Lecture (chapter 1): Introduction

Ernesto F. L. Amaral

January 17, 2018
Advanced Methods of Social Research (SOCI 420)

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

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**

Chapter learning objectives

- Describe the limited but crucial role of statistics in social research
- Distinguish among three applications of statistics
  - Univariate descriptive, bivariate/multivariate descriptive, inferential
- Distinguish between discrete and continuous variables
- Identify/describe three levels of measurement
  - Nominal, ordinal, interval-ratio

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

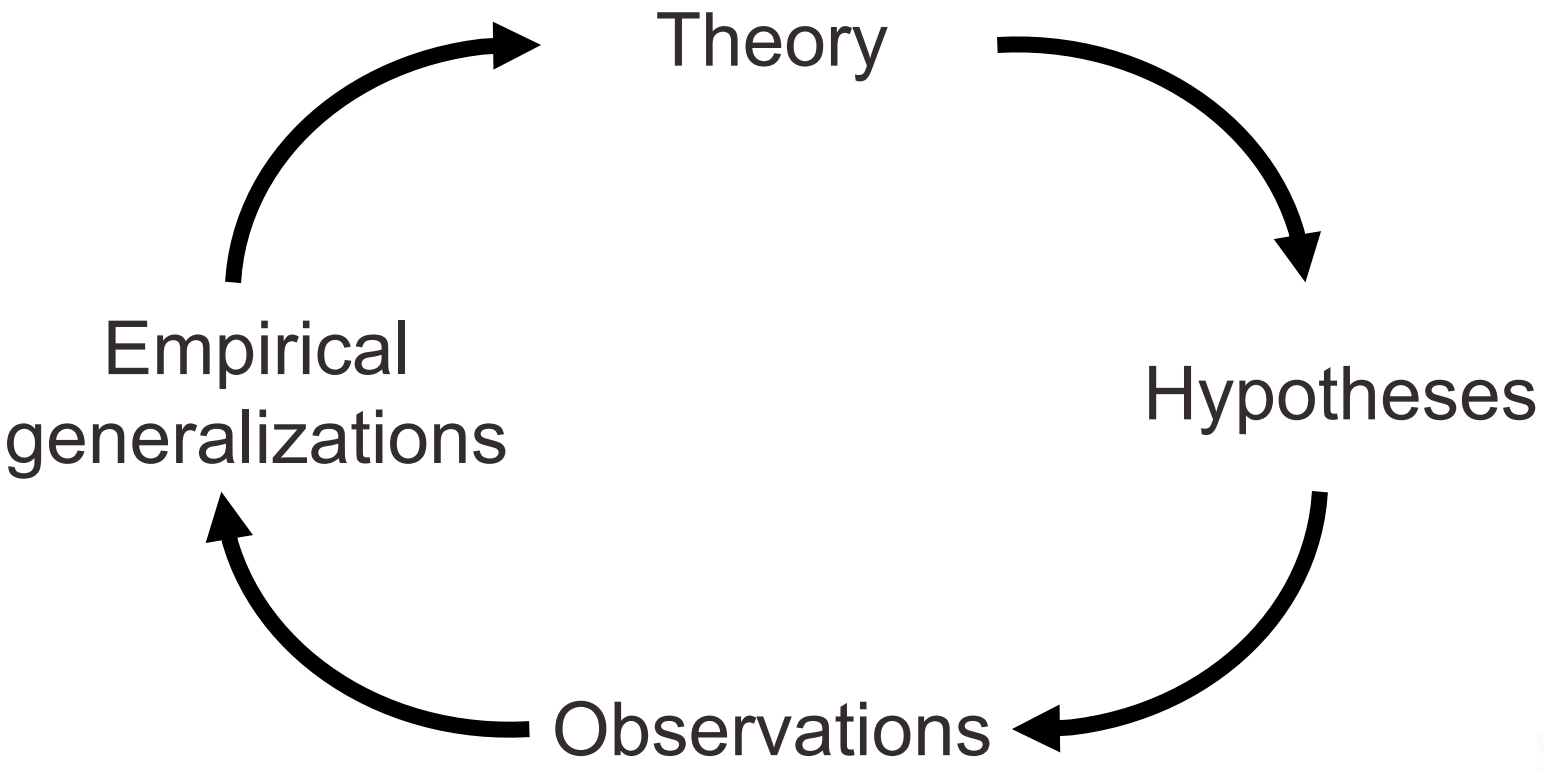
- **Studies without statistics**
  - Some of the most important works in the social sciences do not utilize statistics
  - There is nothing magical about data and statistics
  - Presence of numbers guarantees nothing about the quality of a scientific inquiry
- **Studies with statistics**
  - Data can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

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
- **Hypothesis** is a specific and exact statement about the relationship between variables...

