

Lecture 1: Introduction

Ernesto F. L. Amaral

January 17–22, 2024

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).



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



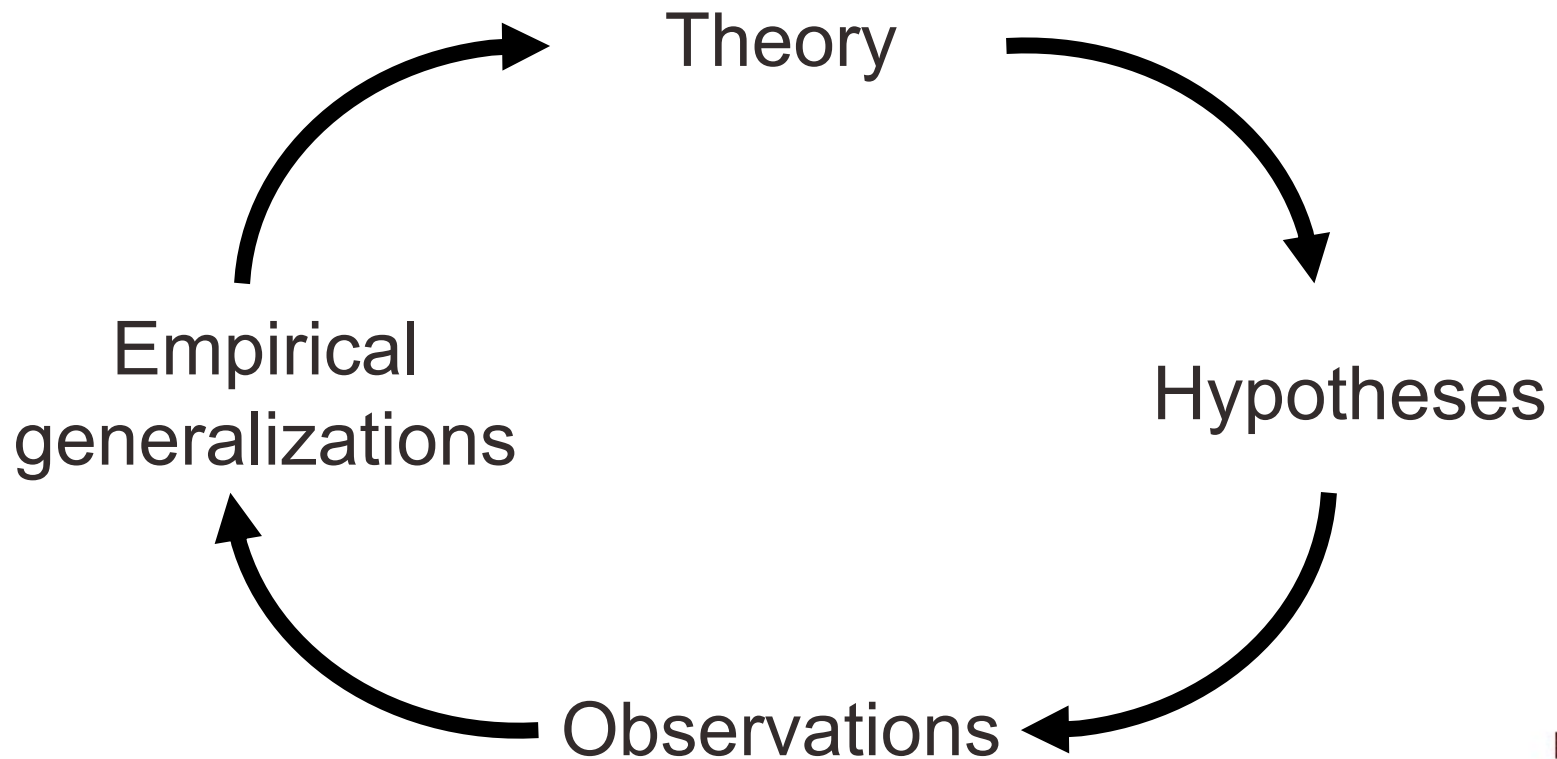
Statistics role in scientific inquiry

- **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

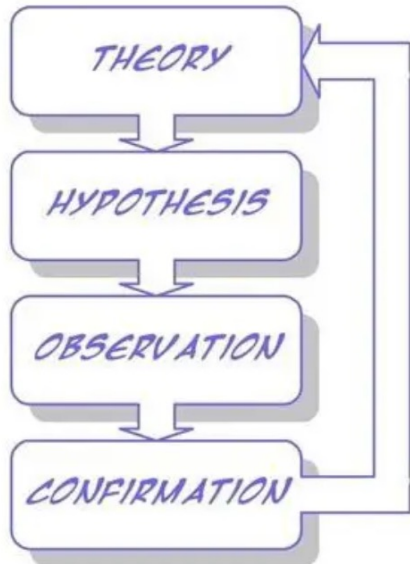
- Scientific theory and research continually shape each other



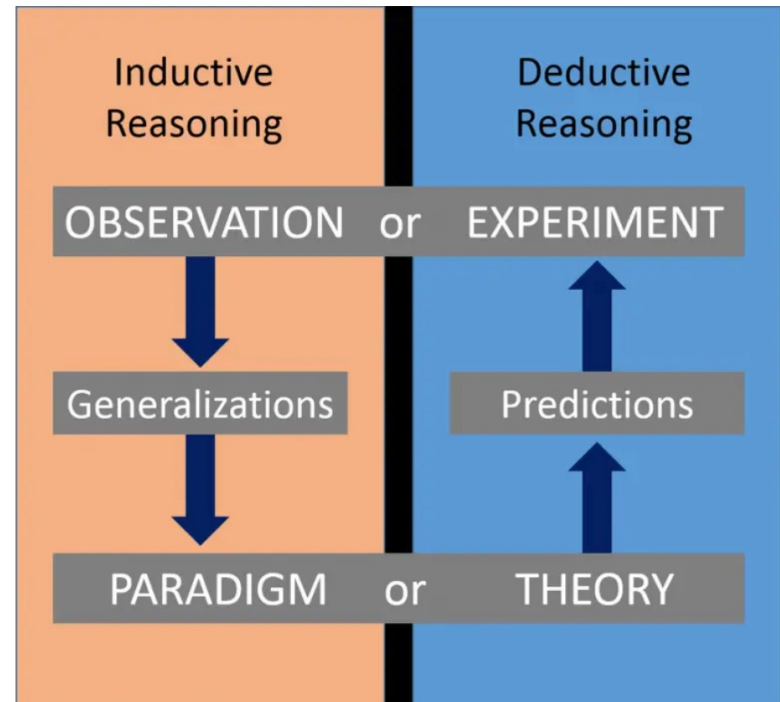
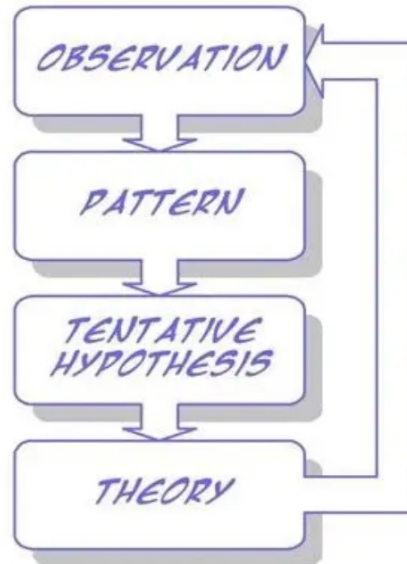
Source: Healey, 2015, p.2.



