1.1 Introductory Concepts

STATISTICAL METHODS I
MAT 152
FALL 2022
D. ROTEN

In this lecture…

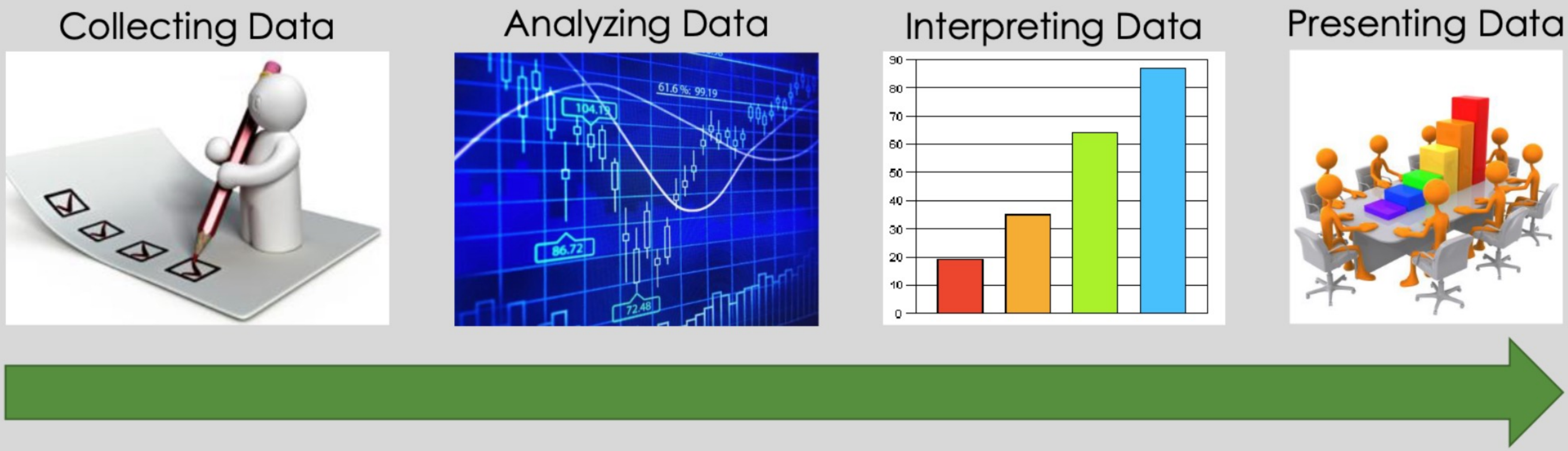
Discuss the terminology associated with statistics

Explore types of data through examples

View data visualization methods

Prerequisites: None

Statistics is a process



Collecting Data

Analyzing Data

Interpreting Data

Presenting Data

What is *statistics*?

- Statistics is:
  - The collection of data
    - What question(s) are being asked?
    - What information needs to be collected?
      - Demographics? Grades? Rainfall totals?
    - How can this information be collected?
      - Surveys? Observations? Instrumentation?
  - The Analysis of data
    - What mathematical tools do we need?
  - The interpretation of data
    - What do our results tell us?
    - Was our hypothesis correct?
  - The presentation of data
    - How can we communicate our findings to others?

Statistics is used to answer questions in almost every field:

Medicine

Economics

Engineering

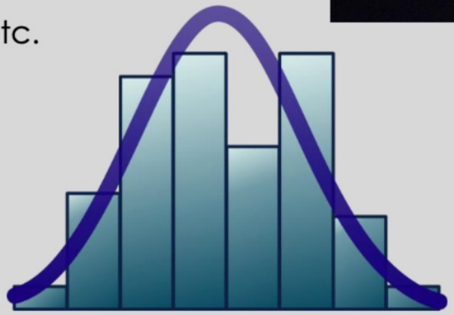
Meteorology

Astronomy

Sports (Sabermetrics)

Construction

Etc.



What is Statistics?

In general, the goal of statistics is to use mathematical concepts/techniques to understand the characteristics of a population.

A population is a collection of people, things, or objects under study.

Often, an entire population is too big to study. (Consider the population of the United States or the wildlife in a national park.) In cases like this, a sample is studied instead.

