Principles of Statistics

Lecture 1

The Language of Statistics

Topics

- Population versus sample
- Parameter versus statistic
- Sample size
- Sample selection (experimental design)
- Types of data
- Descriptive versus inferential statistics

Section 1.2

The difference between the population and a sample of the population

Definitions

- The *population* is everyone or everything you wish to study.
  - Target Population vs. Study Population
- A *variable* is a characteristic of each member of the population.
- A *sample* is a piece of the population.
- A *census* is a study of the entire population.

Definitions (cont.)

- The *sampling error* is the difference between a characteristic of the entire population and a sample of that population.
- The amount of *variation* refers to how different the members of the population are from one another with regard to the variable being studied.

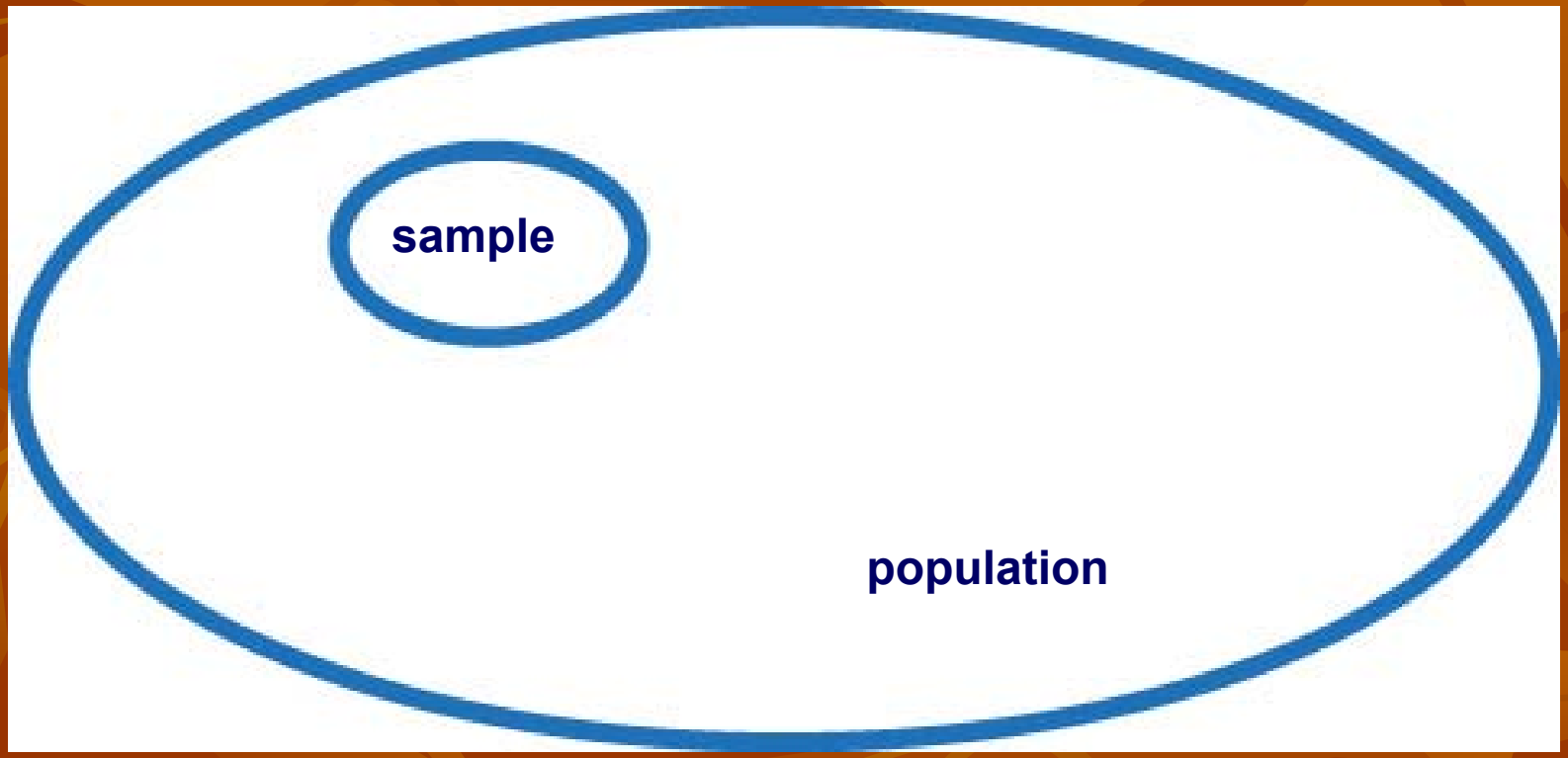
Example – Identifying the Population

- Suppose you wish to study the quality of service at the Marketplace at UCF. What is the population?
  - All UCF students, faculty, staff and guests are potential customers
  - However, some of these individuals may never eat there
  - One possible population: all members of the UCF community who have eaten at the Marketplace at least once in the past year

Example – Identifying Variables of Interest

- Consider the Marketplace example again. What are some variables of interest?
  - Must be related to quality of service
  - Could be measured on each member of the population
  - Some possibilities:
    - Waiting time
    - Quality of food

A *sample* is a piece of the population.



Question: Why sample?

- Too much time and money required to do a census.
- May not be possible to identify all members of the population.
- With destructive testing, a census would leave nothing for further use.

Downside of Sampling

- There will be differences between the sample and the population.
- Greater variation in the measured characteristic among the members of the population translates into greater sampling error.

Upside of Sampling

- Much less time consuming.
- Much less costly.
- The information in a sample may not be a perfect description of the population but is often an adequate reflection.

Example Samples

- Consider the Marketplace example again. Some possible samples might be:
  - Students in this lecture hall who have visited the Marketplace at least once in the past year
  - Faculty in Statistics & Actuarial Science who have visited the Marketplace at least once in the past year

Section 1.3

