# Lecture 01: Introduction to Statistics

魏坤琳

# Why should students in psychology learn statistics?

R. A. Fisher
(1890–1962)

- Psychology is a science
  (APS: Association for Psychological Science)
- Statistics provides objective and systematic methods for extracting information from the results of scientific research
- Statistics is also useful for everyday life

# Example：医疗方法有效?

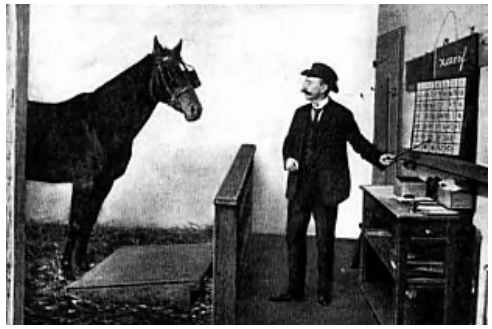
| 200人接受治疗，病情改善 | 75人接受治疗，病情没有改善 |
| --- | --- |

200/(200+75) = 0.73

| 50人没有接受治疗，病情改善 | 15人没有接受治疗，病情没有改善 |
| --- | --- |

50/(50+15) = 0.77


神医



神奇的马 Clever Hans

科学史上最伟大的突破之一：假设检验

“True logic of this world is in the calculus of probabilities.”

— J. C. Maxwell

"The best thing about being a statistician is that you get to play in everyone's backyard."

— John Tukey

# Statistics in the era of big data

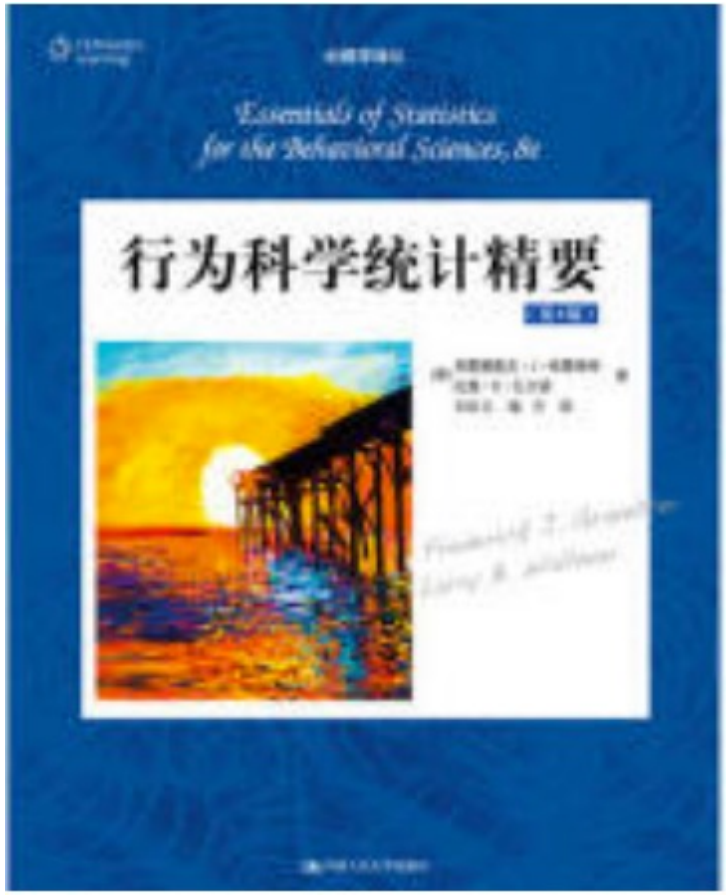
"I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s? The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it."