There are two types of statistics: descriptive statistics and inferential statistics.

- Descriptive statistics deals with organizing and summarizing data using charts, graphs, tables, etc.
- Inferential statistics deals with drawing conclusions from data.

Consider the students at Forsyth Technical Community College…

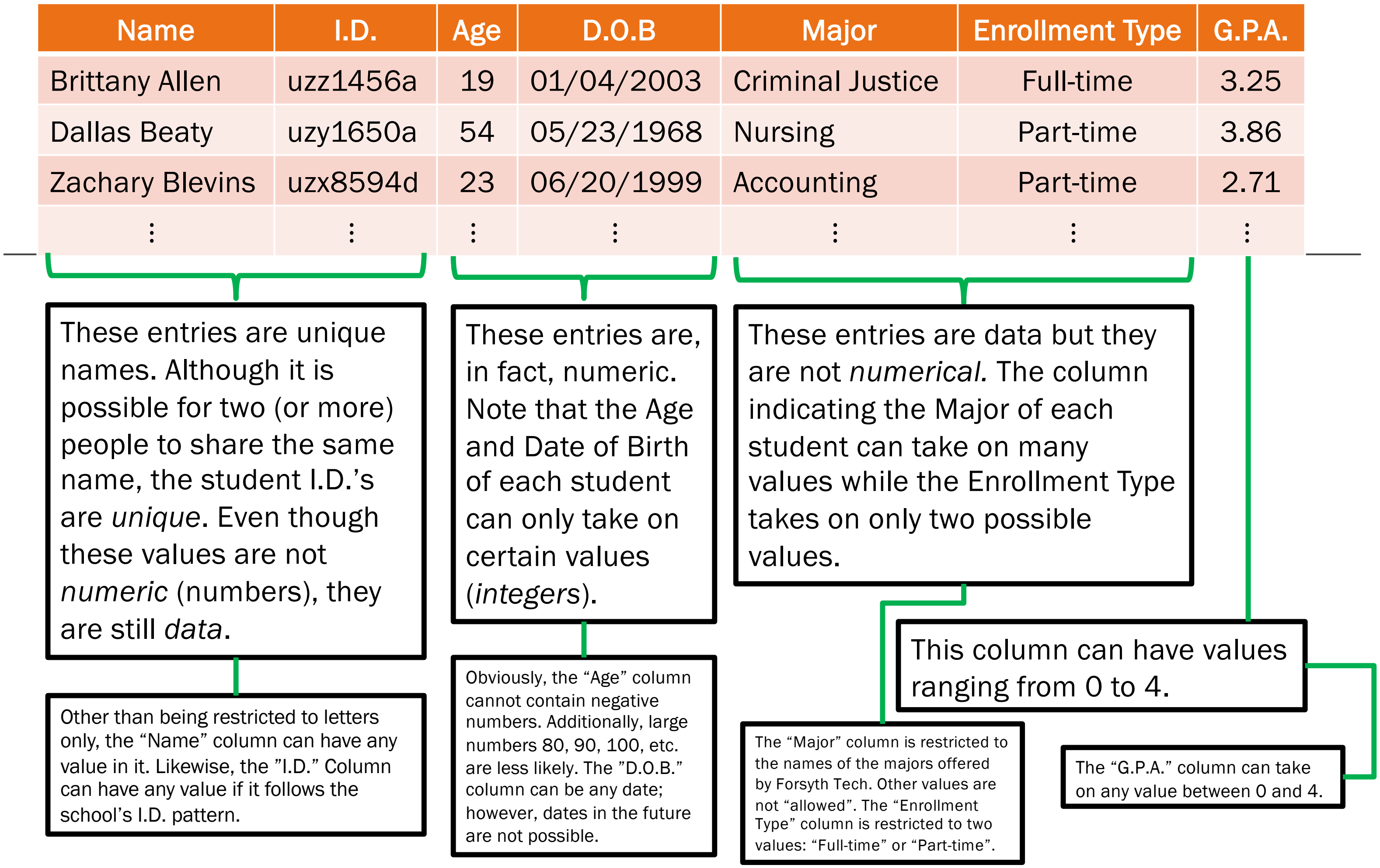
Every student has associated characteristics.