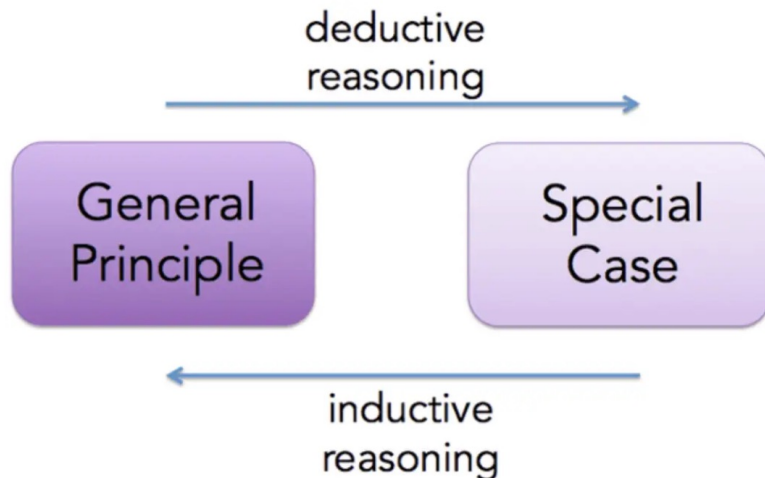
DEDUCTION





INDUCTION



Deductive versus Inductive



<p>I start with theory. I confirm a hypothesis. I tend to do quantitative research.</p>  <p>Deductive</p>	<p>I start with data. I infer conclusions from my data. I tend to do qualitative research.</p>  <p>Inductive</p>
--	---

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.



Coronavirus pandemic, August 24, 2020

#	Country, Other	Total Cases	New Cases	Total Deaths	New Deaths	Total Recovered	Active Cases	Serious, Critical	Tot Cases/ 1M pop	Deaths/ 1M pop	Total Tests	Tests/ 1M pop	Population
	World	23,809,061	+6,189	817,005	+431	16,358,235	6,633,821	61,715	3,054	104.8			
1	USA	5,915,630		181,114		3,217,981	2,516,535	16,483	17,856	547	76,883,479	232,071	331,293,410
2	Brazil	3,627,217		115,451		2,778,709	733,057	8,318	17,046	543	14,144,344	66,473	212,784,888
3	Mexico	563,705	+3,541	60,800	+320	389,124	113,781	3,346	4,365	471	1,263,835	9,787	129,132,739
4	India	3,164,881		58,546		2,403,101	703,234	8,944	2,290	42	35,902,137	25,978	1,382,011,722
5	UK	326,614		41,433		N/A	N/A	72	4,807	610	15,177,265	223,394	67,939,531
6	Italy	260,298		35,441		205,662	19,195	65	4,306	586	8,053,551	133,231	60,448,212
7	France	244,854		30,528		85,199	129,127	399	3,750	468	6,000,000	91,890	65,295,389
8	Spain	420,809		28,872		N/A	N/A	658	9,000	617	8,517,446	182,162	46,757,536
9	Peru	600,438		27,813		407,301	165,324	1,525	18,174	842	3,006,993	91,014	33,038,913
10	Iran	361,150		20,776		311,365	29,009	3,848	4,292	247	3,062,422	36,392	84,150,494
11	Colombia	551,696		17,612		384,171	149,913	1,493	10,825	346	2,508,972	49,231	50,962,919
12	Russia	961,493		16,448		773,095	171,950	2,300	6,588	113	34,600,000	237,077	145,943,991
13	South Africa	611,450		13,159		516,494	81,797	539	10,291	221	3,564,065	59,983	59,418,339
14	Chile	399,568		10,916		372,464	16,188	1,014	20,875	570	2,231,463	116,583	19,140,575
15	Belgium	82,092	+156	9,996	+4	18,242	53,854	89	7,079	862	2,144,563	184,921	11,597,214
16	Germany	236,117		9,336		209,600	17,181	245	2,817	111	10,197,366	121,652	83,824,401
17	Canada	125,647		9,083		111,694	4,870	62	3,325	240	5,169,166	136,782	37,791,278
18	Argentina	350,867		7,366		256,789	86,712	1,960	7,753	163	1,105,878	24,435	45,257,261
19	Indonesia	155,412		6,759		111,060	37,593		567	25	2,056,166	7,506	273,950,524
20	Iraq	207,985		6,519		150,389	51,077	661	5,154	162	1,457,665	36,125	40,350,522

Source: <https://www.worldometers.info/coronavirus/>.

Coronavirus pandemic, August 31, 2021

#	Country, Other	Total Cases	New Cases	Total Deaths	New Deaths	Total Recovered	New Recovered	Active Cases	Serious, Critical	Tot Cases/ 1M pop	Deaths/ 1M pop	Total Tests	Tests/ 1M pop	Population
	World	218,171,757	+278,500	4,527,970	+4,700	195,040,717	+304,214	18,603,070	113,811	27,989	580.9			
1	USA	39,953,651	+6,943	656,482	+89	30,945,115	+650	8,352,054	25,541	119,888	1,970	582,550,800	1,748,051	333,257,237
2	Brazil	20,752,281		579,643		19,692,898		479,740	8,318	96,831	2,705	56,897,224	265,485	214,314,149
3	India	32,808,018	+40,198	438,962	+370	31,982,180	+29,967	386,876	8,944	23,506	314	521,541,098	373,663	1,395,753,675
4	Mexico	3,341,264	+5,564	258,491	+326	2,686,568	+16,627	396,205	4,798	25,603	1,981	9,723,416	74,506	130,505,007
5	Peru	2,149,591		198,263		N/A	N/A	N/A	1,333	64,158	5,917	16,733,426	499,437	33,504,611
6	Russia	6,918,965	+17,813	183,224	+795	6,181,054	+18,624	554,687	2,300	47,388	1,255	178,700,000	1,223,912	146,007,206
7	Indonesia	4,089,801	+10,534	133,023	+532	3,760,497	+16,781	196,281		14,771	480	32,216,075	116,354	276,880,593
8	UK	6,757,650		132,485		5,427,062		1,198,103	982	98,940	1,940	266,714,771	3,905,032	68,300,272
9	Italy	4,534,499		129,146		4,263,960		141,393	548	75,126	2,140	83,728,076	1,387,181	60,358,447
10	Colombia	4,907,264		124,883		4,737,467		44,914	8,155	95,264	2,424	24,121,717	468,271	51,512,348
11	France	6,746,283		114,308		6,225,201		406,774	2,270	103,089	1,747	124,769,146	1,906,579	65,441,374
12	Argentina	5,178,889		111,607		4,869,104		198,178	2,713	113,380	2,443	22,017,526	482,024	45,677,243
13	Iran	4,992,063	+31,319	107,794	+643	4,205,927	+30,522	678,342	7,879	58,565	1,265	28,213,229	330,985	85,240,218
14	Germany	3,950,247	+3,231	92,682	+11	3,738,000	+6,100	119,565	1,096	46,973	1,102	68,329,706	812,527	84,095,254
15	Spain	4,847,298		84,146		4,338,145		425,007	1,685	103,628	1,799	60,618,810	1,295,943	46,775,830
16	South Africa	2,770,575		81,830		2,533,956		154,789	546	46,041	1,360	16,426,011	272,965	60,176,262
17	Poland	2,888,670	+285	75,345	+5	2,657,084	+30	156,241	60	76,423	1,993	19,778,356	523,259	37,798,415
18	Turkey	6,366,438		56,458		5,823,111		486,869	633	74,555	661	76,140,298	891,652	85,392,352
19	Ukraine	2,286,296	+1,356	53,789	+51	2,207,940	+1,257	24,567	177	52,646	1,239	11,980,323	275,866	43,428,075
20	Chile	1,638,675	+345	36,937	+14	1,595,747	+577	5,991	687	84,876	1,913	20,276,691	1,050,240	19,306,720

