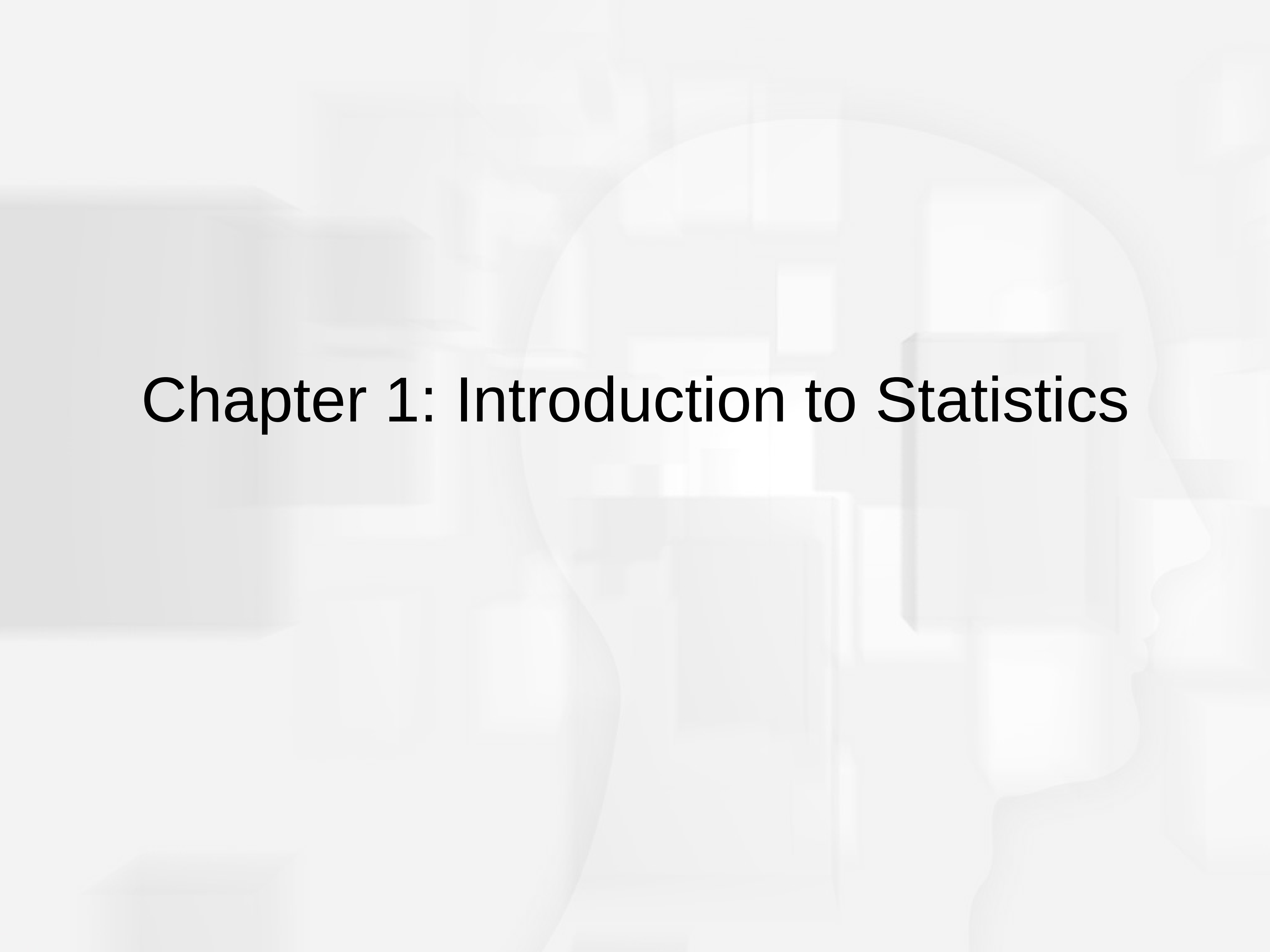


Chapter 1: Introduction to Statistics

Variables

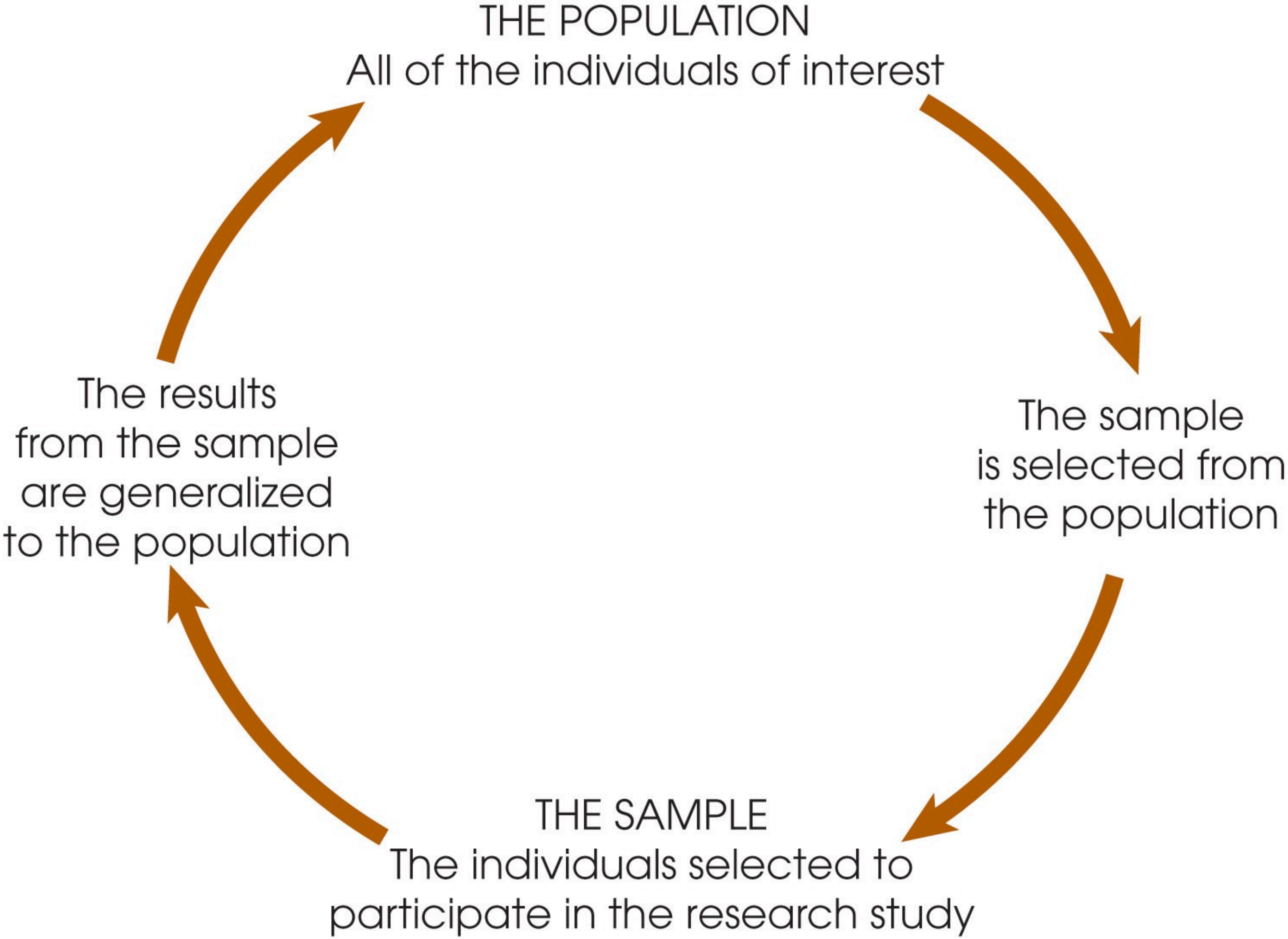
- A **variable** is a characteristic or condition that can change or take on different values.
- Most research begins with a general question about the relationship between two variables for a specific group of individuals.