*by Hal Varian, Google's Chief Economist, 2009*

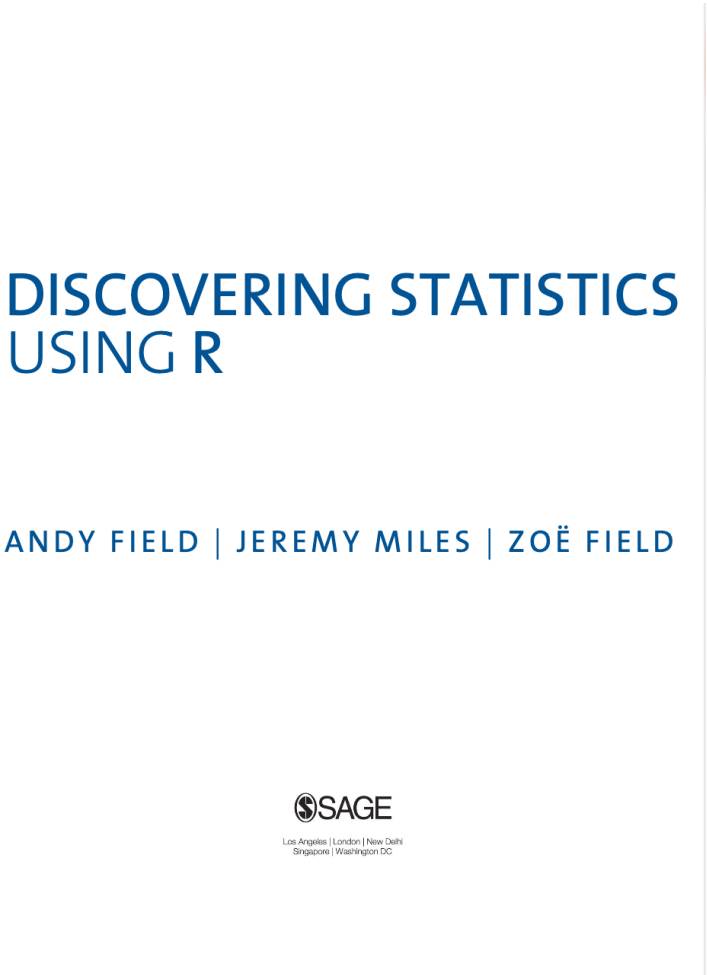
# Syllabus

Where to find study materials: our wechat group or course.pku.edu.cn

# Textbooks

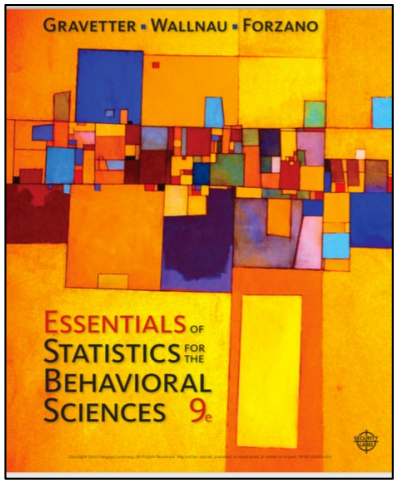



Gravetter & Wallnau, 2016 (8th edition)

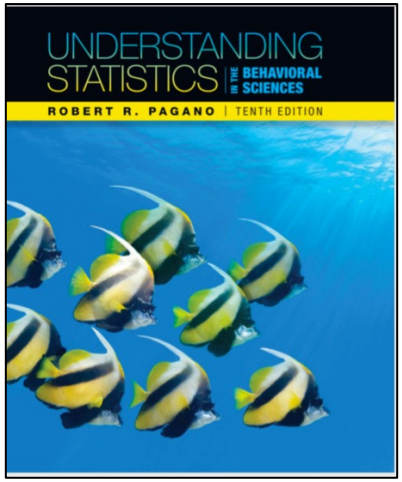


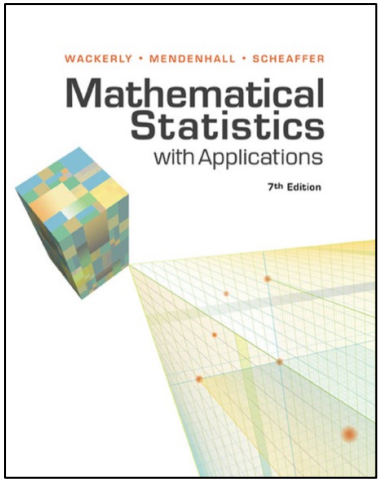

Field, Miles & Filed, 2016

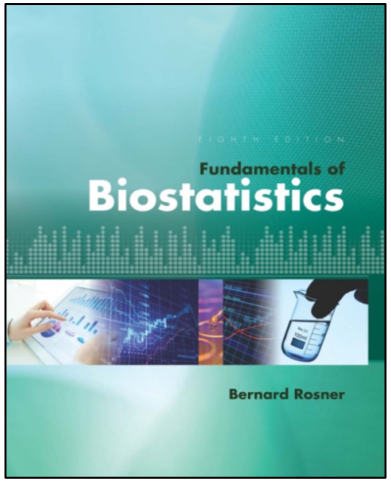
# Readings




**The 9th edition** of "Essentials of statistics for the behavioral sciences" includes a list of *learning objectives* for each section of each chapter.

