Introduction

Javier E. Flores

January 23, 2019

Puzzle

Statistics is a lot like working on a puzzle without having all the jigsaw pieces or knowing what the final picture is.

Puzzle

We **collect** whatever pieces we might immediately have around us, and **analyze** each to determine how it might make the bigger picture.

Puzzle

With too few pieces, the picture is unclear and we are forced to find more...

Puzzle

...until eventually we can accurately **describe** the puzzle's picture!

Figure: My dog, "Mellow"

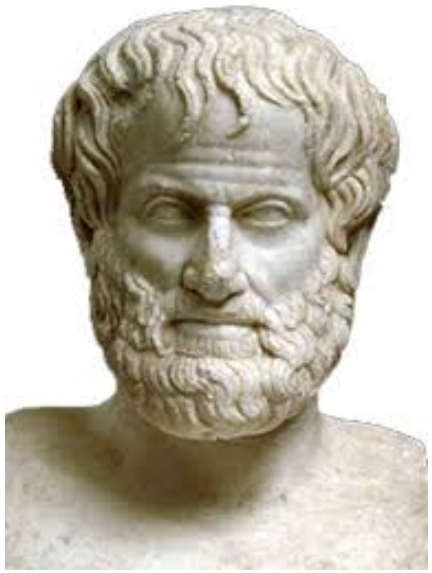
Statistics

Formally defined, **Statistics** is the science of **collecting**, **describing**, and **analyzing data** (e.g. jigsaw puzzle pieces).

However, unlike when working on a puzzle, collected data don't necessarily describe a single picture.

In fact, it is more common for datasets to provide information that answers (describes) multiple questions (pictures)!

Aristotle



"For the things we have to learn before we can do them, we learn by doing them." - Aristotle

Two Real ("Fun") Datasets

In the spirit of Aristotle, we'll begin our foray into statistics by working with two real datasets:

"Student Alcohol Consumption" Data: These data were collected during the 2005-2006 school year from two public schools in Portugal. Data on student attributes such as relationship status, study time, and weekend alcohol consumption are provided.

"Project Blue Book" Data: This dataset contains information on UFO sightings in the United States from the years 1952 to 1969 (during which Project Blue Book, the longest running official U.S. inquiry into UFO sightings, was active).

The Plan

With these datasets, we will:

- Practice loading data into Minitab.
- Describe **cases** and **variables**.
- Classify variables as **categorical** or **quantitative**.
- Practice some basic data operations in Minitab.

Loading Data

To begin, first download each data set.

- Student Alcohol Consumption
- Project Blue Book

Next, read through the relevant instructions provided here. Please note that both of the above datasets are in .csv format.

For now, load only the Student Alcohol Consumption dataset into Minitab.

Student Alcohol Consumption Data Dictionary

School: Gabriel Pereira (GP) or Mousinho da Silveira (MS)

Sex: Female (F) or male (M)

Age: Age (15 - 22 years)

address: Home address type - rural (R) or urban (U)

Pstatus: Parent cohabitation status - living together (T) or apart (A)

Medu: Mother education - none (0), primary education (1), 5th to 9th grade (2), secondary education (3), or higher education (4)

Mjob: Mother job - teacher, health care related, civil services, stay at home, or other

studytime: Weekly study time - $<$ two hours (1), two to five hours (2), five to ten hours (3), or $>$ ten hours (4)

Student Alcohol Consumption Data Dictionary (continued)

failures: Number of past class failures

activities: Extra-curricular activities - yes or no

romantic: In a romantic relationship - yes or no

higher: Wants to take higher education - yes or no

famrel: Quality of family relationships - from very bad (1) to excellent (5)

goout: Go out with friends - from very low (1) to very high (5)

Dalc: Workday alcohol consumption - from very low (1) to very high (5)

Walc: Weekend alcohol consumption - from very low (1) to very high (5)

avg_score: Average of three test scores

Cases and Variables

Now that we've loaded in our data, we'd like to describe the **cases** and **variables**.

A single **case** refers to the individual subject or object we have information on. In the Student Alcohol Consumption data, a case would correspond to a student.

- Cases are generally represented by rows, usually with each case getting a single row.

A **variable** is any characteristic that is recorded for each case. All of the attributes listed in the data dictionary shown previously are variables.

Cases and Variables

Variables may be categorized into one of two general types:

- **Categorical variables** group cases into one of several categories. (e.g. "Mjob").
- **Quantitative variables** record a numeric quantity for each case. (e.g. "Age")

Determining the type of variables contained in our data helps inform how the data should be analyzed.

Creating or Transforming Variables

Minitab allows you to transform or create new variables.

To do so, we first need to create a new variable by creating a new column and typing in the column header. The column header is the variable name.

Next we assign values to each case by:

- manually entering values (not recommended!), or
- by using **Editor** $->$ **Formulas** $->$ **Assign Formula to Column**

Practice:

- Create a new variable "SumScore" that transforms the average of three test scores ("avg_score") into the sum of three test scores.
- Create a new variable "MomHigher" that is "Higher" if the mother's education is higher and "Not Higher" otherwise.

More Variable Types

Earlier I mentioned that variables fall under one of two *general* types. As it so happens, we can further classify variables within each of these general types. Doing so provides additional guidance in selecting an analytic approach.

Categorical variables can be further classified as:

- **Nominal**, where the categories *do not* have a natural ordering (e.g. "Mjob")
- **Ordinal**, where categories *do* have a natural ordering (e.g. "Medu")
- **Binary**, where there are only two exclusive categories (e.g. "Pstatus")

Quantitative variables can be further classified as:

- **Discrete**, where the values are integers, i.e. counting numbers (e.g. "Age")
- **Continuous**, where the values fall anywhere on the real line, i.e. decimal-valued (e.g. "avg_score")

Project Blue Book Data Dictionary

year: Recorded year of sighting

country: Country of sighting (all are US)

state: State of sighting

city: City of sighting

shape: Shape of UFO

duration..seconds.: Duration (in seconds) of sighting

comments: Text description of sighting

Explanatory and Response Variables

Working with data is most fun when you have a specific question you want to answer. For example,

- Do students who more frequently consume alcohol do better on exams?
- Is there a relationship between a student's mother's education level and the student's aspiration for higher education?

Some questions might be answerable using a single variable, but often we are interested in how multiple variables relate to one another.

Explanatory or **predictor** variables are those variables we think will help us explain or predict one or more **response** or **outcome** variables.

Practice

Using the Project Blue Book data, consider the following questions:

- Does the most commonly reported shape of UFO differ by year?
- Are there states in which sightings last longer?

With your group, identify the following for each question:

- The cases
- The explanatory and response variables
- The variable types of the explanatory and response variables.

(Also, have a little fun reading the UFO sighting descriptions ☺)

Wrap-Up

Right now, you should...

- Feel comfortable loading data into Minitab
- Be able to identify cases and variables
- Be able to discern between each of the discussed variable types
- Be able to identify the appropriate explanatory/predictor and response/outcome variables for a scientific question

These notes cover Section 1.1 of the textbook. Please read through the section and its examples, and feel free to continue exploring the data provided in this lecture! ☺