Population

- The entire group of individuals is called the **population**.
- For example, a researcher may be interested in the relation between class size (variable 1) and academic performance (variable 2) for the population of third-grade children.

Sample

- Usually populations are so large that a researcher cannot examine the entire group. Therefore, a **sample** is selected to represent the population in a research study. The goal is to use the results obtained from the sample to help answer questions about the population.



THE POPULATION
All of the individuals of interest

The sample is selected from the population

THE SAMPLE
The individuals selected to participate in the research study

The results from the sample are generalized to the population

Descriptive Statistics

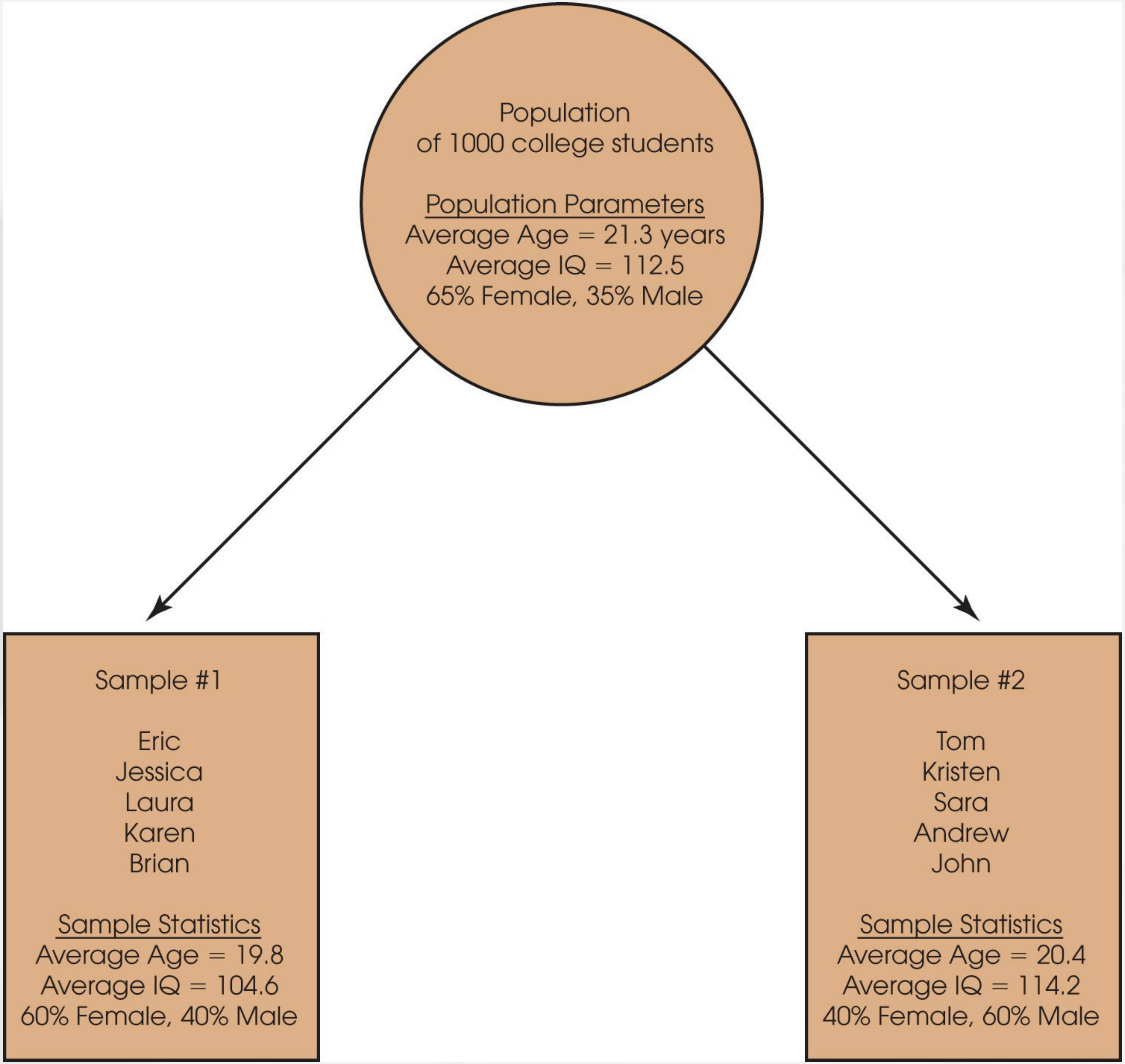
- **Descriptive statistics** are methods for organizing and summarizing data.
  - For example, tables or graphs are used to organize data, and descriptive values such as the average score are used to summarize data.
- A descriptive value for a population is called a **parameter** and a descriptive value for a sample is called a **statistic**.

Inferential Statistics

- **Inferential statistics** are methods for using sample data to make general conclusions (inferences) about populations.
- Because a sample is typically only a part of the whole population, sample data provide only limited information about the population. As a result, sample statistics are generally imperfect representatives of the corresponding population parameters.

Sampling Error

- The discrepancy between a sample statistic and its population parameter is called **sampling error**.
- Defining and measuring sampling error is a large part of inferential statistics.



Population
of 1000 college students

<u>Population Parameters</u>
Average Age = 21.3 years
Average IQ = 112.5
65% Female, 35% Male

Sample #1

Eric
Jessica
Laura
Karen
Brian

<u>Sample Statistics</u>
Average Age = 19.8
Average IQ = 104.6
60% Female, 40% Male

Sample #2

Tom
Kristen
Sara
Andrew
John

<u>Sample Statistics</u>
Average Age = 20.4
Average IQ = 114.2
40% Female, 60% Male

Data

- The measurements obtained in a research study are called the **data**.
- The goal of statistics is to help researchers organize and interpret the data.

Types of Variables

- Variables can be classified as discrete or continuous.
- **Discrete variables** (such as class size) consist of indivisible categories, and **continuous variables** (such as time or weight) are infinitely divisible into whatever units a researcher may choose. For example, time can be measured to the nearest minute, second, half-second, etc.

Real Limits

- To define the units for a continuous variable, a researcher must use **real limits** which are boundaries located exactly half-way between adjacent categories.

Measuring Variables

- To establish relationships between variables, researchers must observe the variables and record their observations. This requires that the variables be **measured**.
- The process of measuring a variable requires a set of categories called a **scale of measurement** and a process that classifies each individual into one category.

Four Types of Measurement Scales

1. A **nominal scale** is an unordered set of categories identified only by name. Nominal measurements only permit you to determine whether two individuals are the same or different.
2. An **ordinal scale** is an ordered set of categories. Ordinal measurements tell you the direction of difference between two individuals.

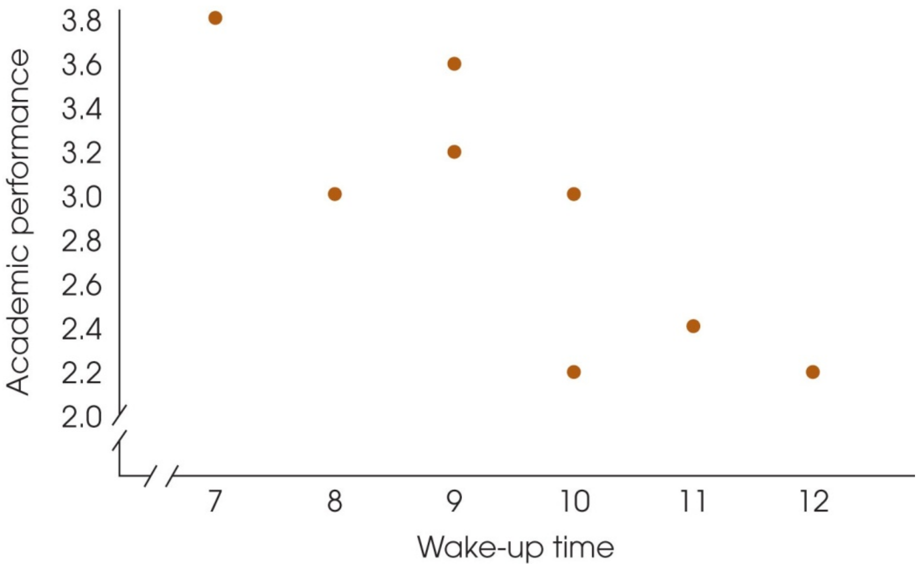
Four Types of Measurement Scales (cont’d.)

