Sections 1-2

Types of Data

PARAMETERS AND STATISTICS

- **Parameter:** a numerical measurement describing some characteristic of a ***population***.
- **Statistic:** a numerical measurement describing some characteristic of a ***sample***.

CATEGORIZING DATA SETS

Data sets are sometimes divided into two categories:

1. ***Quantitative (numerical) data*:** numbers representing counts or measurements.
2. ***Categorical (qualitative or attribute) data*:** names or labels that are not numbers representing counts or measurements.

TYPES OF QUANTITATIVE DATA

Quantitative data is further divided into two types:

1. **Discrete** data results when the data values are quantitative and the possible number of values either finite or "countable."
2. **Continuous** data results from infinitely many possible quantitative values, where the possible number of values is not countable.

COUNTABLE VS. CONTINUOUS

- **Countable Data:** If there are infinitely many values, the collection of values is **countable** if it is possible to count them individually, such as the number of tosses of a coin before getting tails. If you select a particular data value from countable data, there is always a "next" data value.
- **Continuous Data:** It is impossible to count the individual items because they are an a continuous scale, such as the lengths of distances from 0 cm to 12 cm. Continuous data can be measured but not counted. If you select a particular data value from continuous data, there is no "next" value.

LEVELS OF MEASUREMENT

Data can also be classified into four **levels of measurement**.

- nominal
- ordinal
- interval
- ratio

NOMINAL LEVEL OF MEASUREMENT

The **nominal level of measurement** is characterized by data that consists of names, labels, or categories only. The data ***CANNOT*** be arranged in an ordering scheme (such as low to high).

EXAMPLES:

1. Majors of college students.
2. Colors of m&m candy.

ORDINAL LEVEL OF MEASUREMENT

Data are at the **ordinal level of measurement** if they can be arranged in some order, but the differences between data values either cannot be determined or are meaningless.

EXAMPLES:

1. Elementary, Middle, High School, College
2. Freshman, Sophomore, Junior, Senior
3. First Place, Second Place, Third Place.

INTERVAL LEVEL OF MEASUREMENT

Data are at the **interval level of measurement** if they can be arranged in order, and differences between data values can be found and are meaningful. Data at this level *DO NOT* have a ***natural*** zero starting point (where ***none*** of the quantity is present).

EXAMPLES:

1. Temperatures
2. Years
3. Hours

RATIO LEVEL OF MEASUREMENT

Data are at the **ratio level of measurement** if they can be arranged in order, differences can be found and are meaningful, and there is a natural zero starting point (where zero indicates that *none* of the quantity is present). For values at this level, differences and ratios are both meaningful.

EXAMPLES:

1. Mileage on an automobile
2. Distance from home
3. Volume

SUMMARY – LEVELS OF MEASUREMENT

Level of Measurement	Brief Description	Example
Nominal	Categories only. Data cannot be arranged in order.	Eye Colors
Ordinal	Data can be arranged in order, but differences either cannot be found or are meaningless.	Ranks of colleges in U.S. News & World Report
Interval	Differences are meaningful, but there is no natural zero starting points and ratios are meaningless.	Body temperatures in degrees Fahrenheit or Celsius
Ratio	There is a natural zero starting point and ratios make sense.	Heights, lengths, distances, volumes

BIG DATA

- **Big data** refers to data sets so large and so complex that their analysis is beyond the capabilities of traditional software tools. Analysis of big data may require software simultaneously running in parallel on many different computers.
- **Data science** involves applications of statistics, computer science, and software engineering, along with some other relevant fields (such as sociology or finance).

MISSING DATA

- A data value is **missing completely at random** if the likelihood of its being missing is independent of its value or any of the other values in the data set. That is, any data value is just as likely to be missing as any other data value.
- A data value is **missing not at random** if the missing value is related to the reason that it is missing.

CORRECTING FOR MISSING DATA

1. **Delete Cases:** One very common method for dealing with missing data is to delete all subjects having any missing values.
   - If the data are missing completely at random, the remaining values are not likely to be biased and good results can be obtained, but with a smaller sample size.
   - If data are missing not a random, deleting subjects having any missing values can easily result in a bias among the remaining values, so results can be misleading.
2. **Impute Missing Values:** We "impute" missing data values when we substitute values for them.

PRACTICAL CONCEPT

When analyzing sample data with missing values, try to determine ***why*** they are missing, then decide whether it makes sense to treat the remaining values as being representative of the population. If it appears that there are values that are ***missing not at random*** (that is, their values are related to reasons why they are missing), know that the remaining data may well be biased and any conclusions based on those remaining values may well be misleading.