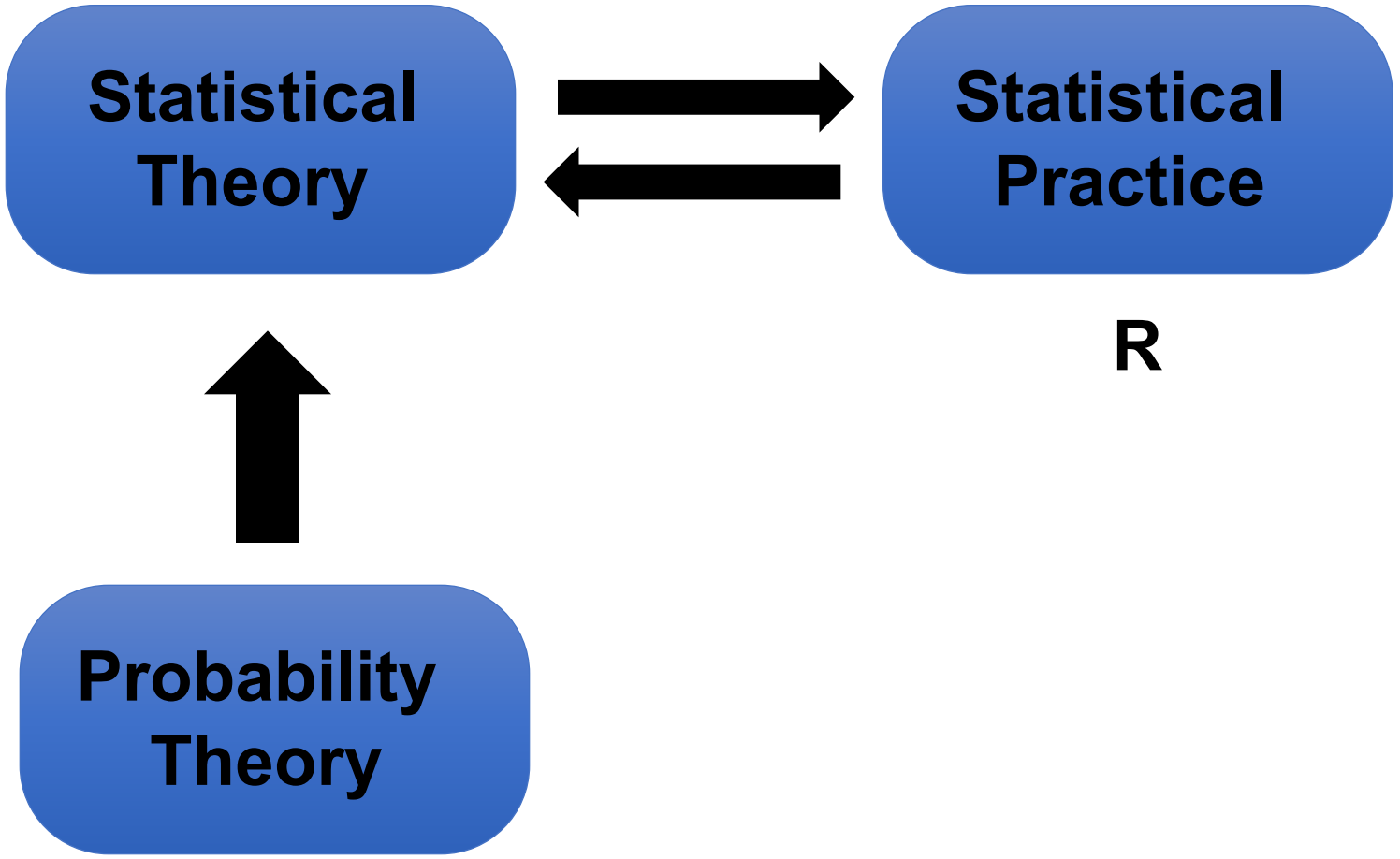


# Useful components of the books

- Learning objectives
- Summary
- Key terms
- Problem sets


Statistical Theory
Statistical Practice
R
Probability Theory

# Schedule (tentative)

| Week | Dates | Lecture | Lab |
|---|---|---|---|
| 1 | 09/10 | Introduction to statistics | |
| 2 | 09/17 | Frequency distributions and R | |
| 3 | 09/24 | Central Tendency, variability and z-Scores | |
| 4 | 10/01 | Holiday (no class, National Day) | |
| 5 | 10/08 | Holiday (no class, moon festival) | |
| 6 | 10/15 | Random sampling and probability | ✓ |
| 7 | 10/22 | Continuous and discrete distributions | ✓ |
| 8 | 10/29 | Intro to hypothesis testing | ✓ |
| 9 | 11/05 | Intro to t statistics | ✓ |
| 10 | 11/12 | t test for two independent samples | ✓ |
| 11 | 11/19 | t test for two related samples | ✓ |
| 12 | 11/26 | One-way ANOVA | ✓ |
| 13 | 12/03 | Repeated-measures ANOVA | ✓ |
| 14 | 12/10 | Two-way ANOVA | ✓ |
| 15 | 12/17 | Mixed-design ANOVA | ✓ |
| 16 | 12/24 | Correlation and regression | ✓ |
| 17 | 12/31 | Review | |

# Evaluation

- Midterm (25–35%) + Final (30–50%) + Homework & Lab Assignments (25–35%)
- Slides in English, homework and exams in Chinese/English.
- Homework: available online after class, due next class
- Cheat-sheet policy for both the midterm and final exam: close book, but a one-page note (A-4, single side) is allowed

Let’s begin!

