	1	.i	4600	35	8470	2	.i	2	.i	2	3740	59	9470	2	5120	45	8390
5	2021	6	1	2	.i	5840	38	6990	5	.i	.i	.i	1	9130	35	6170	2	3500	69	8270
6	2021	7	1	2	.i	3800	40	9470	5	.i	.i	.i	2	4760	31	4670	.i	.i	.i	.i
7	2021	8	1	2	.i	1020	60	7390	5	.i	.i	.i	2	7720	27	3390	2	4230	25	7690
8	2021	9	2	2	.i	230	59	7870	3	.i	.i	2	2	1310	44	9590	2	230	59	8470
9	2021	10	5	2	.i	7020	38	6680	1	1	2	2	.i	.i	.i	.i	2	5860	32	7590
10	2021	12	8	2	.i	800	60	3960	3	.i	.i	2	1	310	39	8680	1	8350	42	1680
11	2021	13	1	2	.i	4850	45	4090	1	.i	2	2	.i	.i	.i	.i	.i	.i	.i	.i
12	2021	14	6	2	.i	4130	16	8680	5	.i	.i	.i	2	9120	35	6180	.i	.i	.i	.i
13	2021	15	5	2	.i	2310	61	7860	1	.i	2	2	.i	.i	.i	.i	1	8310	31	9070
14	2021	16	6	.i	.i	.i	.i	.i	5	.i	.i	.i	2	1240	65	7470	.i	.i	.i	.i
15	2021	17	2	1	.i	4850	45	4580	1	.i	2	2	1	6100	36	8590	2	4760	31	5170
16	2021	18	2	2	.i	2340	38	7890	5	.i	.i	.i	.i	.i	.i	.i	2	350	64	8090
17	2021	19	5	1	.i	4600	35	8470	5	.i	.i	.i	.i	.i	.i	.i	.i	.i	.i	.i
18	2021	21	8	2	.i	110	60	3980	1	.i	2	2	.i	.i	.i	.i	2	4760	31	5290
19	2021	22	6	2	.i	2900	43	7870	5	.i	.i	.i	2	1410	73	7460	2	5320	25	6770
20	2021	23	1	2	.i	2810	54	6570	5	.i	.i	.i	2	120	53	6870	2	735	57	7890
21	2021	24	1	2	.i	4760	31	5580	5	.i	.i	.i	2	4700	38	4670	2	5700	47	4770
22	2021	25	1	2	.i	4930	51	7380	5	.i	.i	.i	2	2200	74	7870	.i	.i	.i	.i
23	2021	26	7	2	.i	5100	24	4260	1	.i	2	2	2	430	39	6290	2	5230	29	5170
24	2021	27	4	2	.i	2810	54	6470	3	.i	.i	2	.i	.i	.i	.i	2	3500	69	8190
25	2021	28	1	2	.i	3850	60	9470	1	.i	2	2	.i	.i	.i	.i	2	5230	29	6870
26	2021	29	4	2	.i	1550	50	770	1	.i	2	2	.i	.i	.i	.i	2	5700	47	7870
27	2021	30	1	2	.i	2320	64	7860	1	.i	2	2	1	430	39	7370	.i	.i	.i	.i
28	2021	31	1	2	.i	2200	74	7870	1	.i	2	2	2	2200	74	7870	.i	.i	.i	.i
29	2021	32	4	2	.i	9620	25	5790	5	.i	.i	.i	2	3850	60	9470	2	1820	71	8370
30	2021	35	5	2	.i	5540	45	6370	4	.i	.i	2	.i	.i	.i	.i	.i	.i	.i	.i
31	2021	36	1	2	.i	3060	80	8180	1	.i	2	2	1	9140	26	6190	.i	.i	.i	.i
32	2021	37	1	2	.i	5800	47	9590	3	.i	.i	2	.i	.i	.i	.i	2	3500	69	8190
33	2021	38	1	2	.i	2550	50	7870	5	.i	.i	.i	2	4700	38	5190	2	5700	47	7980
34	2021	39	1	2	.i	20	50	2370	3	.i	.i	2	2	140	50	2370	2	5700	47	6990
35	2021	40	1	2	.i	5240	31	6890	3	.i	.i	2	.i	.i	.i	.i	2	565	47	2970
36	2021	41	1	2	.i	860	43	6990	3	.i	.i	2	2	6230	44	770	2	7750	35	2980
37	2021	42	1	2	.i	7010	49	3090	3	.i	.i	2	2	1360	65	9570	2	5700	47	8770
38	2021	43	1	2	.i	9130	35	4470	5	.i	.i	.i	2	8010	36	2880	1	4600	35	8470
39	2021	44	1	2	.i	4800	38	7470	1	.i	2	2	2	735	57	1290	2	630	47	6890
40	2021	45	1	2	.i	800	60	7290	5	.i	.i	.i	.i	.i	.i	.i	2	5120	45	3890
41	2021	46	7	2	.i	1006	65	6390	1	.i	1	2	2	1530	70	3360	2	1010	63	3390