3. An **interval scale** is an ordered series of equal-sized categories. Interval measurements identify the direction and magnitude of a difference. The zero point is located arbitrarily on an interval scale.
4. A **ratio scale** is an interval scale where a value of zero indicates none of the variable. Ratio measurements identify the direction and magnitude of differences and allow ratio comparisons of measurements.

Correlational Studies

- The goal of a **correlational** study is to determine whether there is a relationship between two variables and to describe the relationship.
- A **correlational** study simply observes the two variables as they exist naturally.

Student	Wake-up Time	Academic Performance
A	11	2.4
B	9	3.6
C	9	3.2
D	12	2.2
E	7	3.8
F	10	2.2
G	10	3.0
H	8	3.0

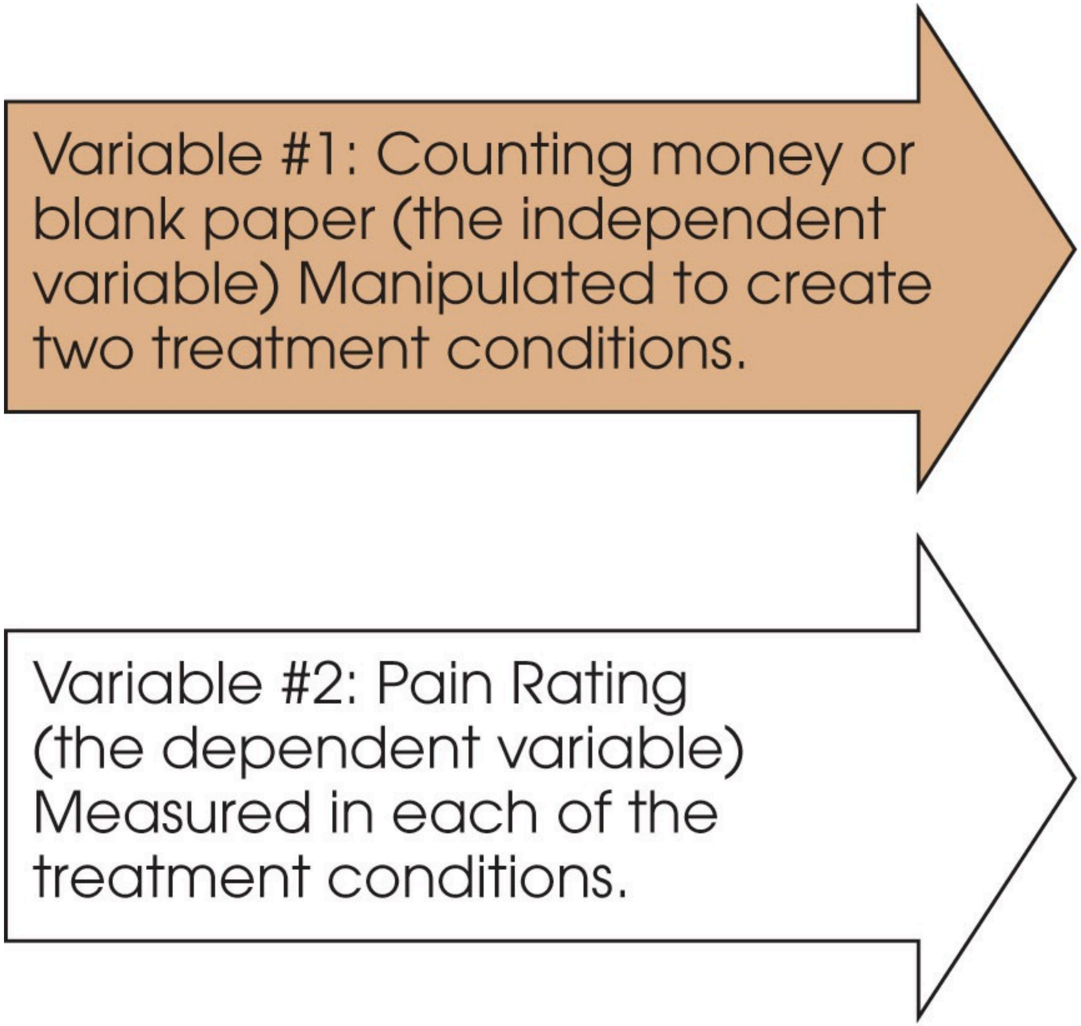


Experiments

- The goal of an **experiment** is to demonstrate a cause-and-effect relationship between two variables; that is, to show that changing the value of one variable causes changes to occur in a second variable.

Experiments (cont'd.)

- In an **experiment**, one variable is manipulated to create treatment conditions. A second variable is observed and measured to obtain scores for a group of individuals in each of the treatment conditions.
  - The measurements are then compared to see if there are differences between treatment conditions. All other variables are controlled to prevent them from influencing the results.
  - The manipulated variable is called the **independent variable** and the observed variable is the **dependent variable**.



Variable #1: Counting money or blank paper (the independent variable) Manipulated to create two treatment conditions.

Variable #2: Pain Rating (the dependent variable) Measured in each of the treatment conditions.

Money	Paper
7	8
4	10
5	8
6	9
6	8
8	10
6	7
5	8
5	8
6	7

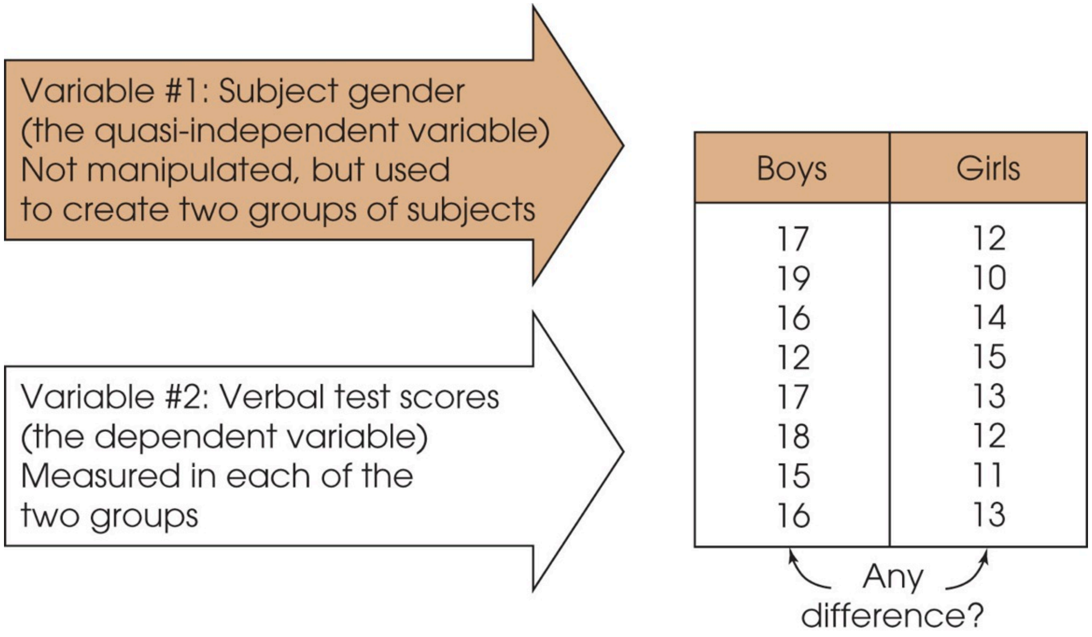
Compare groups of scores

Other Types of Studies

- Other types of research studies, know as non-**experimental** or **quasi-experimental**, are similar to experiments because they also compare groups of scores.
- These studies do not use a manipulated variable to differentiate the groups. Instead, the variable that differentiates the groups is usually a pre-existing participant variable (such as male/female) or a time variable (such as before/after).

