Relevance of Statistics to Understanding and Learning About the World

1) Observe

2) Construct a conjecture (hypothesis)

3) Design an experiment

4) Collect data

5) Analyze data (statistical methods)

6) Conclusion
Sampling and Descriptive Statistics

Statistics is the science of collecting, organizing, and analyzing data for the purpose of estimation and making inferences.

Basic Idea: Learn about a population by studying a small subset of the population.

(Learn means to estimate or infer something about the population)
Two Major Branches of Statistics:

1) Descriptive Statistics
   Use graphical displays and numeric summarizations to represent data

2) Inferential Statistics
   Use analytic methods and theory of probability to draw conclusions or make decisions
Obtaining Data:
Random Sampling

Population - The entire collection of objects or people that one desires to study, usually to make an inference about the population, or to estimate a population parameter.

Sample - a subset of the population.
Two types of populations:

1) Real population - consists of a finite number of members (actual objects) from which data can be collected.

   Ex: Pop. of all college students in U.S.
   Pop. of all legally registered automobiles

2) Conceptual population: exists only as a concept, and can be thought of as having an infinite number of members.

   Ex: Tosses of a fair coin
A population is a collection of objects or outcomes about which we want to learn.

A sample is a subset of the population.

A simple random sample of size \( n \) is a sample chosen in such a way that each subset of size \( n \) from the population is equally likely to comprise the sample.

**Ex:** Pop = Current WVU students

Sample = Subset (e.g. 200) current WVU students

(This is a real population)
Ex: Population = all possible outcomes of tossing a fair coin an infinite number of times
_sample_ = toss the coin 30 times
(This is a conceptual population)

A convenience sample is obtained by some method other than a well-defined random method.

Ex: Population = All K-mart Shoppers
Sample = The next 50 shoppers that exit the Morgantown K-mart
Sampling variation- different random samples can produce different results.

The items (observational units) is a sample are independent if knowing the value of one item does not provide information about the values of the other items.

Items in a simple random sample may be treated as independent, so long as the size of the sample is less than 5% of the size of the population.
Other Sampling Methods

**Stratified Random Sampling** - the population is divided into sub-populations, called strata, and a simple random sample is selected from each stratum.

**Cluster Sampling** - items are selected from the population in groups, or clusters.

A **variable** is a characteristic whose value is observed for each unit in the sample.

Data are values which arise from observing a characteristic, i.e., the values of a variable.
TYPES of DATA

Variables can be classified into one of two types:

1) Numeric

2) Categorical

Numeric variables are variables whose values represent quantities. Numeric variables may be further classified as discrete or continuous.
Discrete numeric variables usually arise by counting.

Continuous numeric variables typically arise by measuring. Their values form a range or interval of possible values.

Ex: Discrete:

- Number of cars in parking lot
- Credit Hours
- Number of books you own
EX: Continuous
Height of a person
Amount of time spent studying
Weight of an apple

Q: Is the variable "Age" discrete or continuous?

Categorical variables are variables whose values are non-numeric.
EX: Gender
Class Rank
Blood Type
Stata21 Course Grade

If the possible values of a categorical variable possess a "natural" ordering, the variable is called a ranked or ordered variable.

EX: Course grade (A, B, C, D, F)
Ex: Speed of Cars on High St.

15  21  13  18  23  19
26  16  71  22  14  21

71 is an outlier
Notation

X (or any capital English letter) is a random variable.

When we think of the sample data before collecting the data, we represent it as $X_1, X_2, \ldots, X_n$ (capital X's).

The actual values of the sample data is represented as $x_1, x_2, x_3, \ldots, x_n$ (lower case letters).
**Population Parameter** - a number that describes some characteristic of a population, e.g., $\mu$ the theoretical mean of the population

**Ex:** Mean age of the population of college students

**Ex:** Mean value of an infinite number of tosses of a fair 6-sided die

**Sample Statistic** - a number that describes some characteristic of a sample, e.g., $\bar{X}$ the mean of the sample

**Ex:** Mean age of a (random) sample of 1000 college students
EX: Mean value of 500 tosses of a fair six-sided die

In order to draw conclusions about a population, or to estimate a population parameter, we need a sample that is representative of the population. Even though we do not have information about the entire pop., the data contained in the sample can be used to make inferences about the entire pop.
EX: We can use the sample mean $\bar{X}$ (calculated from the random sample of 1000 college students) to estimate the (unknown) value of the mean age $\mu$ for the entire population of college students.