Source: <https://www.worldometers.info/coronavirus/>.

Coronavirus pandemic, January 17, 2022

#	Country, Other	Total Cases	New Cases	Total Deaths	New Deaths	Total Recovered	New Recovered	Active Cases	Serious, Critical	Tot Cases/ 1M pop	Deaths/ 1M pop	Total Tests	Tests/ 1M pop	Population
	World	331,459,057	+138,304	5,563,652	+219	269,090,164	+64,428	56,805,241	97,247	42,523	713.8			
1	USA	67,631,191		874,321		43,165,667		23,591,203	25,869	202,490	2,618	862,458,737	2,582,225	333,998,303
2	Brazil	23,083,297		621,261		21,710,831		751,205	8,318	107,419	2,891	63,776,166	296,783	214,891,229
3	India	37,618,271		486,784		35,394,882		1,736,605	8,944	26,852	347	705,411,425	503,527	1,400,939,318
4	Russia	10,834,260		321,990		9,878,371		633,899	2,300	74,191	2,205	246,800,000	1,690,051	146,031,061
5	Mexico	4,385,415	+17,101	301,469	+59	3,478,130	+34,246	605,816	4,798	33,471	2,301	13,163,932	100,471	131,022,844
6	Peru	2,606,126		203,464		N/A	N/A	N/A	1,038	77,378	6,041	23,289,858	691,497	33,680,346
7	UK	15,305,410		152,075		11,497,602		3,655,733	746	223,644	2,222	434,073,111	6,342,723	68,436,401
8	Indonesia	4,272,421		144,174		4,119,472		8,775		15,369	519	67,715,434	243,593	277,986,279
9	Italy	8,790,302		141,391		6,093,633		2,555,278	1,717	145,717	2,344	156,338,495	2,591,622	60,324,574
10	Iran	6,224,196		132,095		6,066,819		25,282	1,313	72,669	1,542	42,908,102	500,962	85,651,435
11	Colombia	5,568,068		131,130		5,258,204		178,734	342	107,659	2,535	31,171,683	602,704	51,719,680
12	France	14,274,528		127,263		9,198,995		4,948,270	3,895	217,943	1,943	211,520,605	3,229,497	65,496,464
13	Argentina	7,197,323		118,231		6,193,473		885,619	2,099	157,024	2,579	30,753,911	670,959	45,835,727
14	Germany	8,045,348		116,411		7,000,000		928,937	3,212	95,553	1,383	89,622,218	1,064,429	84,197,463
15	Poland	4,323,482		102,309		3,800,051		421,122	1,519	114,430	2,708	28,591,765	756,744	37,782,620
16	Ukraine	3,759,530		98,361		3,556,162		105,007	177	86,769	2,270	17,182,817	396,574	43,328,102
17	South Africa	3,560,921		93,451		3,375,859		91,611	546	58,895	1,546	21,815,463	360,811	60,462,270
18	Spain	8,424,503		90,993		5,331,175		3,002,335	2,251	180,077	1,945	66,213,858	1,415,348	46,782,734
19	Turkey	10,522,099		84,920		9,737,610		699,569	1,128	122,722	990	125,433,490	1,462,964	85,739,301
20	Romania	1,911,546		59,257		1,776,122		76,167	485	100,399	3,112	17,974,573	944,065	19,039,551

