# Applied Statistical Methods

## Data and Variables

Xuemao Zhang
East Stroudsburg University

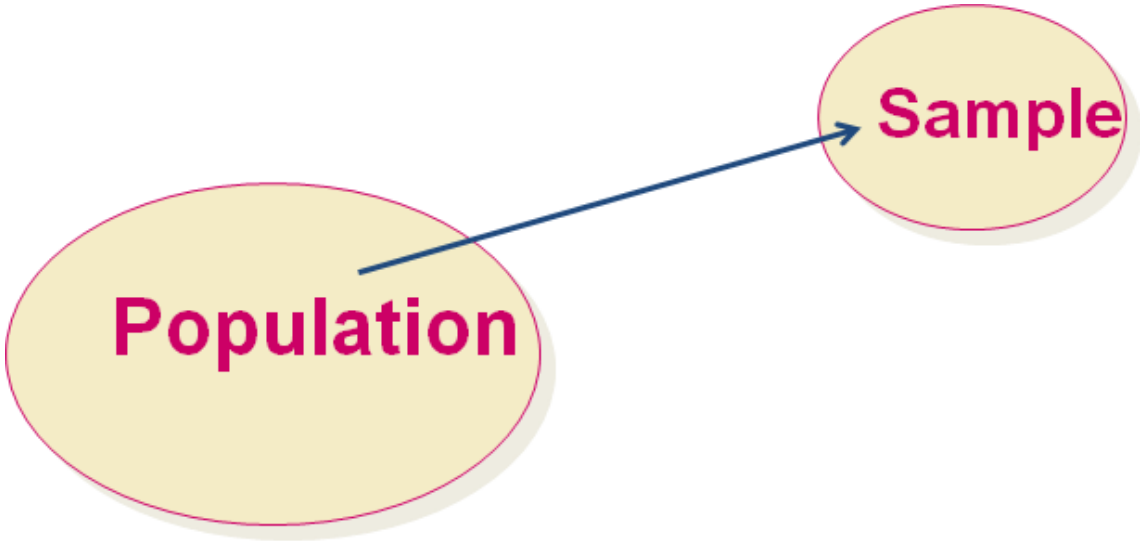
January 20, 2023

# Outline

- Sample and Population
- Types of data
  - Classification of variables and data
- Descriptive statistics and inferential statistics
  - Parameter and statistics
  - Descriptive statistics
  - Inferential statistics

# Sample and Population

- An investigation will typically focus on a well-defined collection of subjects constituting a population of interest.




- Population : The complete collection of all subjects that are being considered.
- Sample: Subcollection of subjects selected from a population.

# Data

- A data set is a collection of measurements of one variable or several variables for some individuals or subjects.
- Often, a data set is a file, in which each column (or field) corresponds to an variable(or attribute) and each row corresponds to measurements of all variables for each subject. This type of data sets is called record-based data.