# Two general purposes of statistics

- To organize and summarize the information in the results of scientific studies
  - for the researchers themselves
  - for communication to others

- To answer specific questions based on the results

# Populations and Samples

- A **population** is the set of all the individuals of interest in a particular study.
- A **sample** is a set of individuals selected from a population, usually intended to represent the population in a research study.

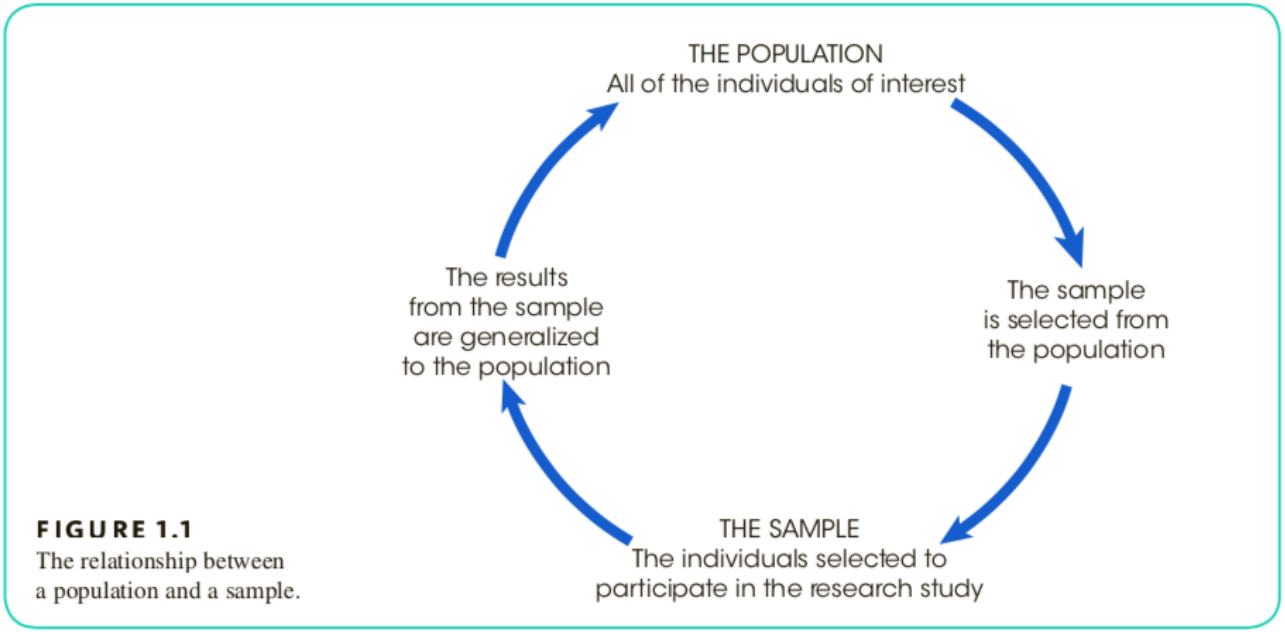



**FIGURE 1.1**
The relationship between a population and a sample.

# Variables and Data

- A **variable** is a characteristic or condition that changes or has different values for different individuals. E.g., hormone level and aggressiveness.
- **Data** (plural) are measurements or observations. A **data set** is a collection of measurements or observations. A **datum** (singular) is a single measurement or observation and is commonly called a score or raw score. E.g., a sample of 100 male PKU students who have been measured for hormone level and aggressiveness.