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

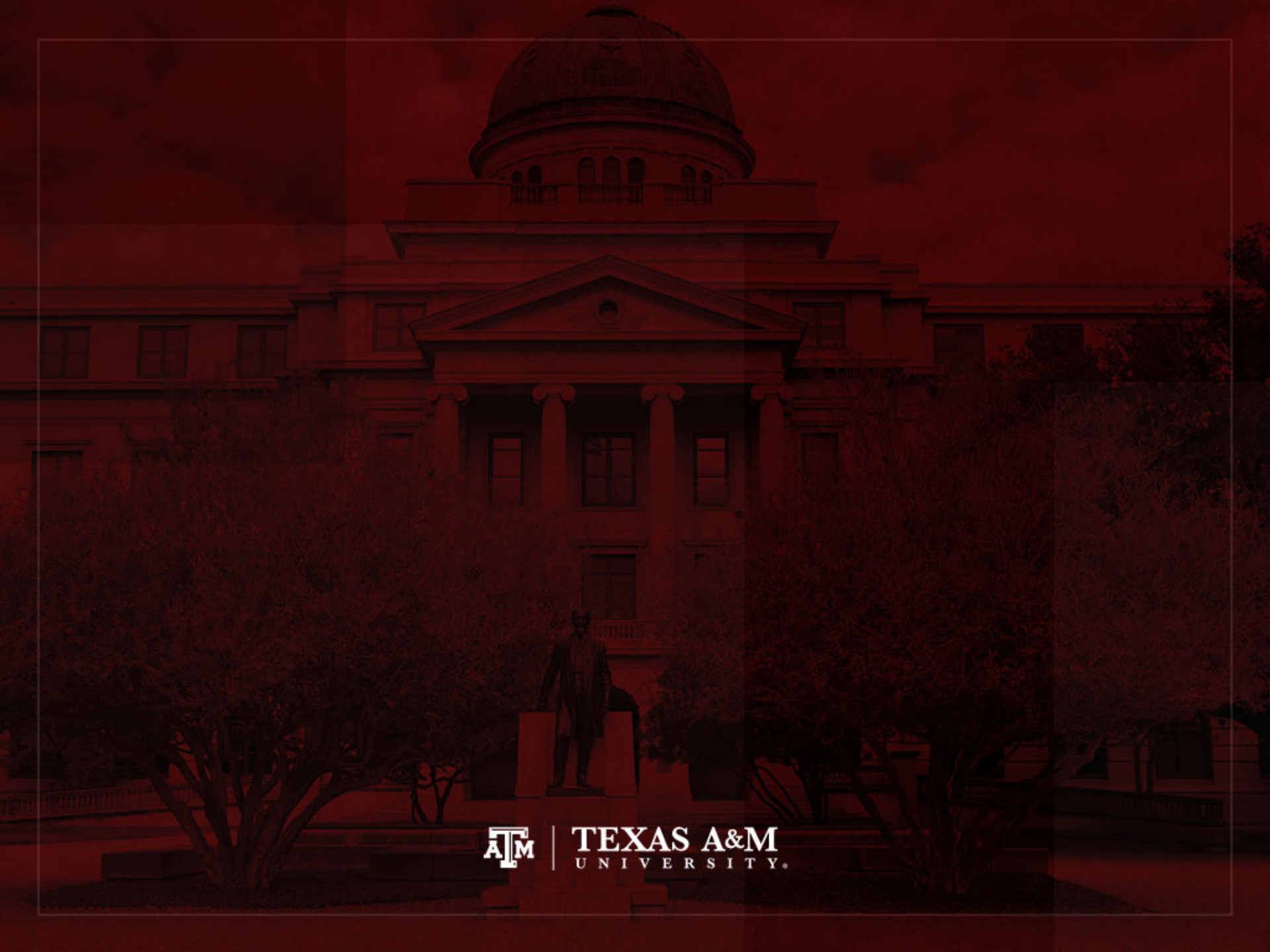
- Statistical analysis of data should be applied after successfully completing earlier phases
 - Rigorous conceptualization and use of theory
 - Well-defined research design and methods
 - Well-conceived research questions
- Review research literature to learn how to
 - Develop and clarify definitions
 - 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	



Association vs. causation

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

Migration  **Earnings**



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



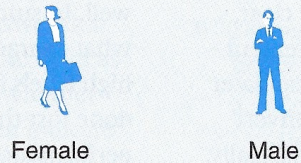
Interval-ratio-level variables

- 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

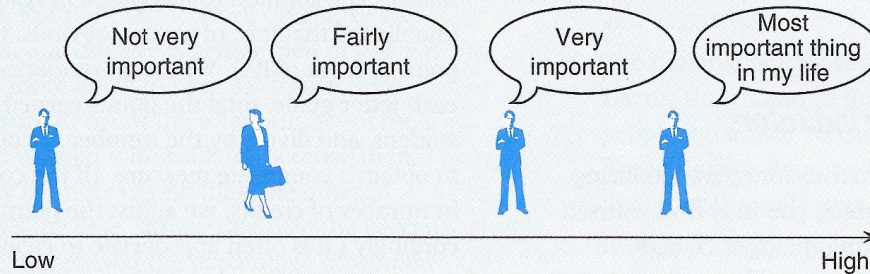


Examples

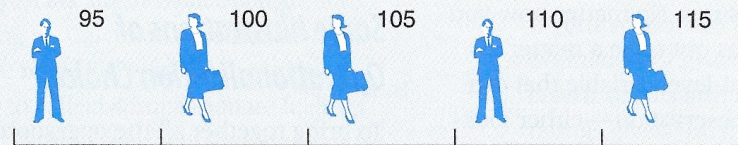
Nominal Measure Example: Gender



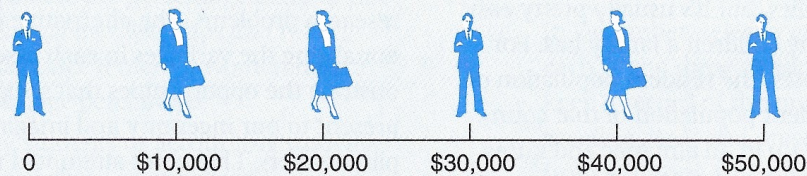
Ordinal Measure Example: Religiosity "How important is religion to you?"



Interval Measure Example: IQ



Ratio Measure Example: Income



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

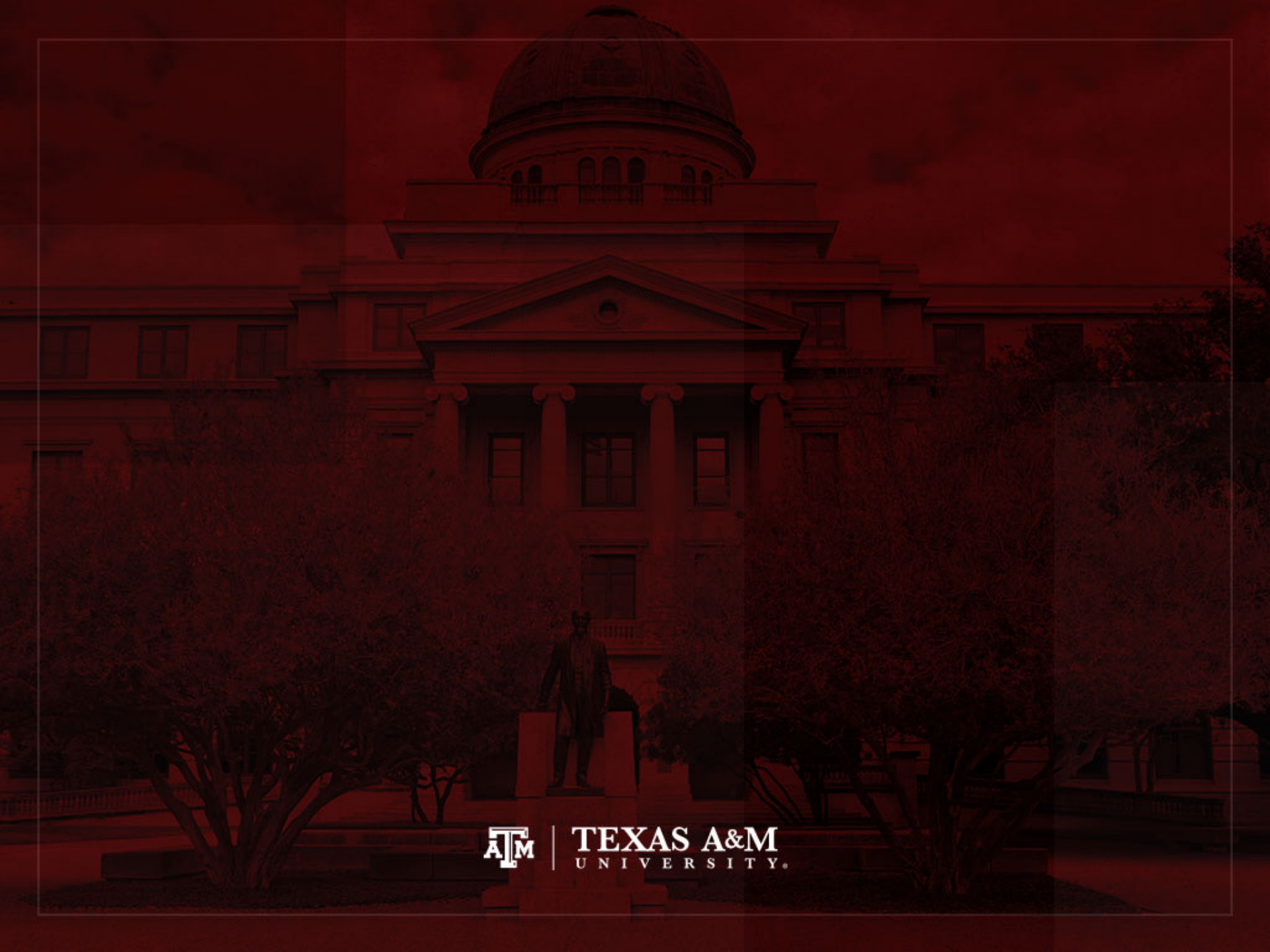
- Always examine the way in which the scores of the variable are actually stated
 - Be careful to look at the way in which the variable is measured before defining its level of measurement
- This is a problem with interval-ratio variables that have been measured at the ordinal level