Other Types of Studies (cont'd.)

- Because these studies do not use the manipulation and control of true experiments, they cannot demonstrate cause and effect relationships. As a result, they are similar to correlational research because they simply demonstrate and describe relationships.

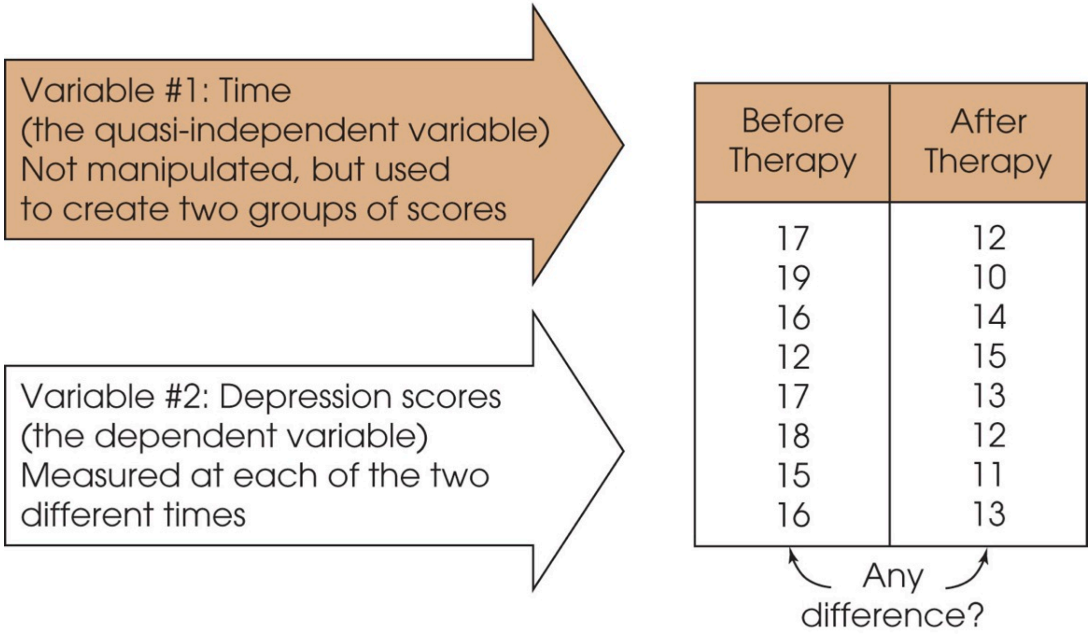


Variable #1: Subject gender
(the quasi-independent variable)
Not manipulated, but used
to create two groups of subjects

Variable #2: Verbal test scores
(the dependent variable)
Measured in each of the
two groups

Boys	Girls
17	12
19	10
16	14
12	15
17	13
18	12
15	11
16	13

Any difference?



Variable #1: Time
(the quasi-independent variable)
Not manipulated, but used
to create two groups of scores

Variable #2: Depression scores
(the dependent variable)
Measured at each of the two
different times

Before Therapy	After Therapy
17	12
19	10
16	14
12	15
17	13
18	12
15	11
16	13

Any difference?

Statistical Notation

- The individual measurements or scores obtained for a research participant will be identified by the letter *X* (or *X* and *Y* if there are multiple scores for each individual).
- The number of scores in a data set will be identified by *N* for a population or n for a sample.
- Summing a set of values is a common operation in statistics and has its own notation. The Greek letter sigma, Σ, will be used to stand for "the sum of." For example, Σ*X* identifies the sum of the scores.

Order of Operations

1. All calculations within parentheses are done first.
2. Squaring or raising to other exponents is done second.
3. Multiplying, and dividing are done third, and should be completed in order from left to right.
4. Summation with the Σ notation is done next.
5. Any additional adding and subtracting is done last and should be completed in order from left to right.

Chapter 2: Frequency Distributions

Frequency Distributions

- After collecting data, the first task for a researcher is to organize and simplify the data so that it is possible to get a general overview of the results.
- This is the goal of descriptive statistical techniques.
- One method for simplifying and organizing data is to construct a **frequency distribution**.

Frequency Distributions (cont'd.)

- A **frequency distribution** is an organized tabulation showing exactly how many individuals are located in each category on the scale of measurement. A frequency distribution presents an organized picture of the entire set of scores, and it shows where each individual is located relative to others in the distribution.

Frequency Distribution Tables

- A frequency distribution table consists of at least two columns - one listing categories on the scale of measurement (*X*) and another for frequency (*f*).
  - In the *X* column, values are listed from the highest to lowest, without skipping any.
  - For the frequency column, tallies are determined for each value (how often each *X* value occurs in the data set). These tallies are the frequencies for each *X* value.
  - The sum of the frequencies should equal *N*.

Frequency Distribution Tables (cont'd.)

- A third column can be used for the proportion (p) for each category: $p = f/N$. The sum of the p column should equal 1.00.
- A fourth column can display the percentage of the distribution corresponding to each *X* value. The percentage is found by multiplying p by 100. The sum of the percentage column is 100%.

Regular Frequency Distribution

- When a frequency distribution table lists all of the individual categories (*X* values) it is called a **regular frequency distribution**.

Grouped Frequency Distribution

- Sometimes, however, a set of scores covers a wide range of values. In these situations, a list of all the *X* values would be quite long - too long to be a “simple” presentation of the data.
- To remedy this situation, a **grouped frequency distribution** table is used.

Grouped Frequency Distribution (cont'd.)

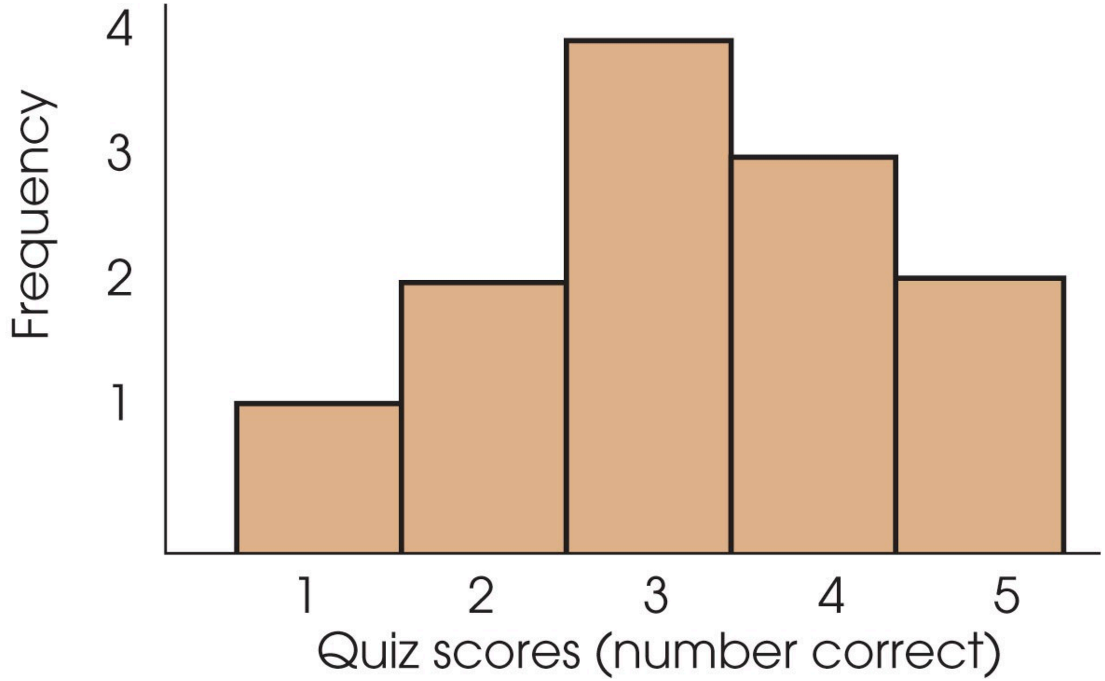
- In a grouped table, the X column lists groups of scores, called **class intervals**, rather than individual values.
- These intervals all have the same width, usually a simple number such as 2, 5, 10, and so on.
- Each interval begins with a value that is a multiple of the interval width. The interval width is selected so that the table will have approximately ten intervals.