Sample/population of scores

# Parameters and Statistics

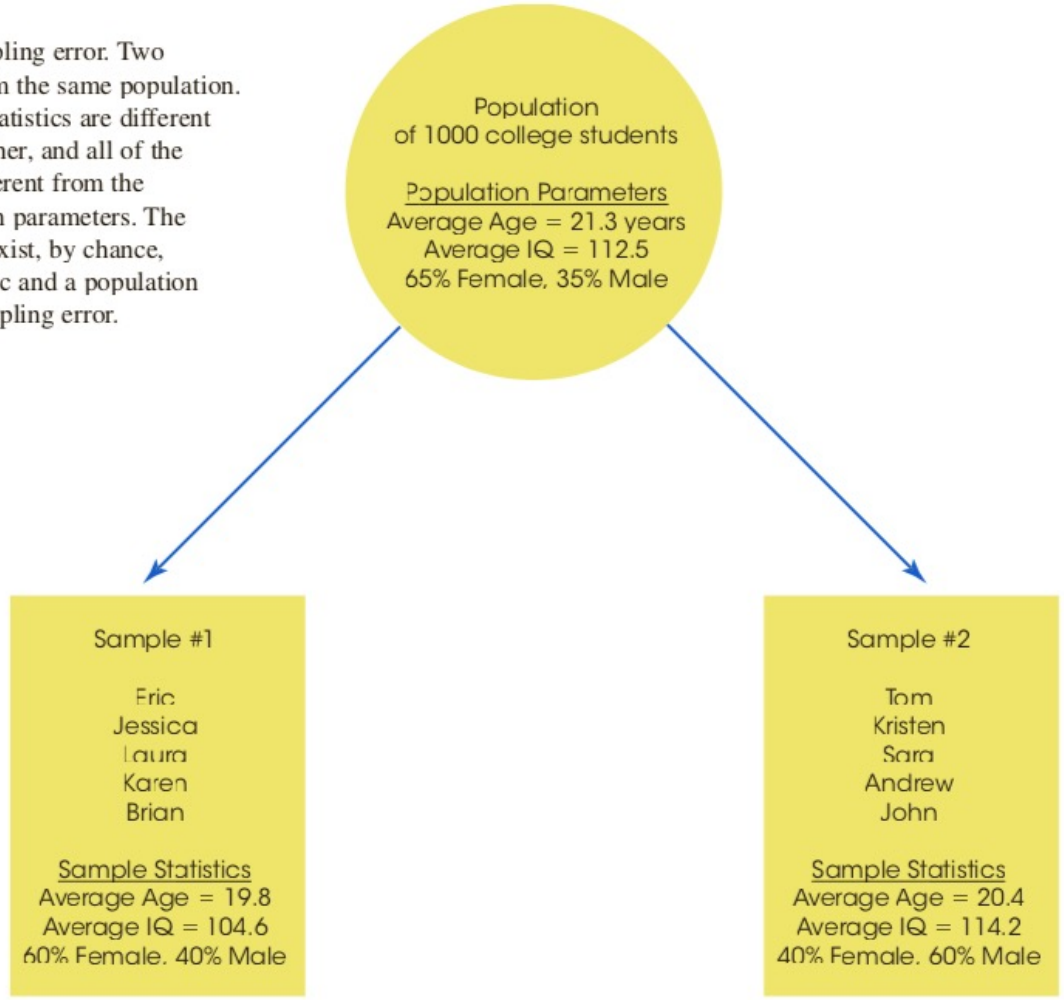
- A **parameter** is a value—usually a numerical value—that describes a population. A parameter is usually derived from measurements of the individuals in the population. E.g. the mean and the variance of a population.
- A **statistic** is a value—usually a numerical value—that describes a sample. A statistic is usually derived from measurements of the individuals in the sample. E.g., the average and the variance of a sample.

# Descriptive and Inferential Statistical Methods

- **Descriptive statistics** are statistical procedures used to summarize, organize, and simplify data.
- **Inferential statistics** consist of techniques that allow us to study samples and then make generalizations about the populations from which they were selected.

- **Sampling error** is the naturally occurring discrepancy, or error, that exists between a sample statistic and the corresponding population parameter.
  (a fundamental problem for inferential statistics)

**FIGURE 1.2**
A demonstration of sampling error. Two samples are selected from the same population. Notice that the sample statistics are different from one sample to another, and all of the sample statistics are different from the corresponding population parameters. The natural differences that exist, by chance, between a sample statistic and a population parameter are called sampling error.




Margin of error?
Confidence interval?