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	<ul style="list-style-type: none"> – Gender – Race/ethnicity – Religion – Marital status 	<ul style="list-style-type: none"> – Classification into categories – <u>Mode</u> 	<ul style="list-style-type: none"> – Counting number in each category (tabulation) – Comparing sizes of categories 	<ul style="list-style-type: none"> – Chi Square – Logistic regression – Multinomial logistic regression
Ordinal	<ul style="list-style-type: none"> – Social class – Attitude scales – Opinion scales 	<ul style="list-style-type: none"> – All of the above – Plus ranking of categories with respect to each other (scale) – Mode, <u>median</u> 	<ul style="list-style-type: none"> – All of the above – Plus judgments of "greater than" and "less than" 	<ul style="list-style-type: none"> – Spearman's Rho – Ordered logistic regression
Interval-ratio	<ul style="list-style-type: none"> – Age – Number of children – Income 	<ul style="list-style-type: none"> – All of the above – Plus description of scores in terms of equal units – Mode, median, <u>mean</u> 	<ul style="list-style-type: none"> – All of the above – Plus mathematical operations (addition, subtraction, multiplication, division, square roots...) 	<ul style="list-style-type: none"> – 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

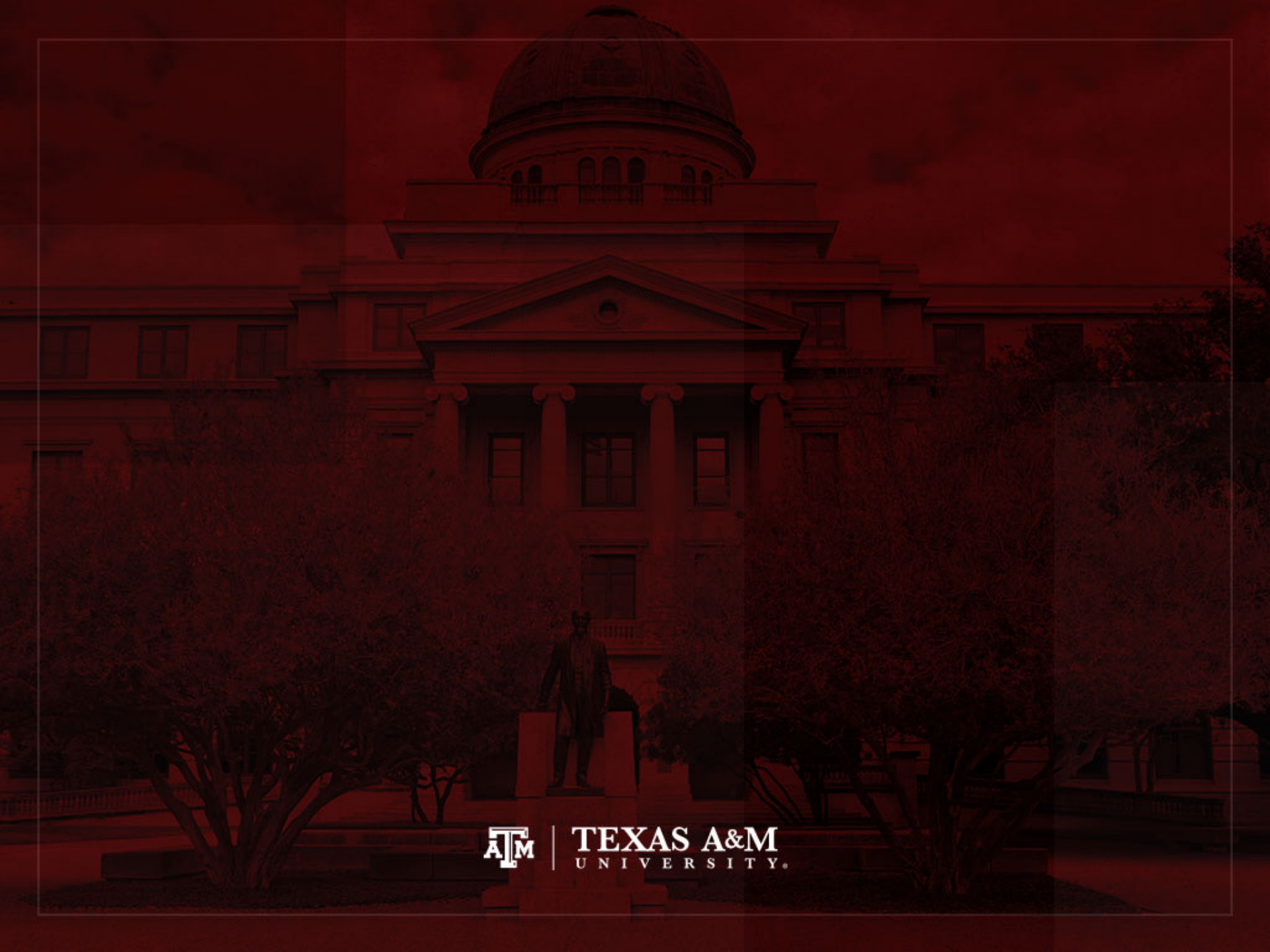
- Social scientists need inferential statistics
 - They almost never have the resources or time to collect data from every case in a population
- Inferential statistics uses data from samples to make generalizations about populations
 - **Population** is the total collection of all cases in which the 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	martye	divorce	widowed	pawrkslf	paocc10	papres10	paidn10	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>