Frequency Distribution Graphs

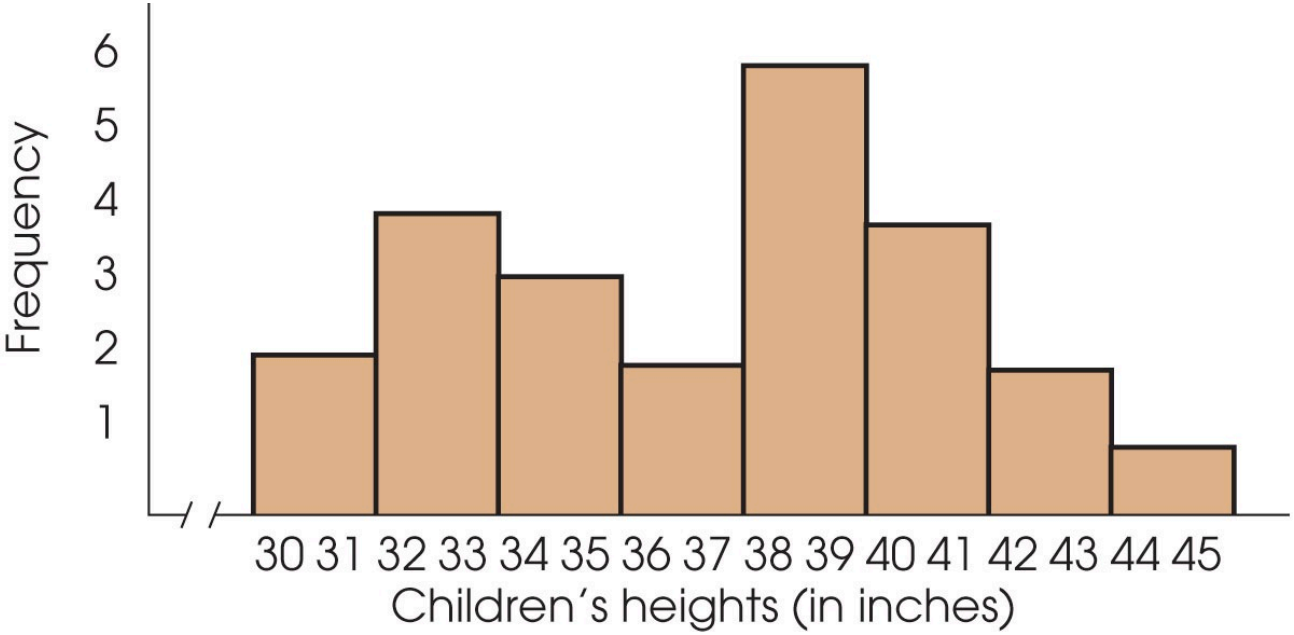
- In a **frequency distribution graph**, the score categories (*X* values) are listed on the *X* axis and the frequencies are listed on the *Y* axis.
- When the score categories consist of numerical scores from an interval or ratio scale, the graph should be either a histogram or a polygon.

Histograms

- In a **histogram**, a bar is centered above each score (or class interval) so that the height of the bar corresponds to the frequency and the width extends to the real limits, so that adjacent bars touch.



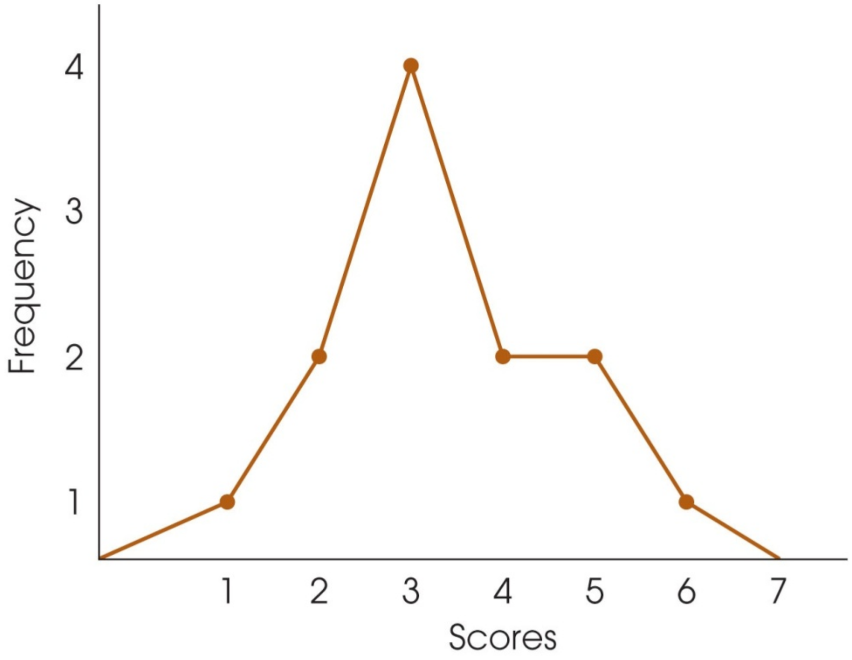
X	*f*
5	2
4	3
3	4
2	2
1	1



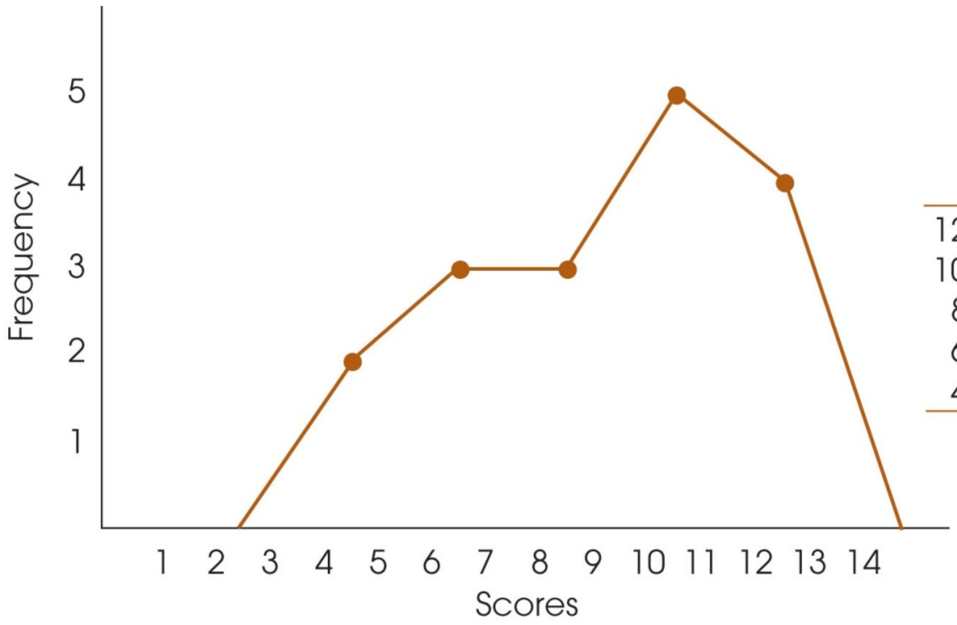
X	*f*
44–45	1
42–43	2
40–41	4
38–39	6
36–37	2
34–35	3
32–33	4
30–31	2

Polygons

- In a **polygon**, a dot is centered above each score so that the height of the dot corresponds to the frequency. The dots are then connected by straight lines. An additional line is drawn at each end to bring the graph back to a zero frequency.