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear c
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3
```

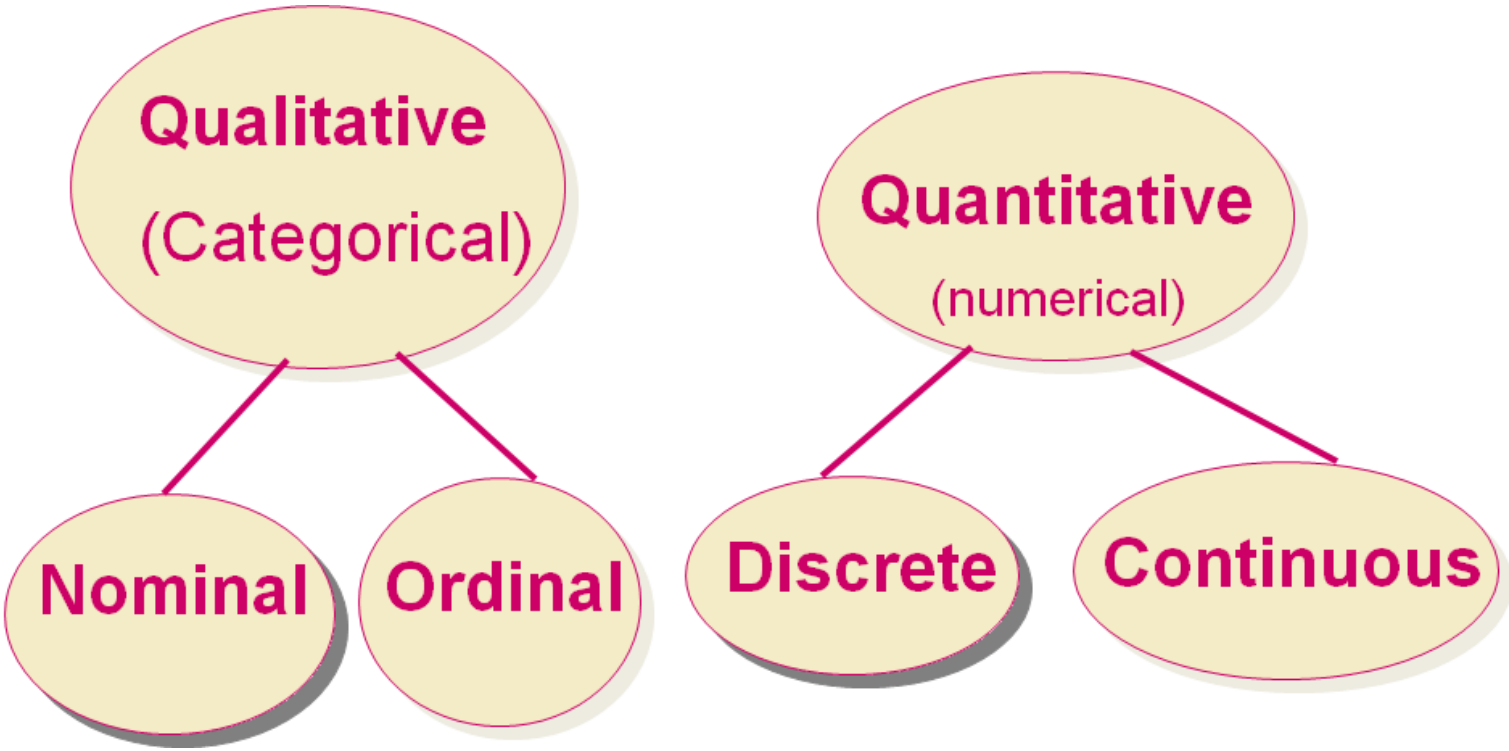
- There are other types of data sets.

# Variables

- A variable (or attribute) is a property or characteristic that can vary from one subject to another or frome one time to another.
  - A data set is obtained by measuring variables.
- Examples
  - Hair color
  - Body temperature
  - time to failure of a computer component.

# Variables

- Classification of variables by the type of measurements



# Categorical Variables

- Categorical variables take category or label/name values, and place an individual into one of several groups.
  - They cannot be used for computations.
- Categorical variables can be further classified using levels of measurement by looking at what is being measured.
  - **Nominal**, when there is no natural ordering among the categories.
    - ⋆ Common examples would be gender, eye color, ethnicity or social security numbers.
  - **Ordinal**, when there is a natural order among the categories, such as, ranking scales or letter grades.
    - ⋆ Examples: Course grades A, B, C, D, or F; Ranks Gold, Silver, Bronze.
    - ⋆ However, ordinal variables are still categorical and do not provide precise measurements.
    - ⋆ Differences between data values either cannot be determined or are meaningless.

# Numerical Variables

- Numerical variables take numerical values, and represent some kind of measurement.
- Numerical variables are often further classified by the number of values:
  - **Discrete**, when the variable takes on a finite or countably infinite number of values.
    - ⋆ Most often these variables indeed represent some kind of **count** such as the number of prescriptions an individual takes daily.
  - **Continuous**, when the variable takes infinitely many values corresponding to the points on a real line interval
    - ⋆ Units should be provided.
    - ⋆ Our precision in measuring these variables is often limited by our instruments.
    - ⋆ Common examples would be height (inches), weight (pounds), or time to recovery (days).

# Univariate and Multivariate data

- A **univariate** data set consists of observations on a single variable.
  - For example, the following sample of lifetimes (hours) of brand D batteries put to a certain use is a numerical univariate data set:

    5.6 5.1 6.2 6.0 5.8 6.5 5.8 5.5
- We have **bivariate** data when observations are made on each of two variables.
  - Example: A data set consists of a (height, weight) pair for each basketball player on a team, with the first observation as (72, 168), the second as (75, 212), and so on.
- **Multivariate** data arises when observations are made on more than one variable (so bivariate is a special case of multivariate).

# Parameter and Statistic

- Parameter is a numerical summary describing some variable of a population.
  - For example, population mean $\mu$, population proportion $p$
- Statistic is a numerical summary describing some variable of a sample
  - A statistic is an estimator of some parameter in a population.
  - For example, sample mean $\bar{x}$, sample proportion $\hat{p}$

# Descriptive statistics

- When desired information is available for all subjects in the population, we have what is called a census.
  - Census is costly and time-consuming. For example, the United States Bureau of the Census is conducting the U.S. Census every ten years.
- DESCRIPTIVE STATISTICS: Procedures used to summarize and describe a set of measurements.
  - Only Descriptive statistics is needed to analyze census data

# Inferential statistics

- INFERENTIAL STATISTICS: Procedures used to draw conclusions or inferences about the population (parameter or distribution) from information contained in a random sample selected from the population.
  - When we cannot enumerate the whole population, we use both Descriptive Statistics and Inferential Statistics.
  - **Probability theory** is necessary for us to make statistical inferences.

# License