The difference between a parameter and a statistic

Definitions

- A *parameter* is a number that describes a characteristic of the population.
  - Constant, Unknown, and of Central Interest
- A *statistic* is a number that describes a characteristic of a sample.
  - Random, Available or Computable from the sample
  - Estimates parameter

Connection Between Parameters & Statistics

- A sample is a piece of the population.
- A parameter is a number that describes a characteristic of the population.
- Typically a statistic is defined in such a way as to be similar in spirit to a parameter but calculated on the sample.
- Ultimately we use the statistic to give us insight on the parameter.

Example – Identifying Parameters of Interest

- Consider again the example in which we were interested in the quality of service at the Marketplace. One of the variables of interest was "waiting time". Parameters of interest might be:
  - Average waiting time of all customers last year
  - Maximum waiting time of all customers last year

Example – Identifying Parameters of Interest

- Another variable of interest was "quality of food". Hence parameters of interest might be:
  - Percentage of all customers last year that rated the food quality as "high"
  - Percentage of all customers last year that rated the food quality as "low"

Example – Identifying Statistics of Interest

- Suppose that students in this lecture hall who had visited the Marketplace at least once in the past year were asked to write down the waiting time of their last visit. Statistics of interest might be:
  - Average waiting time of students in this lecture hall during their last visit
  - Maximum waiting time of students in this lecture hall during their last visit

Short Review of Previous Class

- Population, Variable
- Sample vs. Census
- Sampling Error, Variation
- Parameter vs. Statistic
- Example:

  Consider a project in which the objective is to study the relationship between age and systolic blood pressure.

Topics – Sign Posting

- Population versus sample
- Parameter versus statistic
- Sample size
- Sample selection (experimental design)
- Types of data
- Descriptive versus inferential statistics

Section 1.4

Factors that influence sample size: some sampling and sample size considerations

Definitions

- The *size of the population* is the number of members of the population. Its symbol is *N*.
- The *size of the sample* is the number of members in the sample. Its symbol is *n*.

Factor #1: *N*

- With all other factors remaining constant, a larger *N* generally requires larger *n*.
- Not necessarily the most important factor affecting *n*.

