Lesson 5. What is a Statistical Model? – Part 1

1 Statistical models

- A model is a representation (or simplification) of reality
- A **statistical model** is a mathematical representation of the relationships among random variables
- Some purposes of statistical modeling:
  1. Making predictions, for example:
     - Predicting the price of a car based on its age, mileage, and model
     - Predicting the probability of acceptance to medical school based on GPA
  2. Understanding relationships, for example:
     - After taking mileage into account, how is the age of a car related to is price?
     - How are various measures of a golfer's performance related to the golfer's scoring average?
  3. Testing differences, for example:
     - Is the rate of headache relief for migraine sufferers who take a new medicine sufficiently higher than those in the control group?

Example 1. Suppose we are interested in predicting the price of a used car based on its mileage.

a. In general, do you think more mileage would result in a higher or lower price?
b. If we wanted to describe the relationship with a simple mathematical function, what could we use?

- Statistical models are not deterministic
  1. We don't expect perfectly accurate predictions
  2. We aim to explain as much variability as possible, without overfitting
  3. Even though there's randomness and uncertainty, we will still get meaningful results
     - We will quantify how confident we are in those results
     - "All models are wrong, but some are useful." —George Box, statistician

- Form of a statistical model:

  - Y is the
  - X is the
  - f is a
  - ε is the
    - This is the part of the response variable Y that remains unexplained after considering the predictor X
    - We will frequently assume ε is normally distributed, and it will be important to check this assumption

2 Terminology

- : The people, objects, or cases on which data are recorded.
- : The characteristics measured/recorded about each observational unit.

- : Records numbers (suitable for arithmetic) about the observational unit.
- : Records a category designation about the observational unit.
- : What we call a categorical variable with only two categories.

- : The variable that measures the outcome of interest.
- : The variable(s) whose relationship to the response is being studied. When the primary goal is to make predictions, we call these **predictor variables**.

- : The group we want to make a statement about. The entire pool from which the sample is drawn.
- : A characteristic about the population.
- : The collected data, gathered from a subset of the population.

- _____ : A characteristic of the sample.

- _____ : When the researcher manipulates the explanatory variable by assigning the explanatory group or value to the observational units (also called experimental units or subjects in this setting). Allows for drawing cause-effect conclusions.
- _____ : When the researchers only observe and record information, as opposed to assigning the explanatory variable. Cannot draw cause-effect conclusions.
- _____ : Additional explanatory variables that are not of primary interest but are included in the model to control for their potential effects.

Example 2. You are interested in whether a midshipman's political inclination and GPA help predict his or her major. So you collect a sample of 50 mids, record each one's political inclination, GPA, and major, and analyze the data.

a. What is the population of interest?
b. Identify the response variable and the explanatory variables, and for each one indicate whether it is categorical or quantitative.
c. Is this an experiment or an observational study?
d. If you were to find a significant association between an explanatory variable and the response, would you be able to say there is a cause-effect relationship?
e. If you find that in your sample of 50 mids, the average GPA is 2.8, is 2.8 a parameter or a statistic?