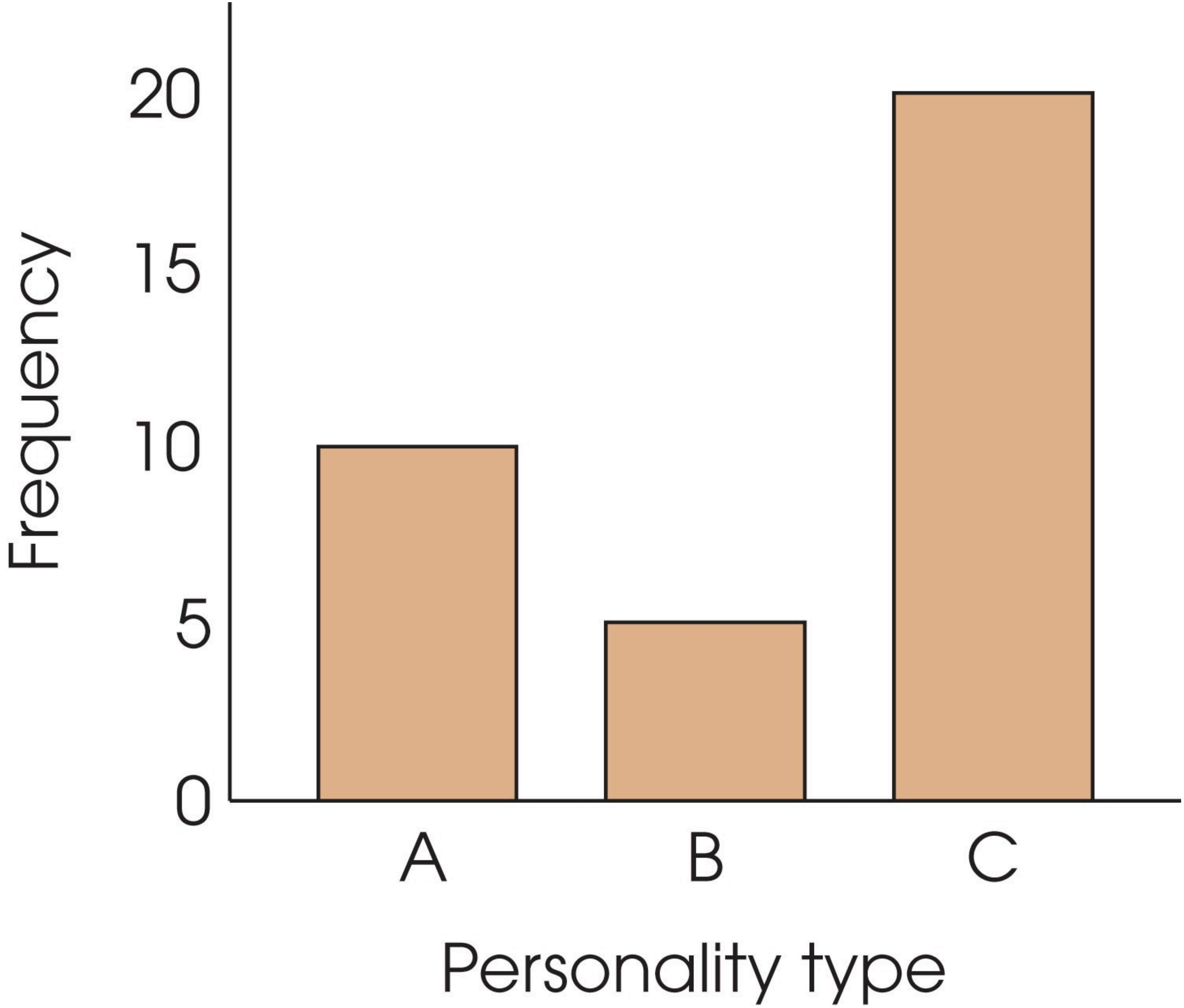
X	*f*
6	1
5	2
4	2
3	4
2	2
1	1



X	*f*
12–13	4
10–11	5
8–9	3
6–7	3
4–5	2

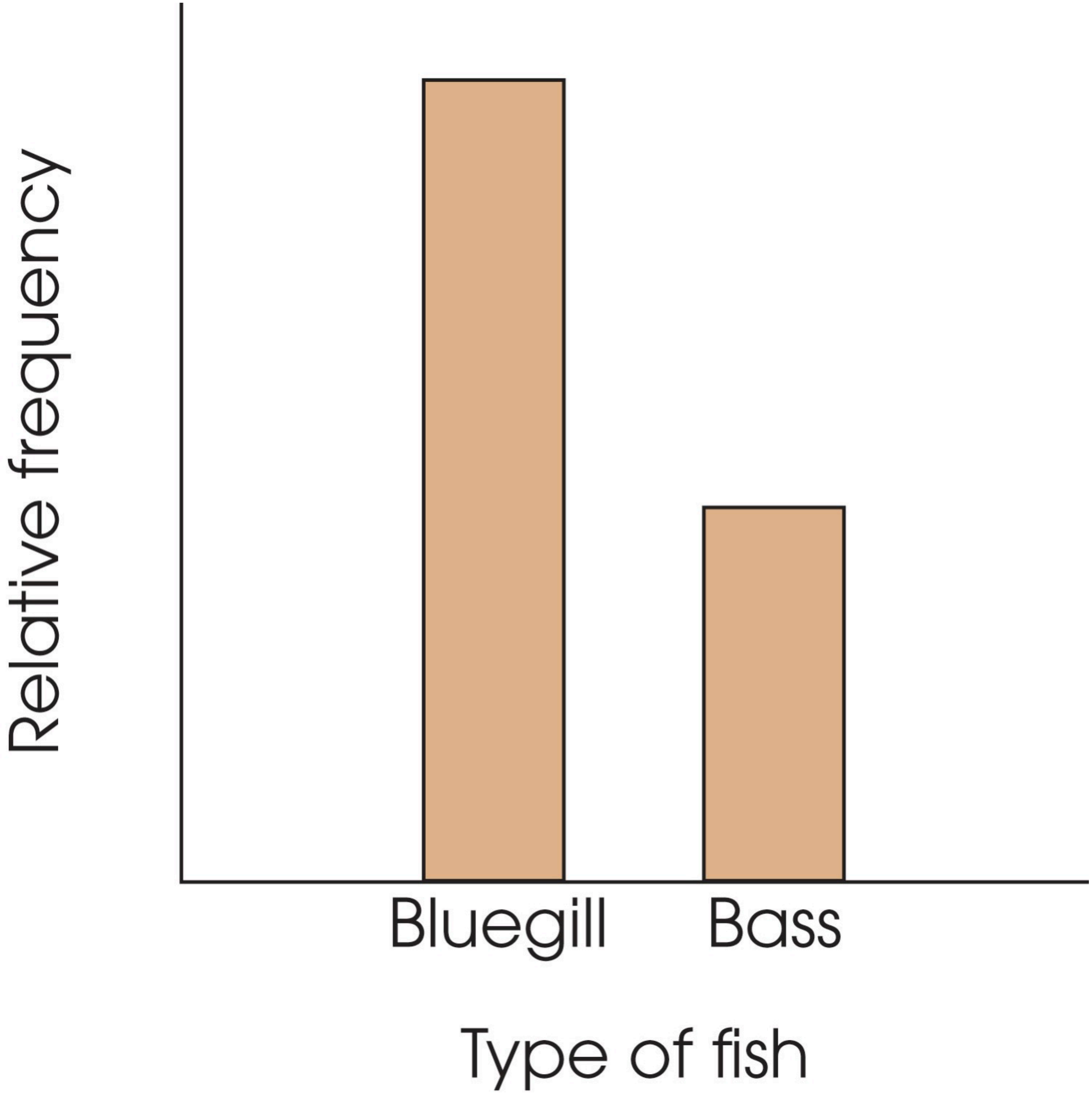
Bar Graphs

- When the score categories (X values) are measurements from a nominal or an ordinal scale, the graph should be a bar graph.
- A **bar graph** is just like a histogram except that gaps or spaces are left between adjacent bars.