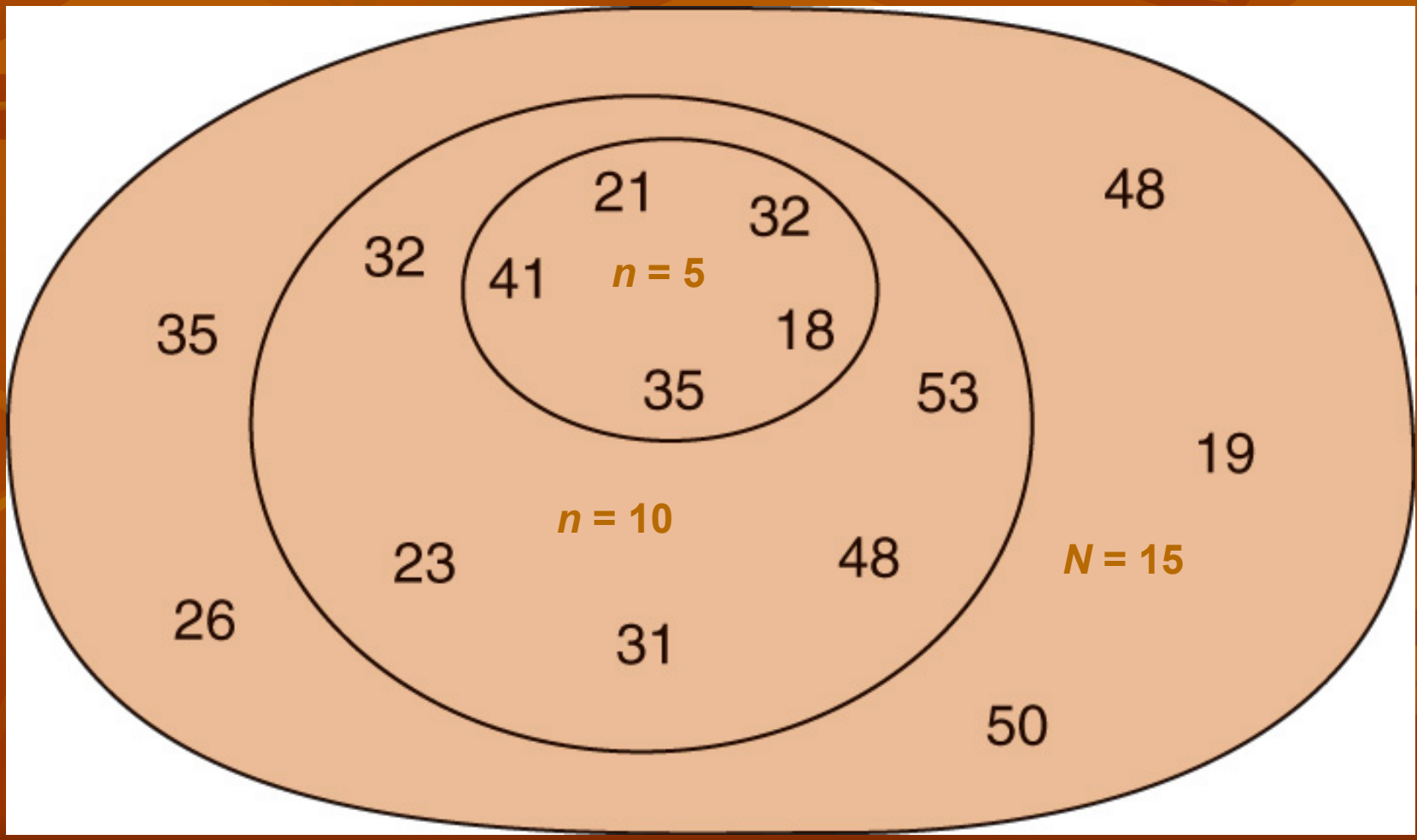
Factor #2: Resources

- Time
- Money
- Equipment
- Personnel
- More of any of these resources usually results in a larger *n*.
- Still not the most important factor influencing *n*.

Factor #3: Error Tolerance

- Larger *n* generally translates into smaller sampling error.
- When $n = N$, there is no sampling error.
- The more costly the errors, the greater *n* must be.
- One of the most important factors affecting *n*.