MY GSS > Search Data

Search Data

Years: to [Select specific years](#)

All ▾

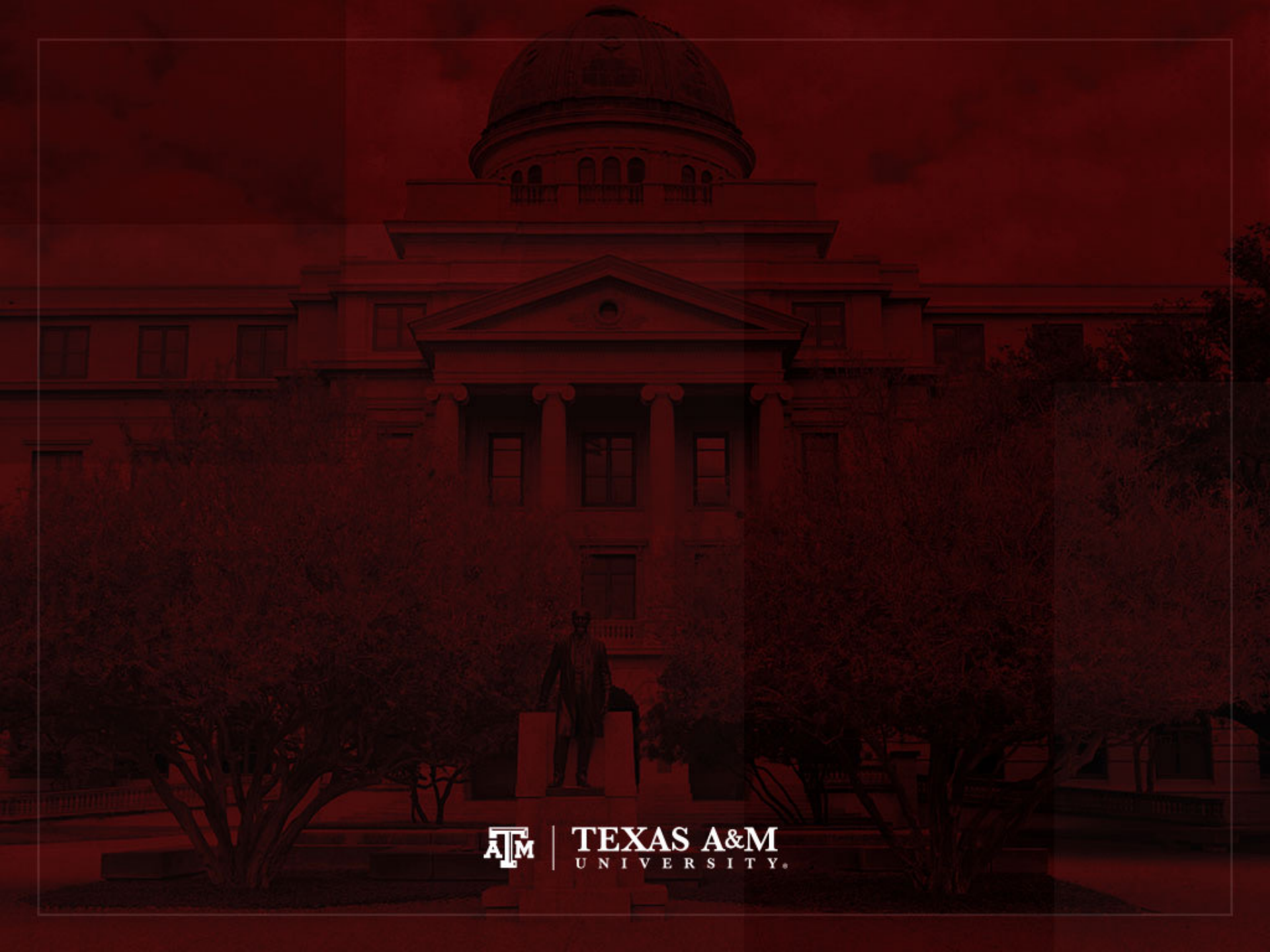
Type Keyword

Filter by:

6404 Results matching criteria

Page 1 of 257 ...

<input type="checkbox"/>	year	GSS year for this respondent		<input type="button" value="+ Add to MyGSS"/>
> Associated Questions				
<input type="checkbox"/>	wrkstat	Labor force status		<input type="button" value="+ Add to MyGSS"/>
> Associated Questions				
<input type="checkbox"/>	hrs1	Number of hours worked last week		<input type="button" value="+ Add to MyGSS"/>
> Associated Questions				
<input type="checkbox"/>	hrs2	Number of hours usually work a week		<input type="button" value="+ Add to MyGSS"/>
> Associated Questions				



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

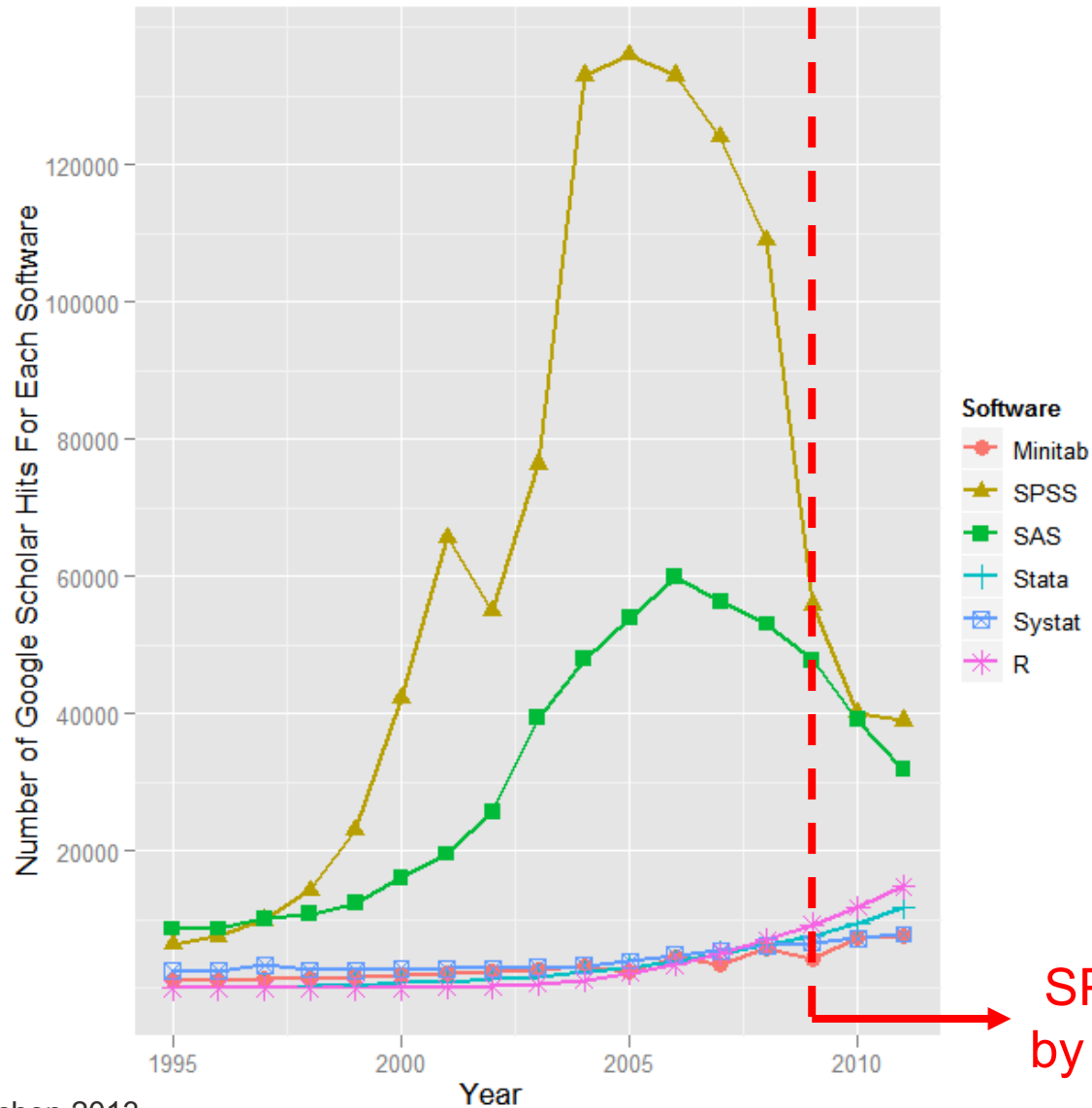
<https://www.r-bloggers.com/forecast-update-will-2014-be-the-beginning-of-the-end-for-sas-and-spss/>