GSS Data Explorer

- This is an online codebook that allows us to search for variables over time

https://gssdataexplorer.norc.org/variables/vfilter




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/

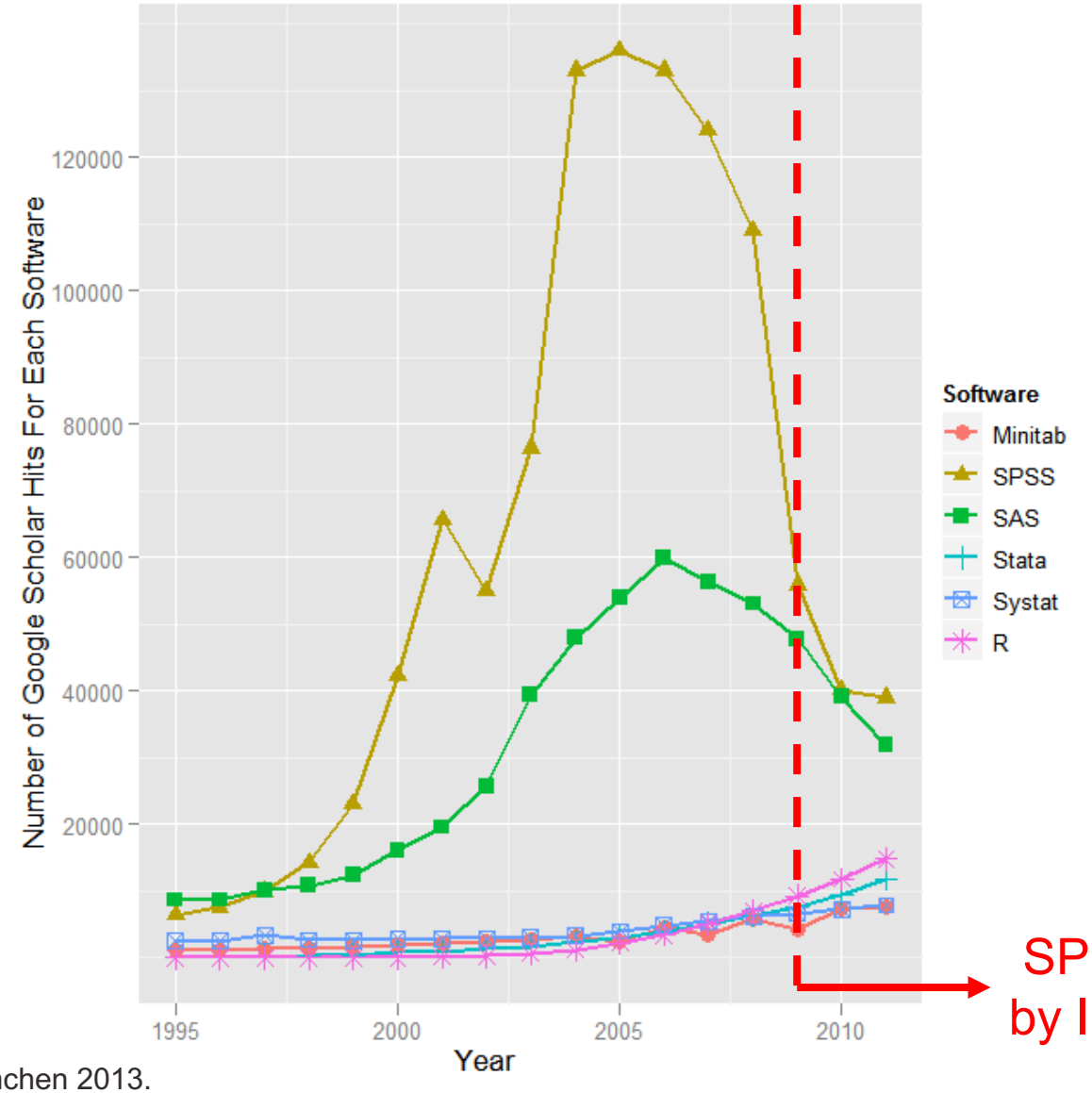
- Forecast Update: Will 2014 be the Beginning of the End for SAS and SPSS?
  - May 14, 2013, by Bob Muenchen