Figure 1.2 – Larger and larger samples (ages)



Example – Sampling Error

- Parameter of interest: mean age (population)

$$\mu = \frac{35 + 26 + 50 + \cdots + 32}{15} = 34.1$$

- Statistic of interest: mean age (sample)

$$\bar{x} = \begin{cases} \dfrac{21 + 41 + 35 + 18 + 32}{5} = 29.4, & n = 5 \\ \dfrac{32 + 23 + 31 + 48 + 53 + 21 + 41 + 35 + 18 + 32}{10} = 33.4, & n = 10 \end{cases}$$

Example - Sampling Error

- Sampling error: $n = 5$

$$|34.1 - 29.4| = 4.7$$

- Sampling error: $n = 10$

$$|34.1 - 33.4| = 0.7$$

Factor #4:Variation

- The less variable the population, the smaller *n* required for a given level of error tolerance.
- The other important factor affecting *n*.

Section 1.5

Selecting the sample

Definitions

- A *biased sample* is a sample that does not fairly represent the population.
- A *simple random sample* (*SRS*) is a sample that has been selected in such a way that all members of the population have an equal chance of being chosen and every sample of size *n* has the same chance of becoming the sample.

Definitions (cont.)

- A *sampling frame* is a list of all members of the population.
- A *table of random numbers* is a list of numbers randomly generated and listed in the order in which they are generated.

Ideal Samples

- Should be a miniversion of the population, i.e., should contain the same key features found in the population. For instance:
  - Should roughly have the same central value
  - Should roughly have the same spread among the values

Example - Biased Sample

- Suppose that in the Marketplace study you sampled the first 50 customers exiting the dining area after noon on 1/10/06. What kind of biases would you expect?
- The average waiting time in the sample would probably be shorter than the population since:
  - the dining hall is not as crowded at the beginning of the semester
  - those individuals leaving at noon probably arrived at 11 am, long before the noon rush

Selecting an Unbiased Sample

- The best way to select an unbiased sample from the most basic populations is to use simple random sampling (*SRS*).

Selecting a *SRS*

- Need a sampling frame.
- Need a way of randomly selecting *n* members from the sampling frame so that
  - Each member has the same chance of being selected
  - All samples of size *n* have the same chance of selection

Some Techniques Used for *SRS*

- Suppose each member of the sampling frame is assigned a unique number (i.e., first member is 1, second member is 2, ...)
  - Generate *n* numbers with a computer using a random number generator that are between 1 and the largest assigned number with no duplicates
  - Use a table of random numbers to select *n* of these unique numbers with the same restrictions as above
  - Write these unique numbers on slips of paper of equal size, place them in a hat, shake vigorously, and remove *n* slips one-at-a-time

Example of SRS – Florida Lottery

- Sampling frame: the numbers 1, 2, 3, . . ., 53.
- Sample size: $n = 6$.
- Sampling mechanism: 53 ping-pong balls labeled 1, 2, . . . , 53, corrected to have equal weight, thoroughly mixed in a large bin and one-by-one 6 balls are forced into 6 tubes.

Section 1.6

Types of data

Definitions

- *Qualitative data* describes a particular characteristic of a sample item. They are most often non-numerical.
- Data that are created by assigning numbers to different categories when the numbers have no real meaning are called *nominal data*.

Definitions (cont.)

- Data that are created by assigning numbers to categories where the order of assignment has meaning are called *ordinal data*.
- *Likert scales* are used to collect information on attitudes, including degree of agreement with a statement, frequency of use, importance of an issue, quality, and likelihood.

Definitions (cont.)

- Data that are inherently numerical are called *quantitative data*.
- *Discrete data* can take on only certain values. These values are often integers or whole numbers. (Count Data)
- *Continuous data* can take one of an infinite number of possible values over an interval on the number line.
  - Interval- and Ratio-Scaled

Qualitative Data

- Also known as categorical data.
- Each possible value of a qualitative variable falls into one of a finite number of categories.
- Is the simplest form of data.

Qualitative Data

- Examples of qualitative data are variables such as:
  - Class standing (freshman, sophomore, junior, etc.)
  - Political affiliation (democrat, republican, independent, etc.)
  - Income class (lower, middle, upper)

Example – Marketplace Study

- Suppose that in the Marketplace study we sampled $n = 10$ students in this class who had dined there at least once in the past year and recorded the "quality of food" they had during their last visit.
- "Quality of food" is a qualitative variable (low, average, high)

Coded Qualitative Data

- Sometimes qualitative variables use numeric values to represent categories (i.e., coded values).
- If these values have no real meaning, these data are referred to as nominal data.
- For example, the values of the variable gender could be coded 1 = “male”, 2 = “female”.

