Introduction to Categorical Data Analysis

Review from the Previous Statistics Class

Components of a Dataset: Elements, Variables and Data

Example – A random sample of students is selected from a large statistics class.
For each student, their Age,
Major (Art, Business or Science) and
Grade (pass, fail or withdraw)
is recorded into a dataset named ***Students***.

Age	Major	Grade
23	Art	Pass
24	Science	Withdraw
30	Business	Pass
27	Business	Pass
19	Art	Fail
28	Business	Pass
20	Science	Pass
23	Business	Pass
18	Science	Pass
22	Science	Pass
... More Data ...		

*The full dataset '**Students**' can be downloaded from Brightspace.*

Dataset and Elements

A **dataset** is a structured collection of data, usually organized in a table format, where:

- **Rows** represent **elements** which are the objects/individuals that we study and collect data about.

- In the example, the **elements** are the **students** we sample from the statistics class.
- The first student (in the 1st row)
  - 23 years old
  - major in Art
  - passes the class

Age	Major	Grade
23	Art	Pass
24	Science	Withdraw
30	Business	Pass
27	Business	Pass
19	Art	Fail
28	Business	Pass
20	Science	Pass
23	Business	Pass
18	Science	Pass
22	Science	Pass
... More Data ...		

Dataset, Variables and Data

A **dataset** is a structured collection of data, usually organized in a table format, where:

- **Columns** represent **variables** which describes the **characteristics** or **features** measured for each element
- In the example, there are three columns corresponding to three **variables**
  - Age
  - Major
  - Grade
- **Data** are the actual values recorded for a variable. E.g.
  - Age takes on integer values above 18
  - Major takes on three possible values: Art, Business and Science
  - Grade takes on three possible values: Pass, Withdraw and Fail

Age	Major	Grade
23	Art	Pass
24	Science	Withdraw
30	Business	Pass
27	Business	Pass
19	Art	Fail
28	Business	Pass
20	Science	Pass
23	Business	Pass
18	Science	Pass
22	Science	Pass
… More Data …		

Types of Variable / Data: Numerical or Categorical

What is a Numerical Variable?

- **A Numerical variable** is a variable that produces **numerical data** which take on **numbers**.
- These numbers represent measurements or quantities
- We can do math with them, like adding or averaging
- In our dataset, **Age is a numerical variable**

What is Categorical Variable?

- **A Categorical variable** is a variable that that produces **categorical data** which take on **labels**
- These numbers represent group or categories
- We classify elements into distinct categories.
- In our dataset, **Major and Grade are categorical variables**