These characteristics vary from student to student.

Consider the first few entries of a database that contains information about every student.

Name	I.D.	Age	D.O.B	Major	Enrollment Type
Brittany Allen	uzz1456a	19	01/04/2003	Criminal Justice	Full-time
Dallas Beaty	uzy1650a	54	05/23/1968	Nursing	Part-time
Zachary Blevins	uzx8594d	23	06/20/1999	Accounting	Part-time
⋮	⋮	⋮	⋮	⋮	⋮

What are the fundamental properties of these characteristics?



Name	I.D.	Age	D.O.B	Major	Enrollment Type	G.P.A.
Brittany Allen	uzz1456a	19	01/04/2003	Criminal Justice	Full-time	3.25
Dallas Beaty	uzy1650a	54	05/23/1968	Nursing	Part-time	3.86
Zachary Blevins	uzx8594d	23	06/20/1999	Accounting	Part-time	2.71
⋮	⋮	⋮	⋮	⋮	⋮	⋮

These entries are unique names. Although it is possible for two (or more) people to share the same name, the student I.D.'s are *unique*. Even though these values are not *numeric* (numbers), they are still *data*.

Other than being restricted to letters only, the "Name" column can have any value in it. Likewise, the "I.D." Column can have any value if it follows the school's I.D. pattern.

These entries are, in fact, numeric. Note that the Age and Date of Birth of each student can only take on certain values (*integers*).

Obviously, the "Age" column cannot contain negative numbers. Additionally, large numbers 80, 90, 100, etc. are less likely. The "D.O.B." column can be any date; however, dates in the future are not possible.

These entries are data but they are not *numerical.* The column indicating the Major of each student can take on many values while the Enrollment Type takes on only two possible values.

The "Major" column is restricted to the names of the majors offered by Forsyth Tech. Other values are not "allowed". The "Enrollment Type" column is restricted to two values: "Full-time" or "Part-time".

This column can have values ranging from 0 to 4.

The "G.P.A." column can take on any value between 0 and 4.

Variables

These students in this dataset are described by a set of variables: Name, I.D., Age, D.O.B, Major, and Enrollment Type.

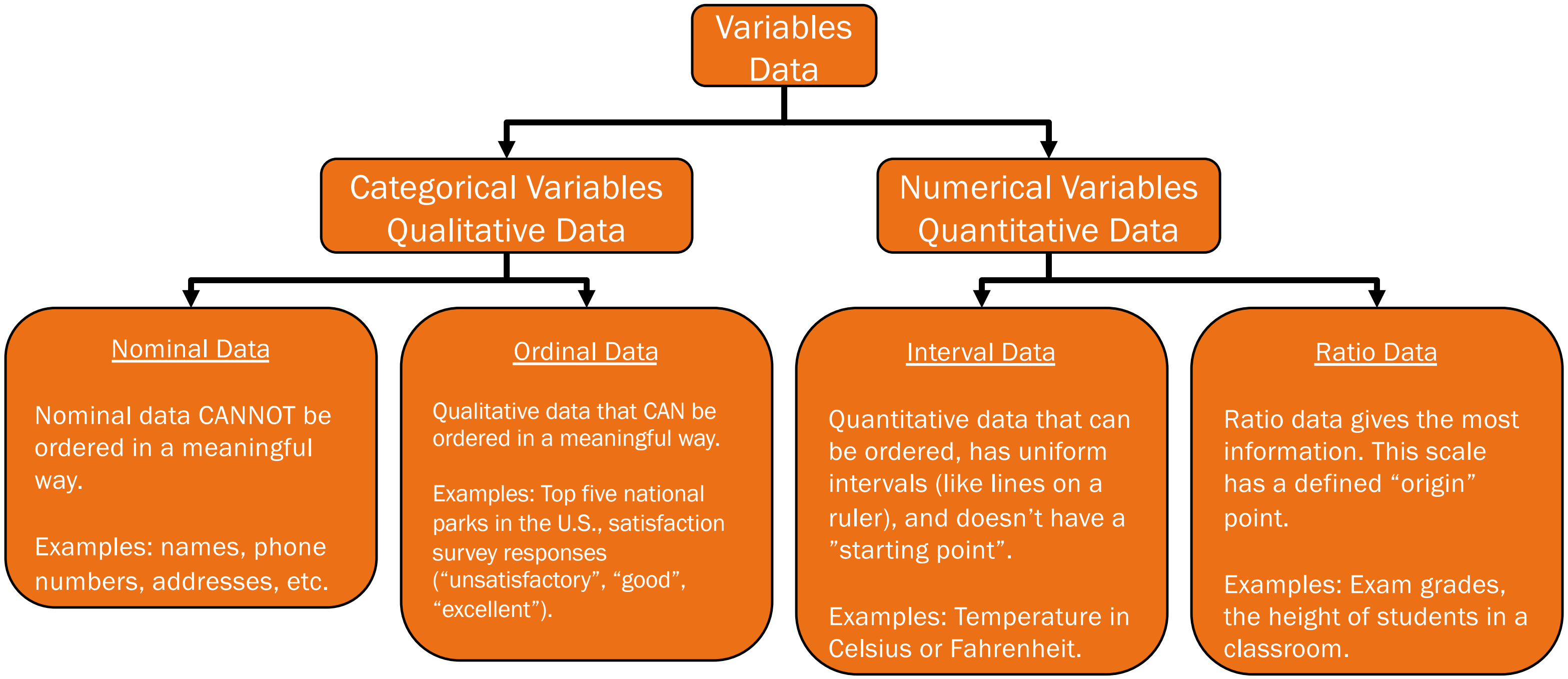
Just like this dataset, any population or sample can be described with a set of variables. These variable can be *restricted* based on the values that they contain.