- 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/>

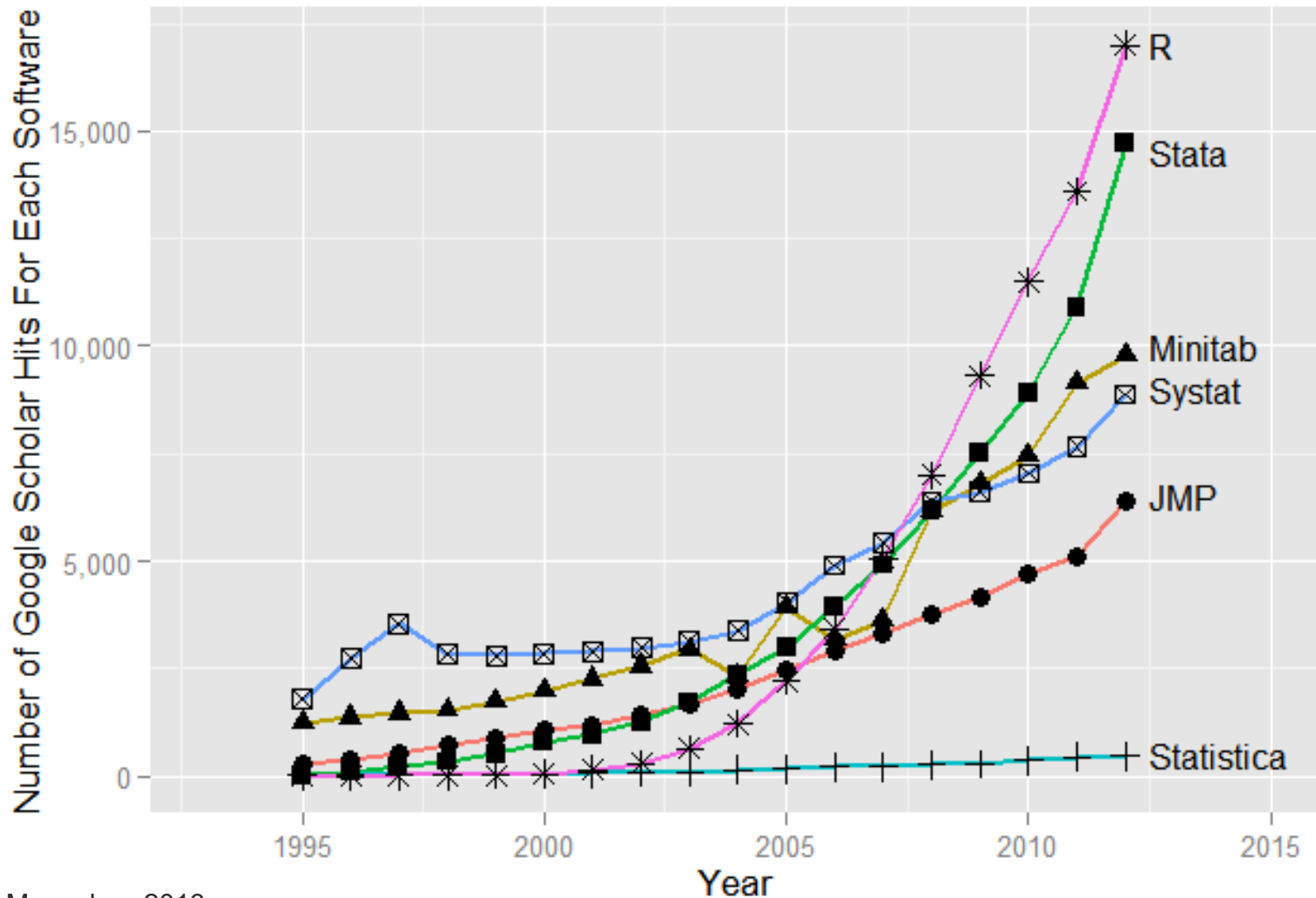
Scholarly use of data analysis software



SPSS was acquired by IBM in 2009

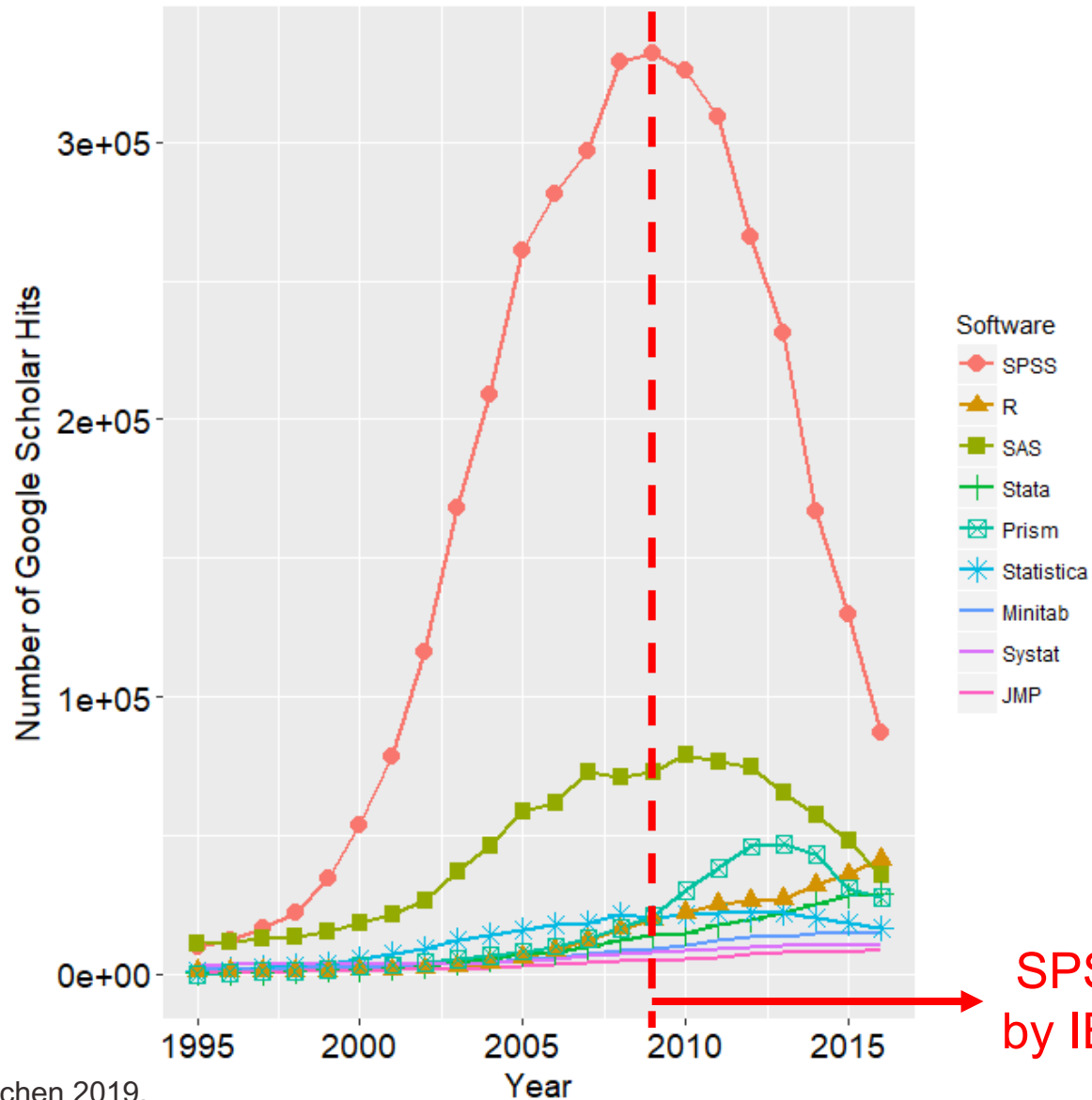
Source: Muenchen 2013.

Scholarly use of data analysis software, SAS and SPSS removed



Source: Muenchen 2013.