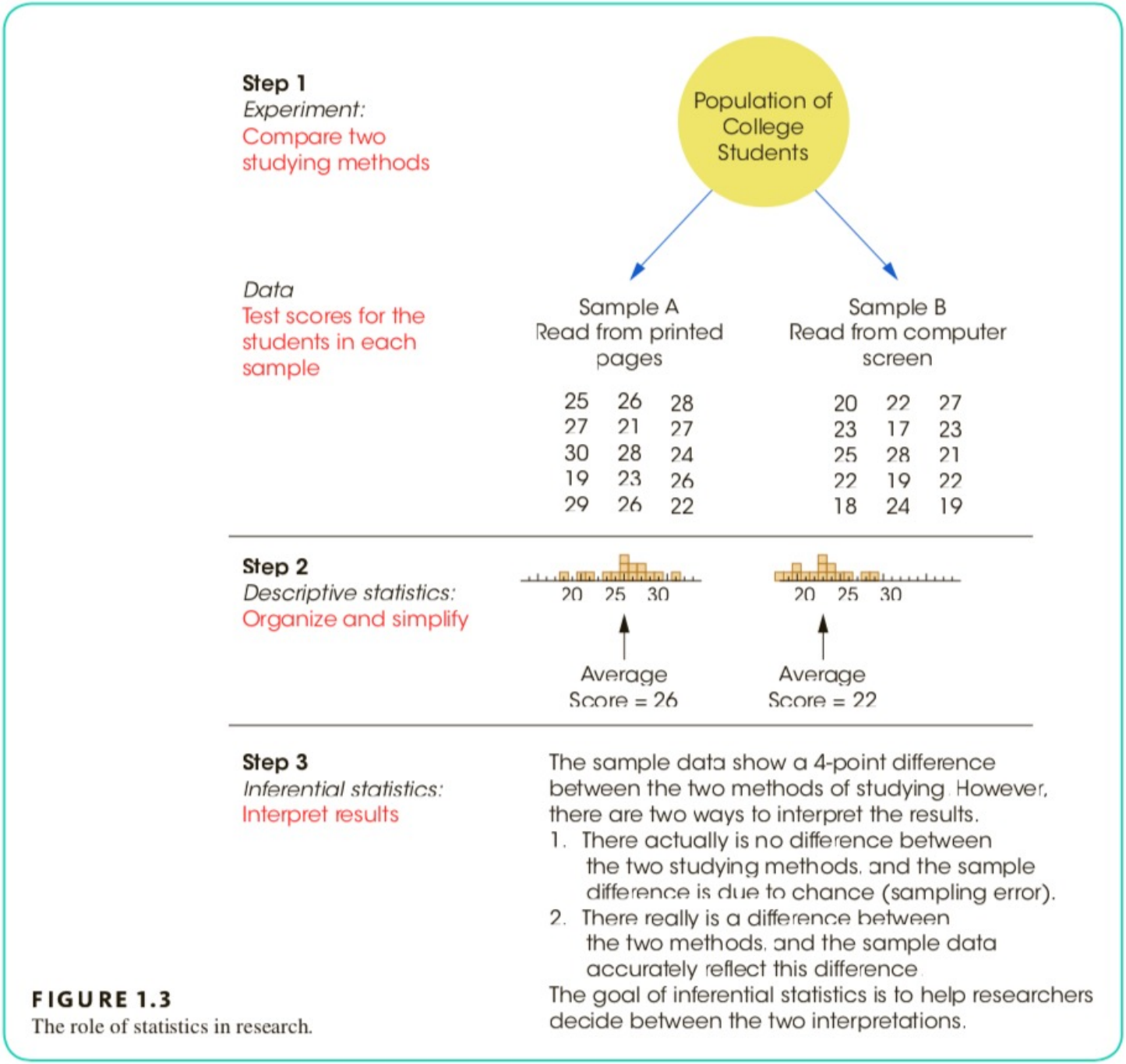



**FIGURE 1.3**
The role of statistics in research.

# Constructs and Operational Definitions

- **Constructs** are internal attributes or characteristics that cannot be directly observed but are useful for describing and explaining behavior. e.g. intelligence, anxiety, happiness, consciousness, self
- An **operational definition** (1) identifies a measurement procedure (a set of operations) for measuring an external behavior and (2) uses the resulting measurements as a definition and a measurement of a hypothetical construct.

# Discrete and Continuous Variables

- A **discrete variable** consists of separate, indivisible categories. No values can exist between two neighboring categories.
  e.g. number of people, academic major (psychology, biology, math, etc.)
- For a **continuous variable**, there are an infinite number of possible values that fall between any two observed values. A continuous variable is divisible into an infinite number of fractional parts.
  e.g. time, space, height, weight, blood pressure

- **Real limits** are the boundaries of intervals for scores that are represented on a *continuous* number line. The real limit separating two adjacent scores is located exactly halfway between the scores. Each score has two real limits. The **upper real limit** is at the top of the interval, and the **lower real limit** is at the bottom.