There are two types of variables: Numerical and Categorical.

- Numerical variables take on values with equal units such as weights (in pounds) or time (in hours).
- Categorical variables place the person or thing into a category.

The data that makes up these two variable types can be further broken down.

Types of Data



Variables
Data

Categorical Variables
Qualitative Data

Numerical Variables
Quantitative Data

Nominal Data

Nominal data CANNOT be ordered in a meaningful way.

Examples: names, phone numbers, addresses, etc.

Ordinal Data

Qualitative data that CAN be ordered in a meaningful way.

Examples: Top five national parks in the U.S., satisfaction survey responses (“unsatisfactory”, “good”, “excellent”).

Interval Data

Quantitative data that can be ordered, has uniform intervals (like lines on a ruler), and doesn’t have a ”starting point”.

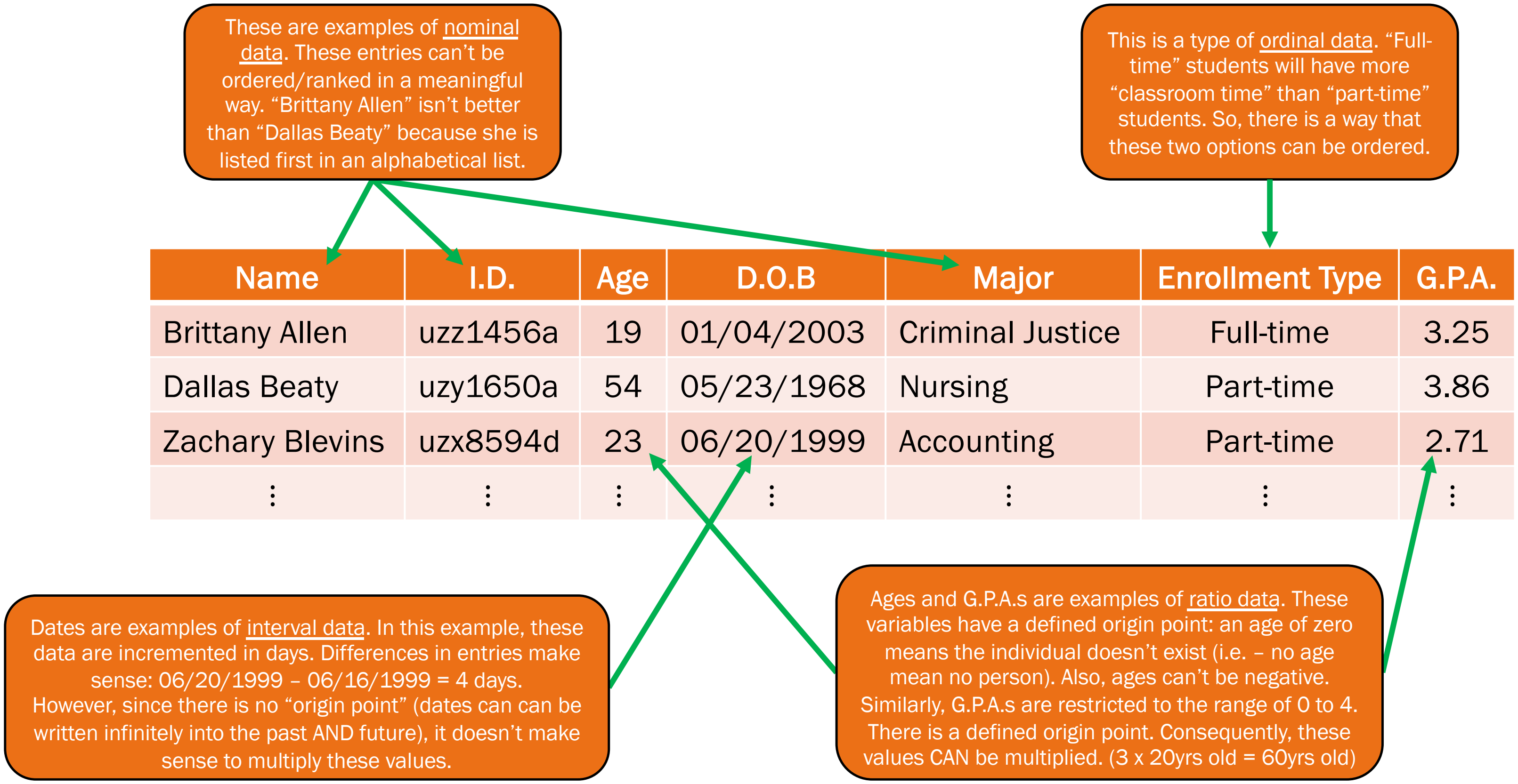
Examples: Temperature in Celsius or Fahrenheit.

Ratio Data

Ratio data gives the most information. This scale has a defined “origin” point.

Examples: Exam grades, the height of students in a classroom.

Let’s classify the types of variables and data in our previous example…



These are examples of nominal data. These entries can’t be ordered/ranked in a meaningful way. “Brittany Allen” isn’t better than “Dallas Beaty” because she is listed first in an alphabetical list.

This is a type of ordinal data. “Full-time” students will have more “classroom time” than “part-time” students. So, there is a way that these two options can be ordered.

Name	I.D.	Age	D.O.B	Major	Enrollment Type	G.P.A.