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/

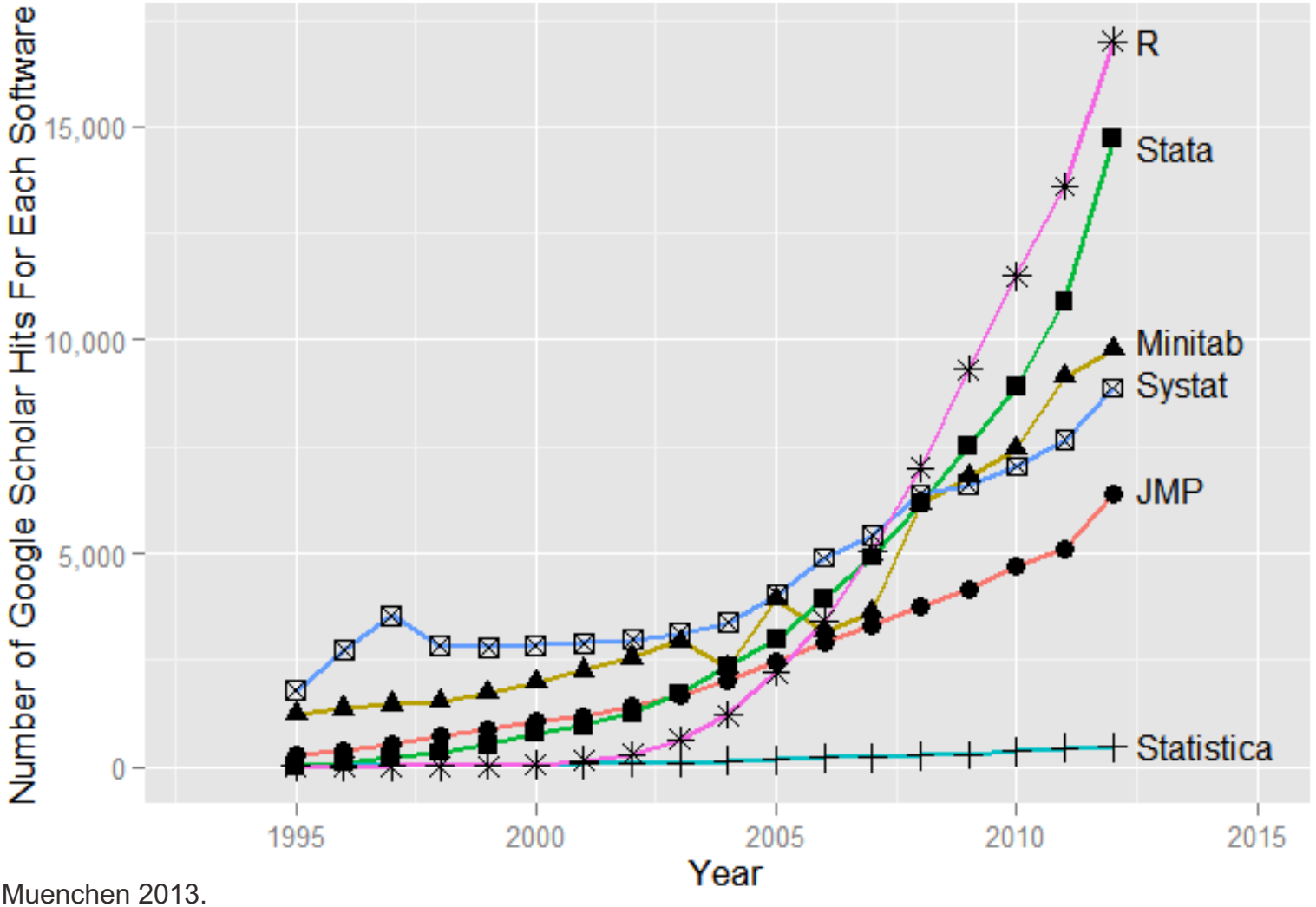
Scholarly use of data analysis software



Source: Muenchen 2013.