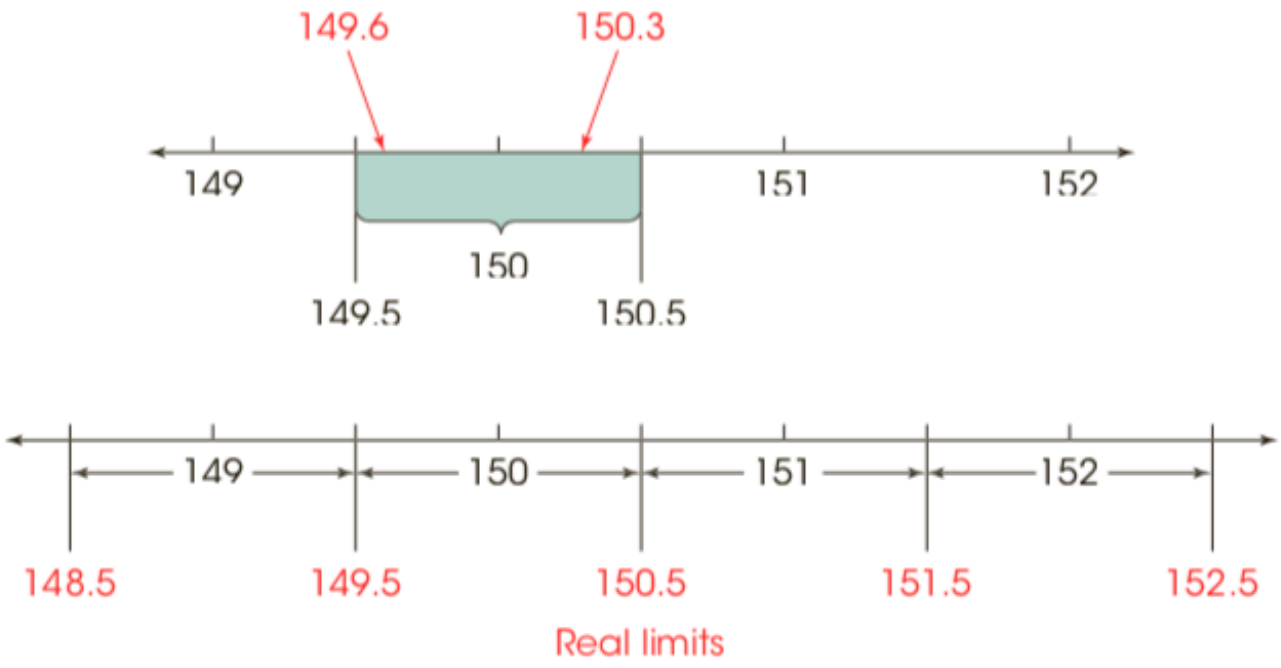



**FIGURE 1.4**
When measuring weight to the nearest whole pound, 149.6 and 150.3 are assigned the value of 150 (top). Any value in the interval between 149.5 and 150.5 is given the value of 150.

# Scales of Measurement (four levels)

- A **nominal scale** consists of a set of categories that have different names. Measurements on a nominal scale label and categorize observations, but do not make any quantitative distinctions between observations.

  e.g. gender, academic major, room number

# Scales of Measurement (four levels)

- An **ordinal scale** consists of a set of categories that are organized in an ordered sequence. Measurements on an ordinal scale rank observations in terms of size or magnitude.

  e.g. T-shirt sizes (small, medium, large), food preferences (rare, medium rare, medium etc).

# Scales of Measurement (four levels)

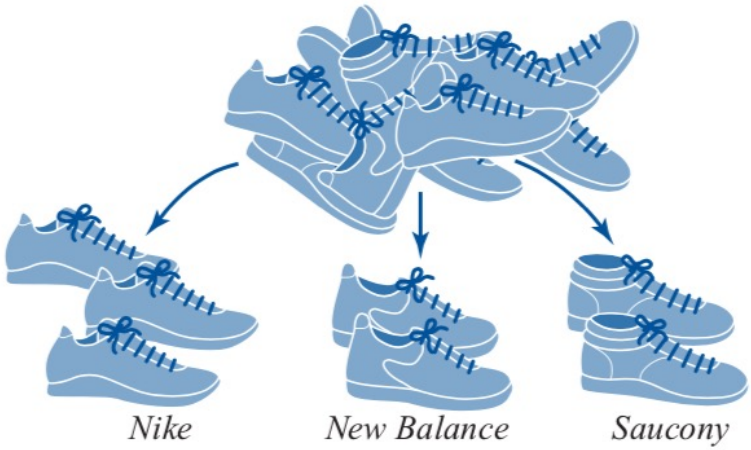
- An **interval scale** consists of ordered categories that are all intervals of exactly the same size. Equal differences between numbers on scale reflect equal differences in magnitude. However, the zero point on an interval scale is arbitrary and does not indicate a zero amount of the variable being measured (the absence of the trait to be measured).

  e.g. Fahrenheit or Celsius temperature scale; year (2000), IQ scores, SAT scores, Likert scale, Z score, day time (8 AM)

# Scales of Measurement (four levels)

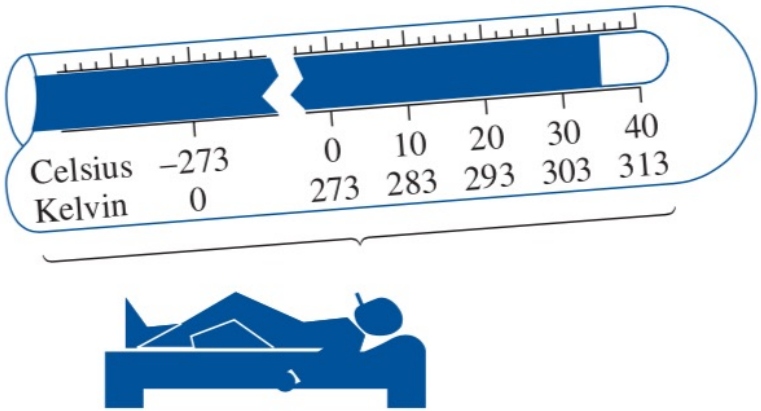
- A **ratio scale** is an interval scale with the additional feature that a score of zero indicates none of the variable being measured. With a ratio scale, ratios of numbers do reflect ratios of magnitude.

  e.g. Kelvin temperature scale, response time, accuracy




**Nominal**
Units of scale are categories. Objects are measured by determining the category to which they belong. There is no magnitude relationship between the categories.