Coded Qualitative Data (Cont.)

- If numbers have meaning, i.e., if they represent ordered categories, these are refereed to as ordinal data.
- The numbers represent a relative ordering.
- However, you are generally unable to tell how far apart the values are based on the relative ordering. (i.e.,Differencing does not make sense.)
- For example, the values of the variable income level could be coded 1 = “low”, 2 = “middle” and 3 = “high”.

Coded Qualitative Data (Cont.)

- However, if person A has income level 1, person B income level 2 and person C income level 3, then we don't know for sure if the difference between A's & B's income is the same as B's & C's.

Coded Qualitative Data (Cont.)

- A common type of question that yields ordinal data uses a Likert scale.
- For instance, to measure agreement with a particular issue or statement you might use the ordinal scale:
  - 1 = “strongly disagree”
  - 2 = “disagree”
  - 3 = “undecided”
  - 4 = “agree”
  - 5 = “strongly agree”

Quantitative Data

- Comes in two flavors:
  - Discrete
  - Continuous
- Discrete data are often integers or whole numbers resulting from counting the number of times something occurs such as:
  - The number of students attending the large lecture of STA 2014 on a given day
  - The number of individuals who support the war in Iraq in a recent telephone poll

Quantitative Data (cont.)

- Data measured using a Likert scale is sometimes treated as discrete.
- Doing this imposes an assumption on the scale that says the difference in opinion between “strongly agree” and “agree” is the same as that between “agree” and “undecided”.

Quantitative Data (cont.)

- Continuous data are most often the result of measurement such as:
- Interval-Scaled
  - Temperature
- Ratio-Scaled
  - Heights of UCF students
  - Weights (actual not ideal) of UCF students
  - The waiting time for a parking space of UCF students

Quiz 1 – Thursday 1/19/05

- Sections 1.2, 1.3, 1.4, 1.5 & 1.6
- Closed book and closed notes.
- Two cheat sheets allowed

Review: Sample

- Factors that Affect Sample Size
  - Population Size
  - Resources
  - Error Tolerance
  - Variation
- Select a Sample
  - Biased vs. Fair Sample
  - Simple Random Sample (SRS)

Review: Types of Data

- Qualitative Data
  - Nominal Data
  - Ordinal Data
    - Likert Scales
- Quantitative Data
  - Discrete Data – Count data
  - Continuous Data
    - Interval-Scaled Data
    - Ratio-Scaled Data

Topics – Sign Posting

- Population versus sample
- Parameter versus statistic
- Sample size
- Sample selection (experimental design)
- Types of data
- Descriptive versus inferential statistics

Section 1.7

The difference between descriptive statistics and inferential statistics

Definitions

- The *tools of descriptive statistics* allow you to summarize the data.
- An *inference* is a deduction or a conclusion about the population.
  - Statistical Inference vs. Guess
  - Reliability Measure
- The *techniques of inferential statistics* allow us to draw inferences or conclusions about the population from the sample.
- We use *probability* theory to calculate the likelihood of observing or selecting a particular sample from a population.

Data's Collected . . . What's Next?

- Summarize the data: descriptive statistics.
- Make conclusions: inferential statistics.
- Descriptive and inferential statistics are complementary tools supporting each other.
- However, their tasks are quite different.