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

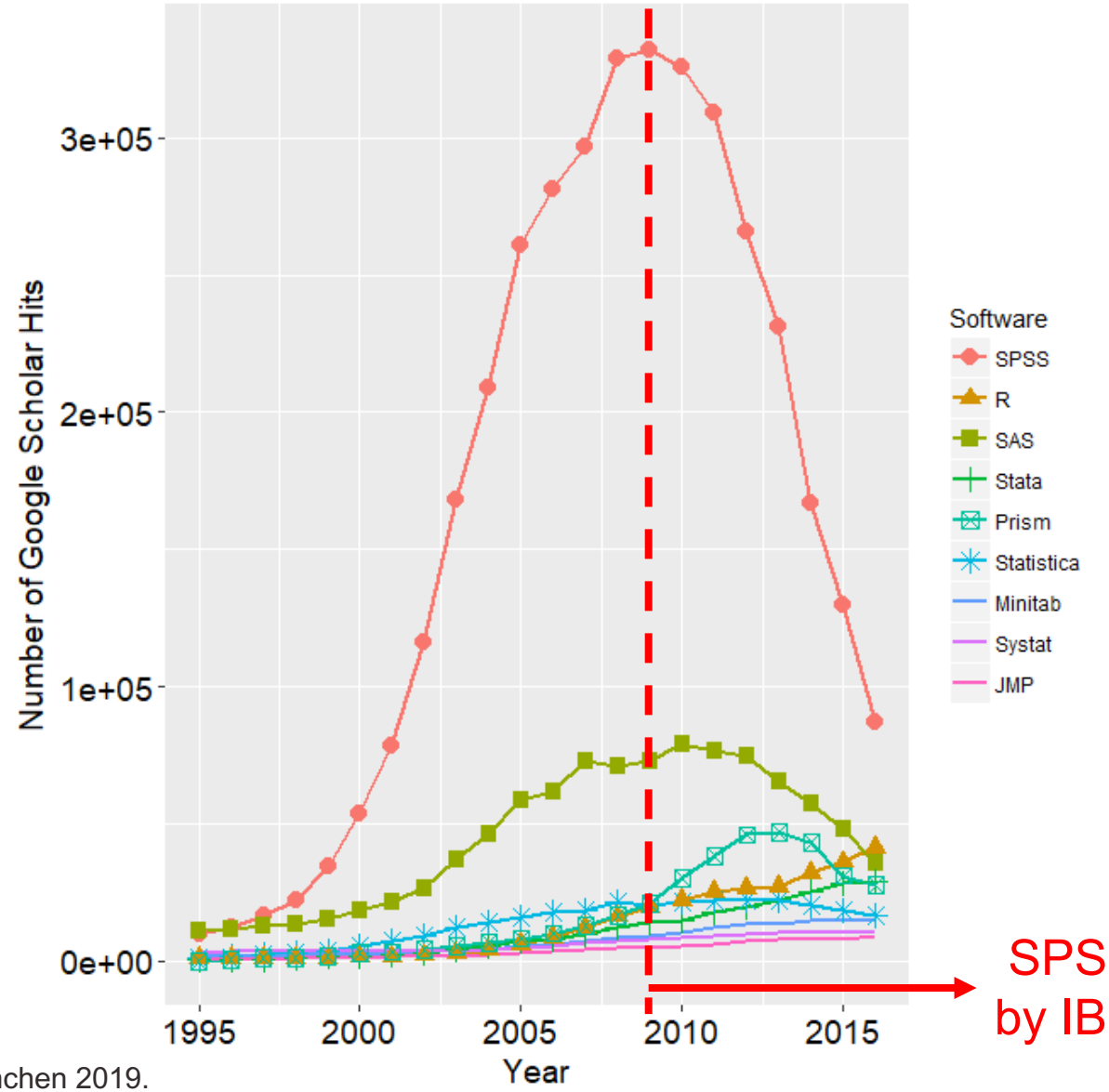
Scholarly use of data analysis software, SAS and SPSS removed