Variables and cases

- **Variables**
  - Characteristics that can change values from case to case
  - E.g. gender, age, income, political party affiliation...

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

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	

Observations

- **Observations** 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
  - Cases 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
- As we discussed, statistics is one of the most important links between theory and research

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)

Types of variables

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

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

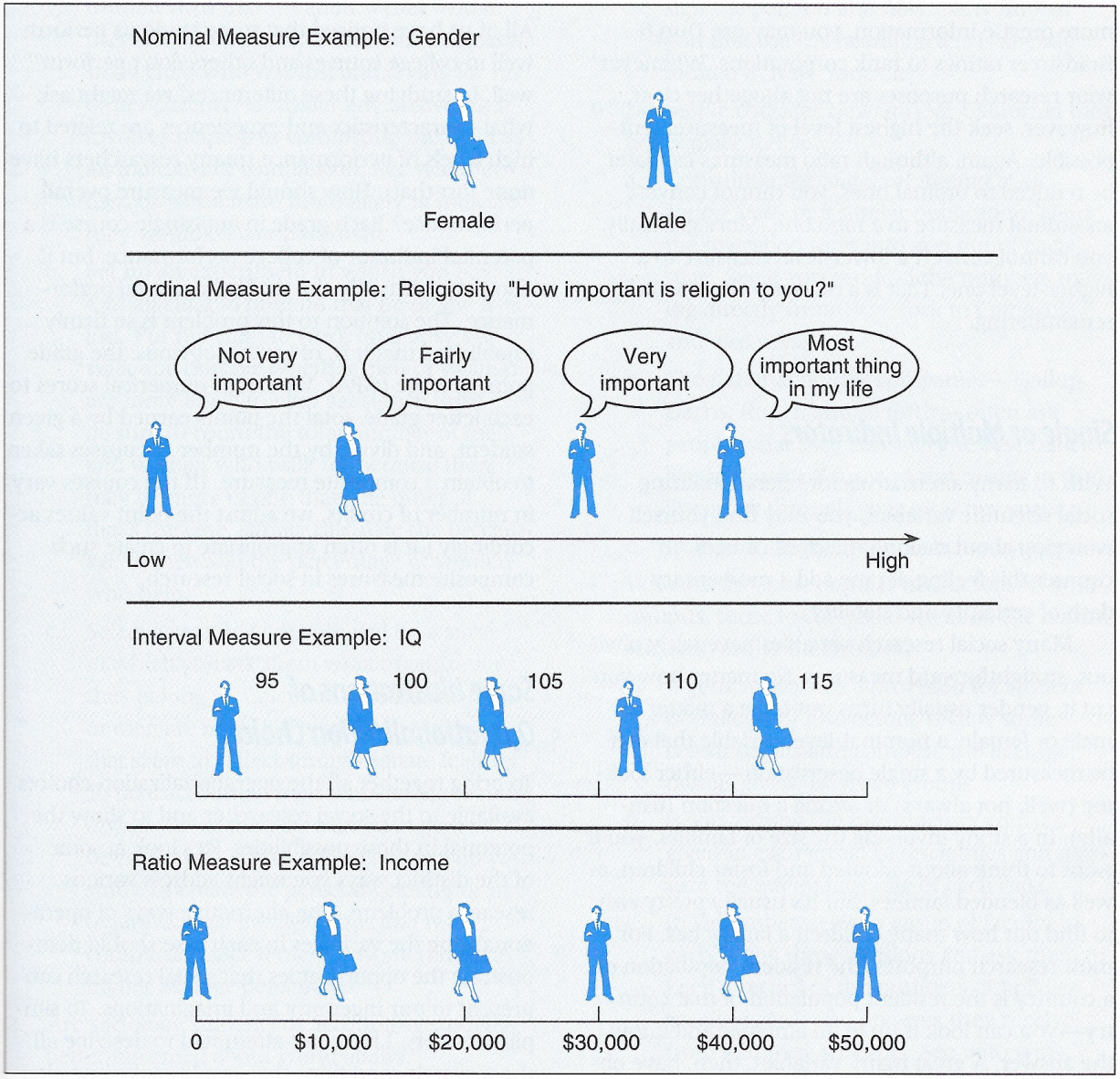
- Attitude and opinion scales
  - Prejudice, alienation, political conservatism...
- Likert scale:
  - (1) strongly disagree; (2) disagree; (3) neither agree nor disagree; (4) agree; (5) strongly agree
- Into which of the following classes would you say you belong?

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.