Frequency
20
15
10
5
0
A
B
C
Personality type

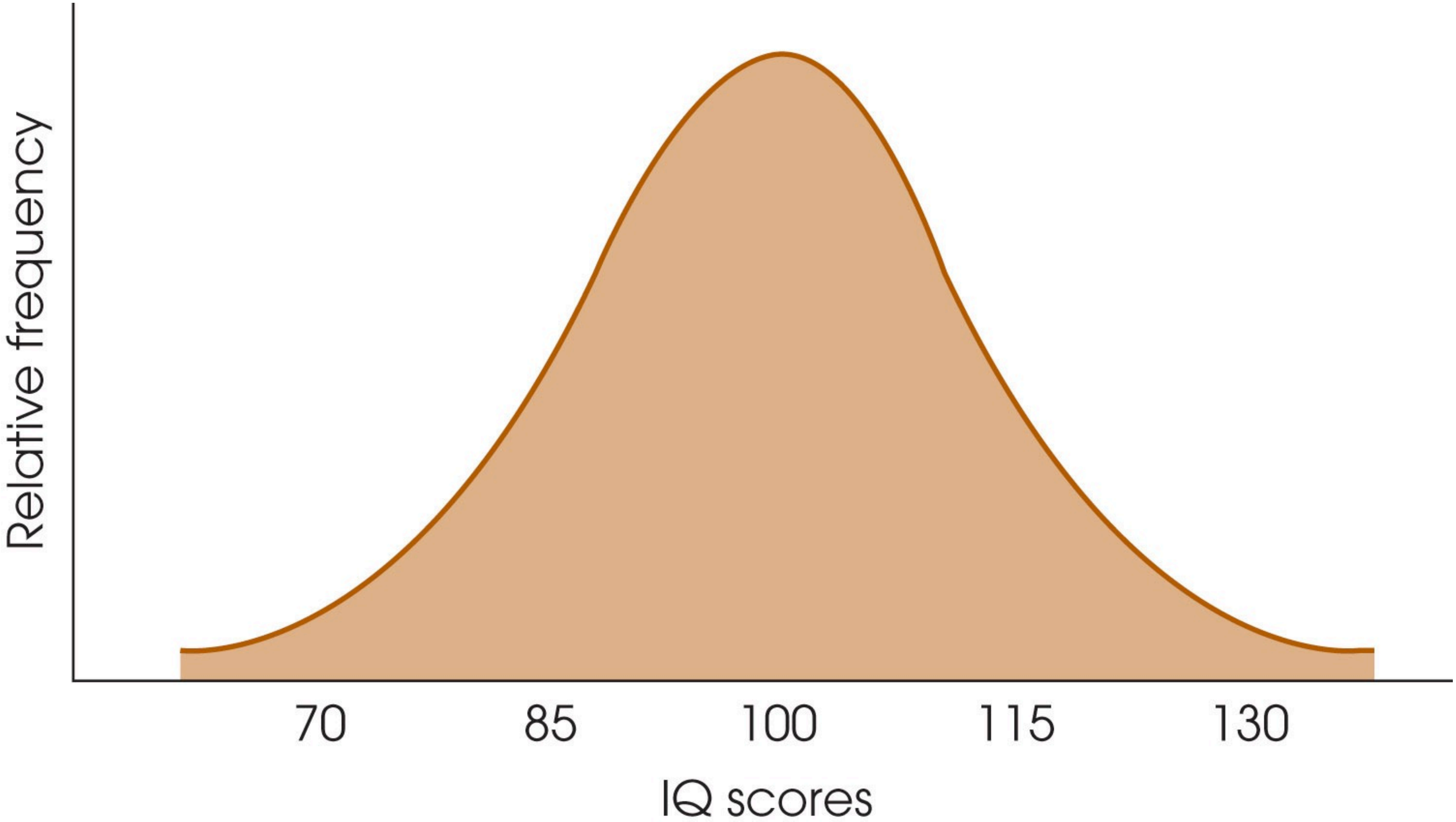
Relative Frequency

- Many populations are so large that it is impossible to know the exact number of individuals (frequency) for any specific category.
- In these situations, population distributions can be shown using **relative frequency** instead of the absolute number of individuals for each category.


Relative frequency
Bluegill
Bass
Type of fish

Smooth Curve

- If the scores in the population are measured on an interval or ratio scale, it is customary to present the distribution as a **smooth curve** rather than a jagged histogram or polygon.
- The smooth curve emphasizes the fact that the distribution is not showing the exact frequency for each category.


Relative frequency
70
85
100
115
130
IQ scores

Frequency Distribution Graphs

- Frequency distribution graphs are useful because they show the entire set of scores.
- At a glance, you can determine the highest score, the lowest score, and where the scores are centered.
- The graph also shows whether the scores are clustered together or scattered over a wide range.

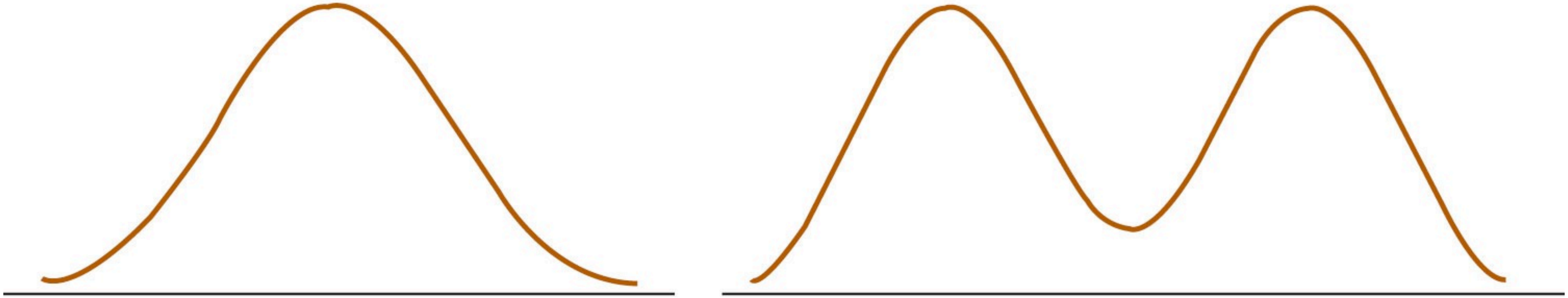
Shape

- A graph shows the **shape** of the distribution.
- A distribution is **symmetrical** if the left side of the graph is (roughly) a mirror image of the right side.
- One example of a symmetrical distribution is the bell-shaped normal distribution.
- On the other hand, distributions are **skewed** when scores pile up on one side of the distribution, leaving a "tail" of a few extreme values on the other side.