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)

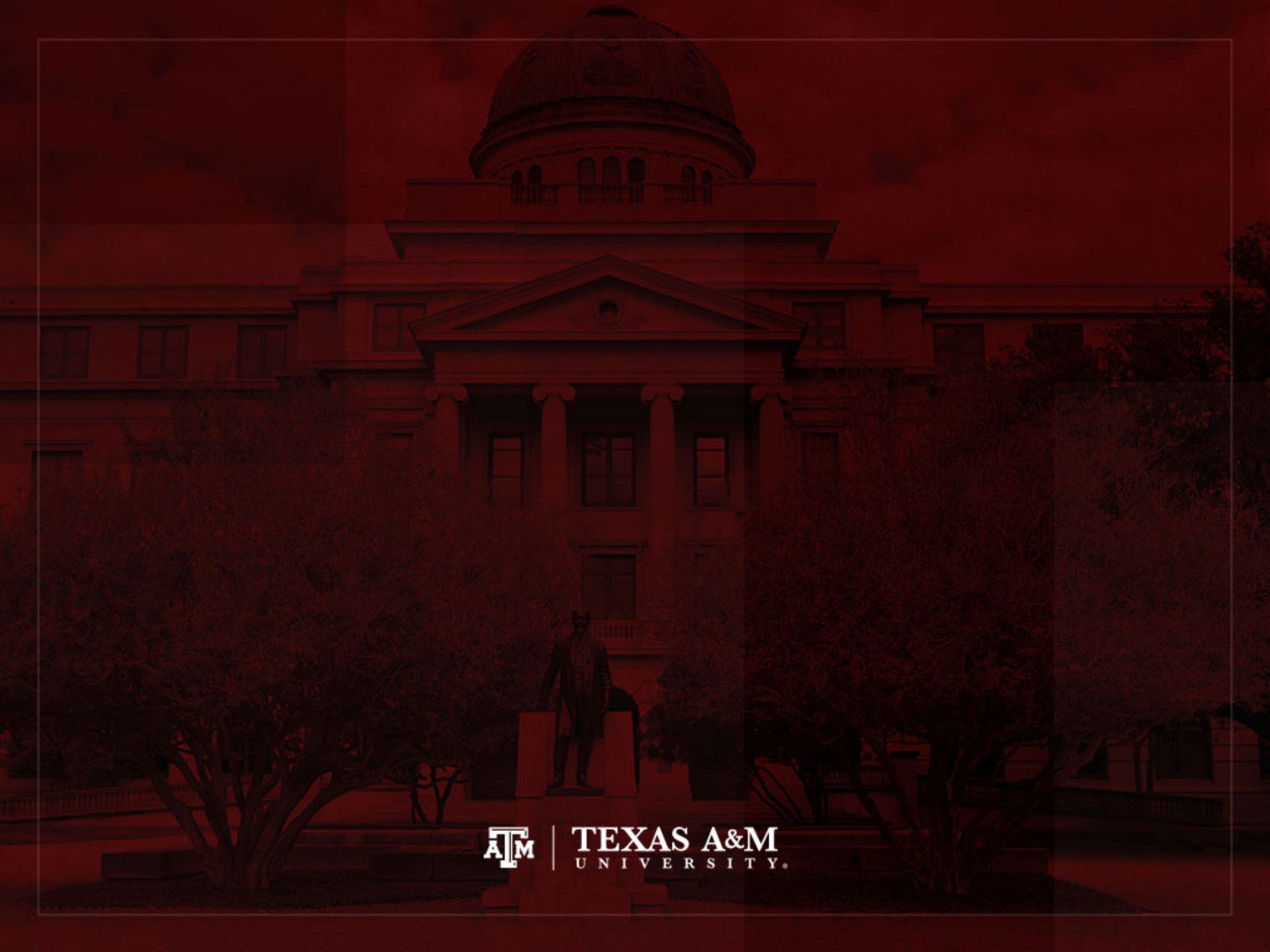
  https://www.stata.com/customer-service/short-term-license

- 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