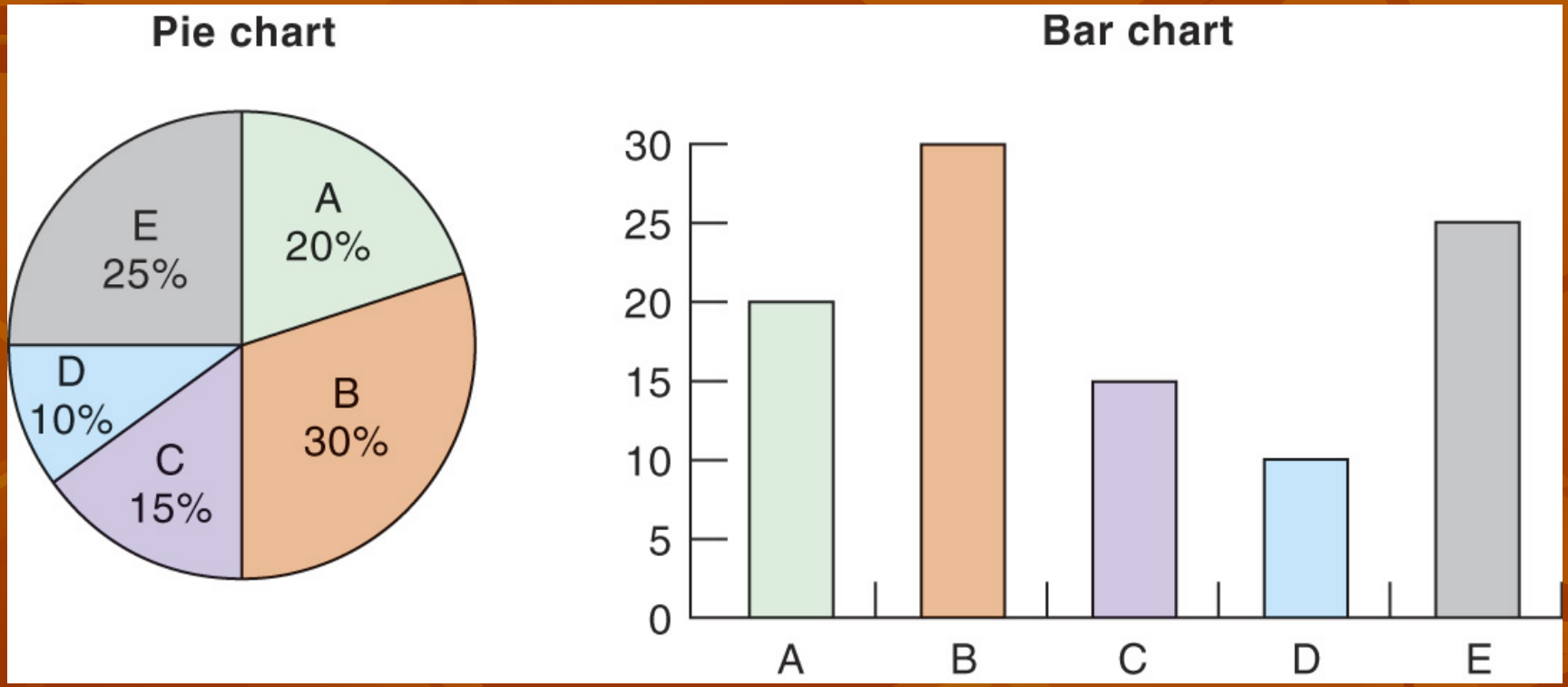
Descriptive Statistics

- Usually the first tools you will use.
- The focus is on the sample with no extrapolation to the population.
- Comes in two flavors:
  - Graphical or visual descriptive tools
  - Numerical descriptive tools

Graphical Tools

- Bar charts.
- Pie charts.
- Histograms.
- Many others.
- Provide a visual summary of the sample.
- A picture is worth a 1,000 words (or numbers)