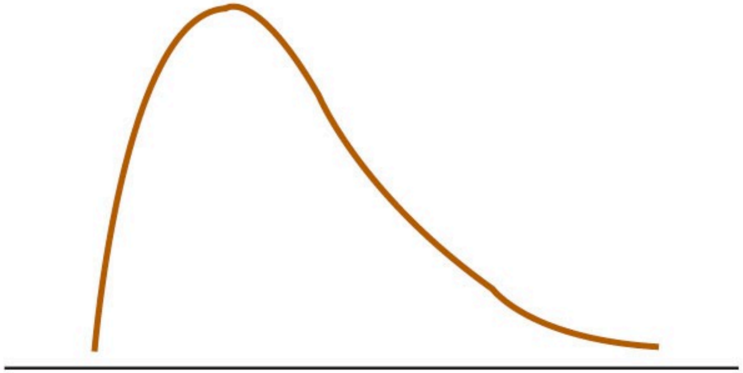
Positively and Negatively Skewed Distributions

- In a **positively skewed** distribution, the scores tend to pile up on the left side of the distribution with the tail tapering off to the right.
- In a **negatively skewed** distribution, the scores tend to pile up on the right side and the tail points to the left.

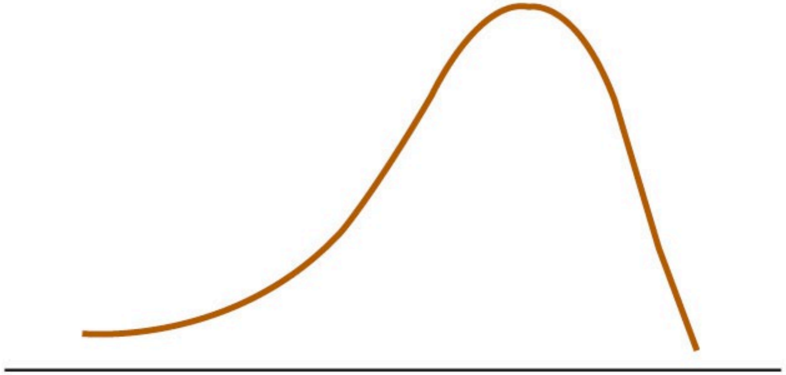
Symmetrical distributions



Skewed distributions



Positive skew



Negative skew

Percentiles, Percentile Ranks, and Interpolation

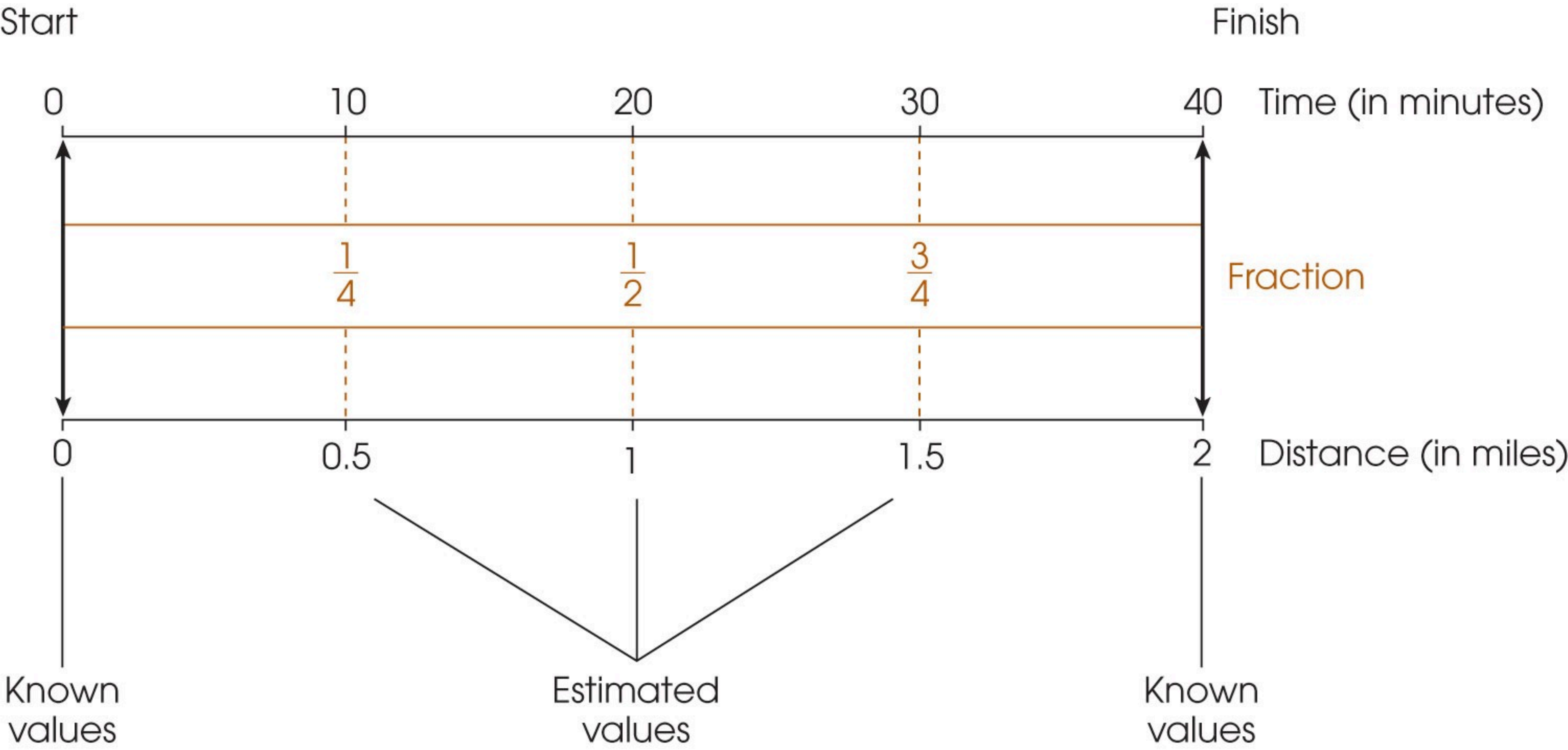
- The relative location of individual scores within a distribution can be described by percentiles and percentile ranks.
- The **percentile rank** for a particular *X* value is the percentage of individuals with scores equal to or less than that *X* value.
- When an *X* value is described by its rank, it is called a **percentile**.

Percentiles, Percentile Ranks, and Interpolation (cont'd.)

- To find percentiles and percentile ranks, two new columns are placed in the frequency distribution table: One is for cumulative frequency (*cf*) and the other is for cumulative percentage (c%).
- Each cumulative percentage identifies the percentile rank for the upper real limit of the corresponding score or class interval.

Interpolation

- When scores or percentages do not correspond to upper real limits or cumulative percentages, you must use interpolation to determine the corresponding ranks and percentiles.
- **Interpolation** is a mathematical process based on the assumption that the scores and the percentages change in a regular, linear fashion as you move through an interval from one end to the other.


Start
Finish
0
10
20
30
40
Time (in minutes)
1/4
1/2
3/4
Fraction
0
0.5
1
1.5
2
Distance (in miles)
Known values
Estimated values
Known values

Stem-and-Leaf Displays

- A **stem-and-leaf** display provides an efficient method for obtaining and displaying a frequency distribution.
  - Each score is divided into a **stem** consisting of the first digit or digits, and a **leaf** consisting of the final digit.
  - Then, go through the list of scores, one at a time, and write the leaf for each score beside its stem.

Stem-and-Leaf Displays (cont’d.)

- The resulting display provides an organized picture of the entire distribution. The number of leaves beside each stem corresponds to the frequency, and the individual leaves identify the individual scores.

TABLE 2.3

A set of $N = 24$ scores presented as raw data and organized in a stem and leaf display.

Data			Stem and Leaf Display	
83	82	63	3	23
62	93	78	4	26
71	68	33	5	6279
76	52	97	6	283
85	42	46	7	1643846
32	57	59	8	3521
56	73	74	9	37
74	81	76		