Citations per year for each software



SPSS was acquired by IBM in 2009

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/soci420-24spring/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)

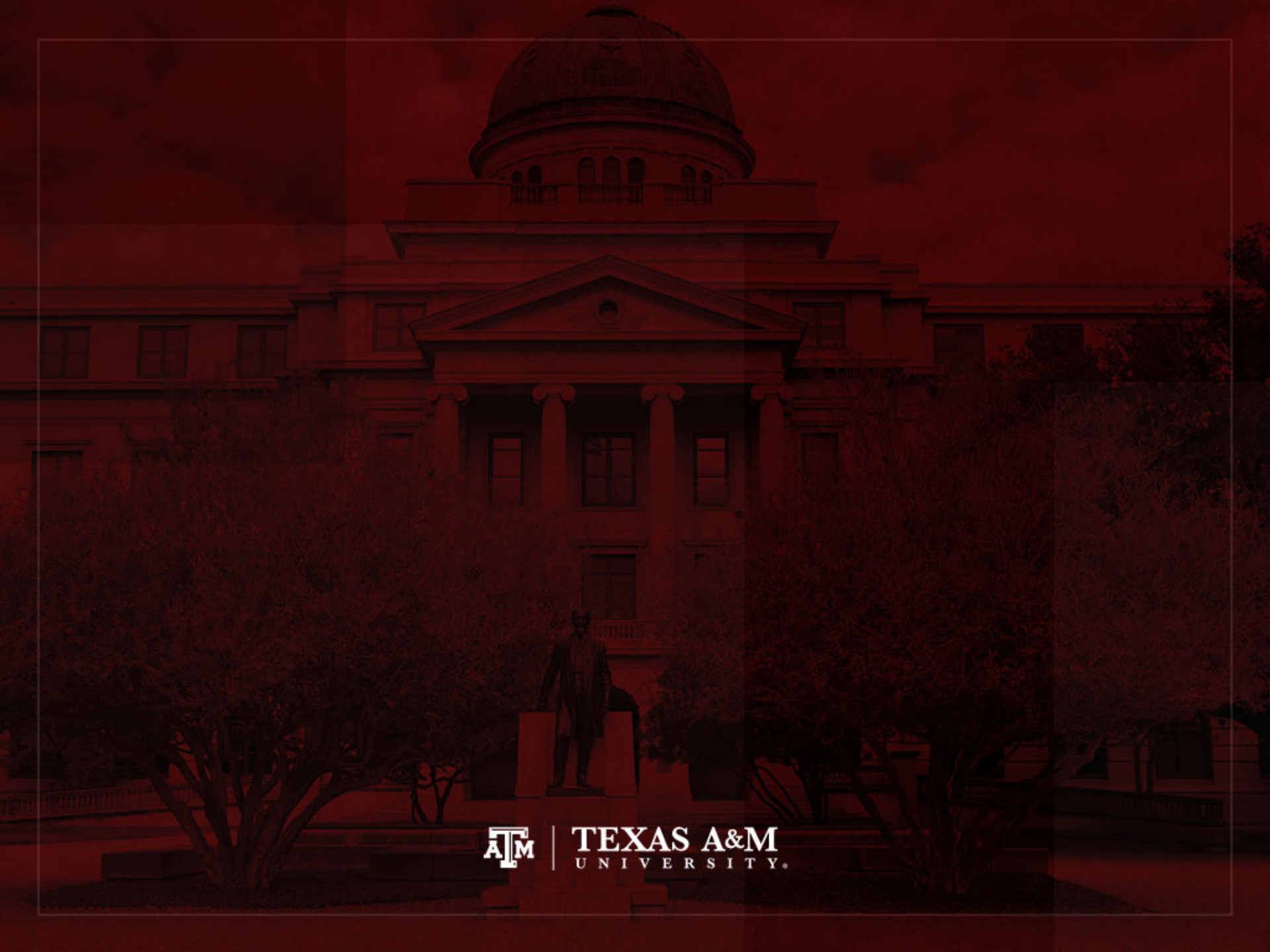
<https://www.stata.com/order/new/edu/gradplans/student-pricing>



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.