# Chapter 1: Introduction

Shiwen Shen

University of South Carolina

2017 Summer I

# What Is Statistics?

**Definition: Statistics** is the science of data; how to interpret data, analyze data, and design studies to collect data.

- Statisics is used in all disciplines; not just in engineering.
- "Statistics get to play in everyone else's back yard." (John Tukey)
- John Tukey: `https://en.wikipedia.org/wiki/John_Tukey`

# Statistics Examples

1. In a reliability (time to failure) study, engineers are interested in describing the time until failure for a electronic device.
2. In an agricultural experiment, researchers want to know which of four fertilizers produces the highest corn yield.
3. In a clinical trial, physicians want to determine which of two drugs is more effective for treating HIV in the early stages of the disease.
4. In a social network analysis, researchers want to know the group patterns among all the users.

# What Do Statisticians Do?

- Statisticians use their skills in mathematics and computing to formulate **statistical models** and analyze **data** for a specific problem at hand.
- Models are then used to **estimate** important quantities of interest, to **test** the validity of proposed conjectures, and to **predict** future behavior.
- Being able to identify and model sources of **variability** is an important part of statistics.

# Example: Variability Matters!

Suppose that I am trying to predict

$$Y = \text{MATH 141 final score}$$

for incoming freshmen enrolled in MATH 141. I randomly sample 50 freshmen students and for each of them, I will record the following variables:

$$x_1 = \text{SAT MATH score}$$

and

$$x_2 = \text{high school GPA}$$

# Example: Variability Matters! (cont.)

A **deterministic model** would take the form

$$Y = f(x_1, x_2),$$

where $f()$ is a function of $x_1$ and $x_2$. ($f()$ could be linear or in other shape.) This model suggests that for a student with values $x_1$ and $x_2$, we would compute $Y$ exactly if the function $f$ was known. For example,

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Clearly, this is not realistic.

# Example: Variability Matters! (cont.)

A **statistical model** for $Y$ might look like something like this:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

where $\epsilon$ is a term that accounts for not only measurement error but also

1. all of the other variables not accounted for (e.g. major, difficulty of exam, study habits, etc).
2. the error induced by assuming a linear relationship between $Y$ and $x_1$ and $x_2$.

# Example: Variability Matters! (cont.)

**Discussion 1:**

- Is this sample of students representative of some larger population? After all, we would like our model/predictions to be useful on a larger scale (and not simply for these 50 students).
- This is the idea behind **statistical inference**. We would like to use sample information to make statements about a larger (relevant) population.

# Example: Variability Matters! (cont.)

**Discussion 2:**

- How sould we estimate $\beta_0$, $\beta_1$, and $\beta_2$ in the model above?
- If we can do this, then we can produce **predictions** of $Y$ on a student-by-student basis (e.g. for future students, etc.)
- This may be of interest to academic advisers who are trying to model the success of their incoming students.
- We can also characterize numerical **uncertainty** with our predictions.
- **Probability** is the "mathematics of unvertainty" and forms the basis for all of statistics.

# Who Is Engineer?

An **engineer** is someone who solves problems of interest to society by the efficient application of scientific principles. The steps in the engineering method are as follows:

1. Develop a clear and concise description of the problem.
2. Identify the important **factors** that affect this problem.
3. Propose a **model** for the problem, using scientific or engineering knowledge of the phenomenon being studied.
4. Conduct appropriate **experiments** and collect **data** to test the model proposed.
5. Refine the **model** on the basis of the observed **data**.
6. Manipulate the **model** to assist in developing a solution to the problem, and draw the conclusion if possible.