Three levels of measurement

Levels	Examples	Measurement procedures	Mathematical operations permitted
Nominal	Gender Race Religion Marital status	Classification into categories	Counting number in each category, comparing sizes of categories
Ordinal	Social class Attitude scales Opinion scales	Classification into categories plus ranking of categories with respect to each other	All of the above plus judgments of ”greater than” and ”less than”
Interval-ratio	Age Number of children Income	All of the above plus description of scores in terms of equal units	All of the above plus all other mathematical operations (addition, subtraction, multiplication, division, square roots...)

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

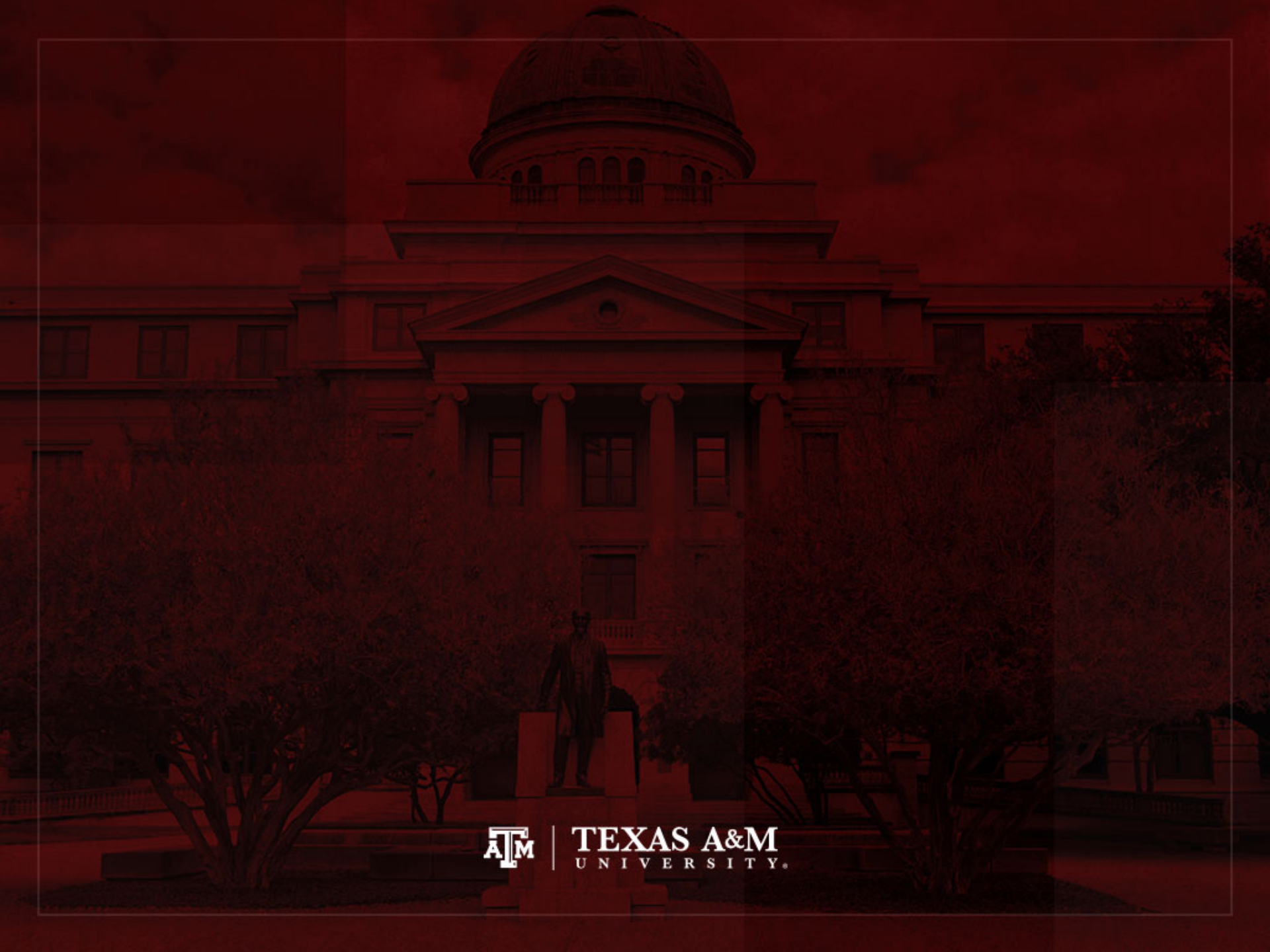
- Nominal-level (e.g. marital status) and ordinal-level (e.g. capital punishment support) variables are almost always **discrete**

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


TEXAS A&M
UNIVERSITY