Figure 1.3 – Pie chart and bar chart



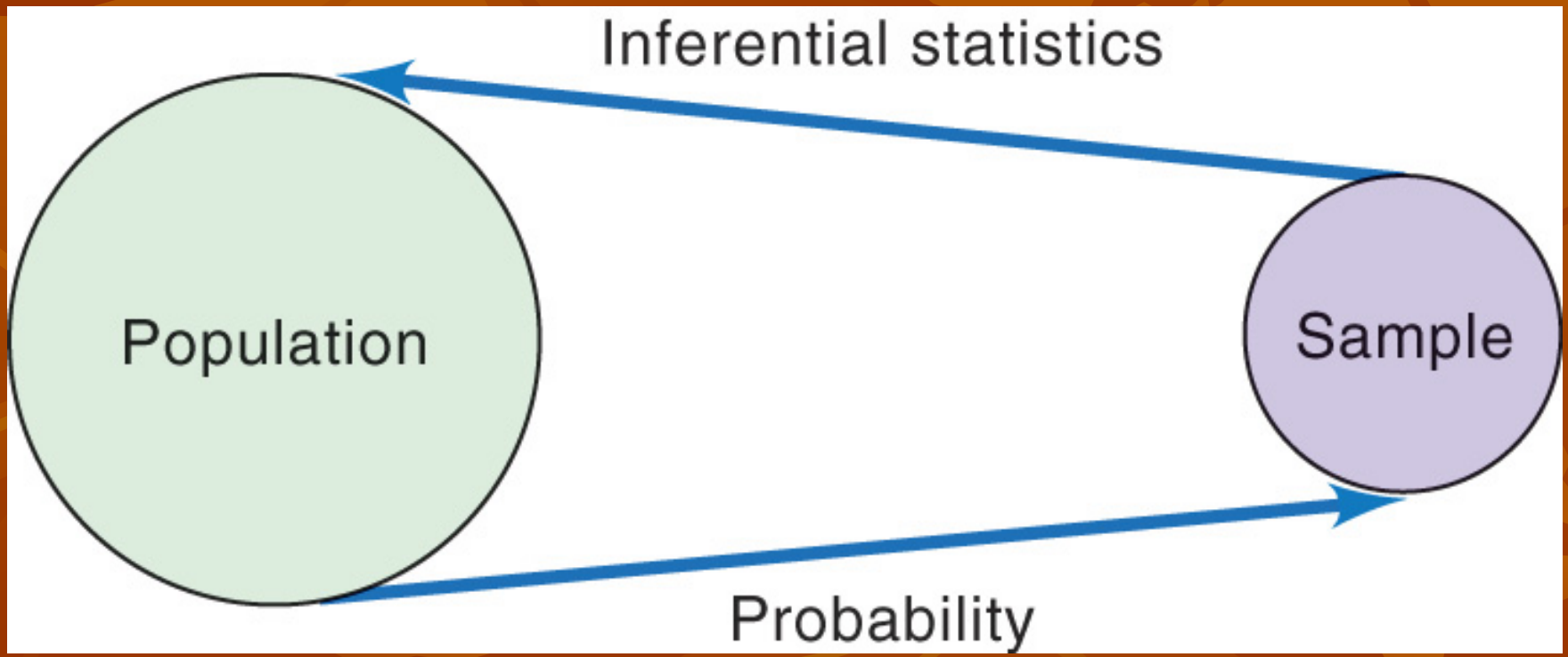
Numerical Tools

- Mean.
- Median.
- Mode.
- Maximum and minimum values.
- Many others.
- Provide a numerical summary of the sample.

Inferential Statistics

- We wish to draw conclusions (inferences) about the population from information in the sample.
- For instance, in hypothesis testing the process is as follows:
  - Make a statement about the population
  - Collect a sample
  - Use probability theory to compute the likelihood of the observed sample from the population
  - Decide to accept of reject the statement using inferential statistics

Figure 1.4 – Relationship between probability and inferential statistics



Inferential Tools

- Statistical Inference is the process of making inference about the population based on the information in a sample.
- Two Forms of Statistical Inference
  - Confidence intervals
  - Hypothesis tests

Basic Summation Notation

- If we refer to the individual data values in a sample of size n as $X_1, X_2, \cdots, X_n$

  Then we could write the sum using the sigma notation

$$X_1 + X_2 + \cdots + X_n = \sum_{i=1}^{n} X_i$$

- For example, service times (in minutes) 23, 16, 5, 8, 25, 19, 10, 5, 4, 13. Then

$$(1) \quad X_1 + X_2 + \cdots + X_{10} = 23 + 16 + \cdots + 13 = \sum_{i=1}^{10} X_i$$

$$(2) \quad |X_1 - 10| + |X_2 - 10| + \cdots + |X_{10} - 10|$$
$$= |23 - 10| + |16 - 10| + \cdots + |13 - 10|$$
$$= \sum_{i=1}^{10} |X_i - 10|$$