Introduction to Statistics and the Structure of Data

Ryan Miller

What do we mean by 'Statistics'?

In a course about statistics, we should begin by asking ourselves: "What does the term 'statistics' even mean?"

- *"Statistics is the science of collecting, describing, and analyzing data."* Our Lock5 textbook
- *"Statistics is the grammar of science."* Karl Pearson (1936), who established the world's first stats department
- *"Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write."* H.G.Wells (1946), considered "the father of science fiction"
- *"Statistics are no substitute for judgment."* Henry Clay (1852), US politician - secretary of state, senator, and presidential candidate

Personally, I see statistics as the science of making evidence-based generalizations in the face of uncertainty

Two Real Datasets

Like many other fields, modern statistics has become reliant on computers. Thus, we begin our study of statistics by familiarizing ourselves with two real datasets using the software program Minitab.

Our plan:

1. Practice loading these datasets into Minitab
2. Describe their **cases** and **variables**
3. Classify variables as **categorical** or **quantitative** (numeric)
4. Practice some basic data manipulations in Minitab

Happy Planet Data: The Happy Planet Index seeks to assess how well nations are doing at achieving long, happy, and sustainable lives by using data from various sources

Antiquities Act Data: Data used in the FiveThirtyEight article: "Trump Might be the First President to Scrap a National Monument"

Importing Data into Minitab

Link: How to enter and import data in Minitab

Happy Planet Data Dictionary

- **Country**: Name of the country
- **Region**: Code for the region, 1 = Latin America, 2 = Western Nations, 3 = Middle East, 4 = Sub-Saharan Africa, 5 = South Asia, 6 = East Asia, 7 = Former Communist Countries
- **Happiness**: 0 to 10 score from Gallop World Poll data
- **LifeExpectancy**: Average life expectancy (years) from UN Department of Economic and Social Affairs
- **Footprint**: A measure of ecological footprint from *The Edition of the Global Footprint Networks National Footprint Accounts*, higher numbers indicate greater environmental impact
- **HLY**: Happy Life Years - a combined measure of life expectancy and well-being
- **HPI**: Happy Plant Index - a 0-100 score
- **HPIRank**: HPI rank of the country
- **GDPperCapita**: Gross Domestic Product per capita
- **HDI**: Human Development Index from the UN Human Development Report Office
- **Population**: Population (in millions)

Cases and Variables

- **Case** refers to the subject/object/unit that the data has information about
  - Cases are generally represented by rows, usually with each case getting a single row (depending on your research question the data might not come in this format!)
- A **variable** is any characteristic that is recorded for each case (generally a column)

When deciding how to analyze data, it is useful to classify each variable - a variable's type dictates how we should analyze it

- **Categorical variables** divide the cases into groups - they record the category of a case, for example: *Region*
- **Quantitative variables** record a numeric quantity for each case, for example: *LifeExpectancy*

Creating or Transforming Variables

Minitab allows you to transform or create variables. Begin by typing in header of a new column to name the new variable, you can fill the cells (values for each case) either by manually entering values (not recommended!) or by using:

Editor -> Formulas -> Assign Formula to Column

Practice:

1. Create a new variable "TotalGDP" that transforms "GDPperCapita" into the country's total GDP
2. Create a new variable "Asia" that is "Asia" if the country's region is "South Asia" or "East Asia", and "NotAsia" otherwise

More Specific Categorical and Quantitative Variables

Distinguishing categorical from quantitative is important, but sometimes we need to be more specific to ensure the best statistical approach is used

Categorical variables can be further characterized as:

- **Nominal variables** - have no natural order, for example: *Region* in the Happy Planet Data
- **Binary variables** - two exclusive categories, for example: "benign", "malignant"
- **Ordinal variables** - have a natural order, for example: "low", "medium", "high"

Quantitative variables can be further characterized as:

- **discrete** - for example: the integers (1, 2, . . .) or the number of kittens in a litter
- **continuous** - for example: the real numbers (1, 1.1, 1.05, . . .) or the combined weight of the litter

Antiquities Act Data Dictionary

- **current_name**: Name of the piece of land
- **states**: State(s) or territory where the land is located
- **original_name**: Original name of the piece of land (if included)
- **current_agency**: Current land management agency. NPS = National Parks Service, BLM = Bureau of Land Management, USFS = US Forest Service, FWS = US Fish and Wildlife Service, NOAA = National Oceanic and Atmospheric Administration
- **action**: Type of action taken on land
- **date**: Date of action
- **year**: Year of action
- **pres_or_congress**: The president or congress that issued the action
- **acres_affected**: Acres affected by the action, national monuments that cover ocean are listed in square miles

Explanatory and Response Variables

We typically analyze data in order to answer questions, for example:

- Are countries in some regions happier than others?
- Is a higher per capita GDP related to happiness?

Our questions might be answerable using a single variable, but often they requires us to relate multiple variables

- **Explanatory** or predictor variables are hypothesized to help explain or predict one or more **response** or outcome variables

Practice

Using the Antiquities Act data, consider the following questions:

1. How does the acreage of national monuments established under the Antiquities Act differ by year?
2. Do actions taken under the Antiquities Act differ by political party?

With your group, identify the following for each question:

- The cases
- The explanatory and response variables
- The type of variable describing the explanatory and response variables

Discussion

Suppose I collect data on this class and I record your expected graduation year (ie: 2019, 2020, . . .)

- Is this variable quantitative or categorical?
- *Should we analyze* this variable as quantitative or categorical?
  - We often use the **mean** or an **average** to analyze a quantitative variable
  - We often use **proportions** or **percentages** to analyze a categorical variable

Take a minute or two to discuss these questions with your group

Grey Areas

Occasionally there are situations where a variable is technically one type, but it makes more sense to analyze it as another, for example:

- The variable "Year" might technically be a discrete quantitative variable, but if we only have data for 2 or 3 years it makes more sense to treat it as a categorical variable
- A Likert scale variable might technically be an ordinal categorical variable, but it often makes sense to translate it into numeric scores and treat it as quantitative

Conclusion

Right now you should...

1. Feel comfortable loading data into Minitab
2. Be able to identify cases and variables
3. Be able to discern between categorical and quantitative variables
4. Be able to determine the relevant explanatory and response variables for a scientific question

These notes cover Section 1.1 of the textbook, I encourage you to read through the section and examples