Brittany Allen	uzz1456a	19	01/04/2003	Criminal Justice	Full-time	3.25
Dallas Beaty	uzy1650a	54	05/23/1968	Nursing	Part-time	3.86
Zachary Blevins	uzx8594d	23	06/20/1999	Accounting	Part-time	2.71
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Dates are examples of interval data. In this example, these data are incremented in days. Differences in entries make sense: 06/20/1999 – 06/16/1999 = 4 days. However, since there is no “origin point” (dates can can be written infinitely into the past AND future), it doesn’t make sense to multiply these values.

Ages and G.P.A.s are examples of ratio data. These variables have a defined origin point: an age of zero means the individual doesn’t exist (i.e. – no age mean no person). Also, ages can’t be negative. Similarly, G.P.A.s are restricted to the range of 0 to 4. There is a defined origin point. Consequently, these values CAN be multiplied. (3 x 20yrs old = 60yrs old)

Summary

Statistics is the study of populations using data collected from samples. Descriptive statistics pertains to organizing and displaying data while inferential statistics focuses on drawing conclusions about populations from samples.

Members of populations (and samples) can be described with a set of variables. There are categorical variables which pertain to non-numeric characteristics and numerical variables which pertain to numeric characteristics.

Data can be further broken down based on the level of measurement it represents. Nominal and ordinal data pertain to categorical variables while interval and ratio data pertain to numerical variables.