**Ordinal**
Possesses the property of magnitude. Can rank-order the objects according to whether they possess more, less, or the same amount of the variable being measured. Thus, can determine whether $A > B$, $A = B$, or $A < B$.




**Interval and Ratio**
*Interval:* Possesses the properties of magnitude and equal intervals between adjacent units. Can do same determinations as with an ordinal scale, plus can determine whether $A - B = C - D$, $A - B > C - D$, or $A - B < C - D$.
*Ratio:* Possesses the properties of magnitude, equal interval between adjacent units, and an absolute zero point. Can do all the mathematical operations usually associated with numbers, including ratios.

**figure 2.1** Scales of measurement and their characteristics.

# An example: pain ratings

1）Visual Analog Scale (VAS)

- A 10-cm line, one end corresponding to “no pain”, the other end corresponding to “the worst pain”. Mark your perceived pain now on the line. Then we measure it using a ruler.

2）Numeric Rating Scale (NRS)

- Rate your pain from 0 to 10 integer numbers, 0 means “no pain”, 10 means “the worst pain you can imagine”.

3） Verbal rating scale (VRS)

- Asks participants to choose a word or phrase from a list that best describes their pain intensity. A typical version includes: No pain; Mild pain; Moderate pain; Severe pain; Very severe (or extreme) pain.

# Three Data Structures, Research Methods, and Statistics

Different data structures resulted from different research methods and require different statistical techniques

## Data Structure 1. One Group with One or More Separate Variables Measured for Each Individual: Descriptive Research

**Descriptive research** or the **descriptive research strategy** involves measuring one or more separate variables for each individual with the intent of simply describing the individual variables.

Three separate variables measured for each individual in a group of eight students.

| | Weekly Number of Student Fast-Food Meals | Number of Hours Sleeping Each Day | Number of Hours Studying Each Day |
|---|---|---|---|
| A | 0 | 9 | 3 |
| B | 4 | 7 | 2 |
| C | 2 | 8 | 4 |
| D | 1 | 10 | 3 |
| E | 0 | 11 | 2 |
| F | 0 | 7 | 4 |
| G | 5 | 7 | 3 |
| H | 3 | 8 | 2 |

# Data Structure 2. One Group with Two Variables Measured for Each Individual: The Correlational Method

In the **correlational method**, two different variables are observed to determine whether there is a relationship between them.

Limitation of the correlational method: **Correlation does not necessarily mean cause-and-effect**.

(a)

| Student | Facebook Time | Academic Performance |
|---|---|---|
| A | 4 | 2.4 |
| B | 2 | 3.6 |
| C | 2 | 3.2 |
| D | 5 | 2.2 |
| E | 0 | 3.8 |
| F | 3 | 2.2 |
| G | 3 | 3.0 |
| H | 1 | 3.0 |

(b)

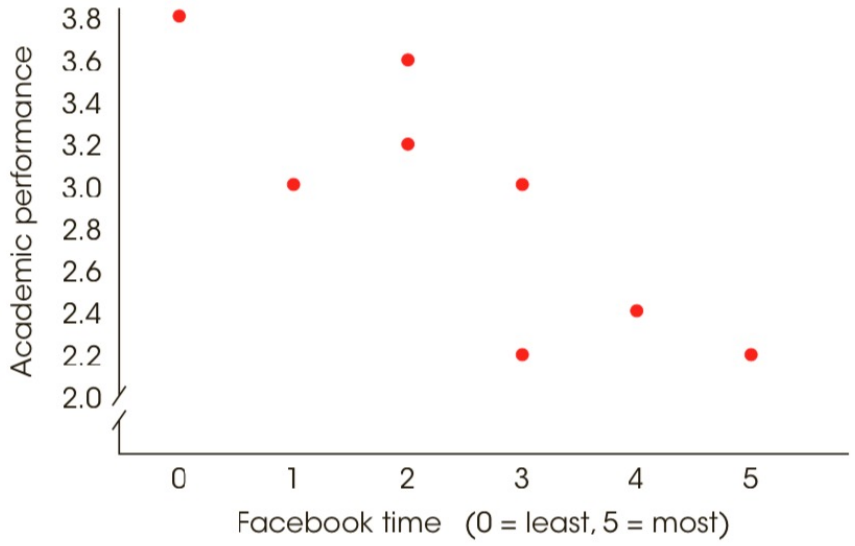



**FIGURE 1.5**

One of two data structures for studies evaluating the relationship between variables. Note that there are two separate measurements for each individual (Facebook time and academic performance). The same scores are shown in a table (a) and a graph (b).

## Data Structure 3. Comparing Two (or More) Groups of Scores: Experimental and Nonexperimental Methods

Example of experimental methods

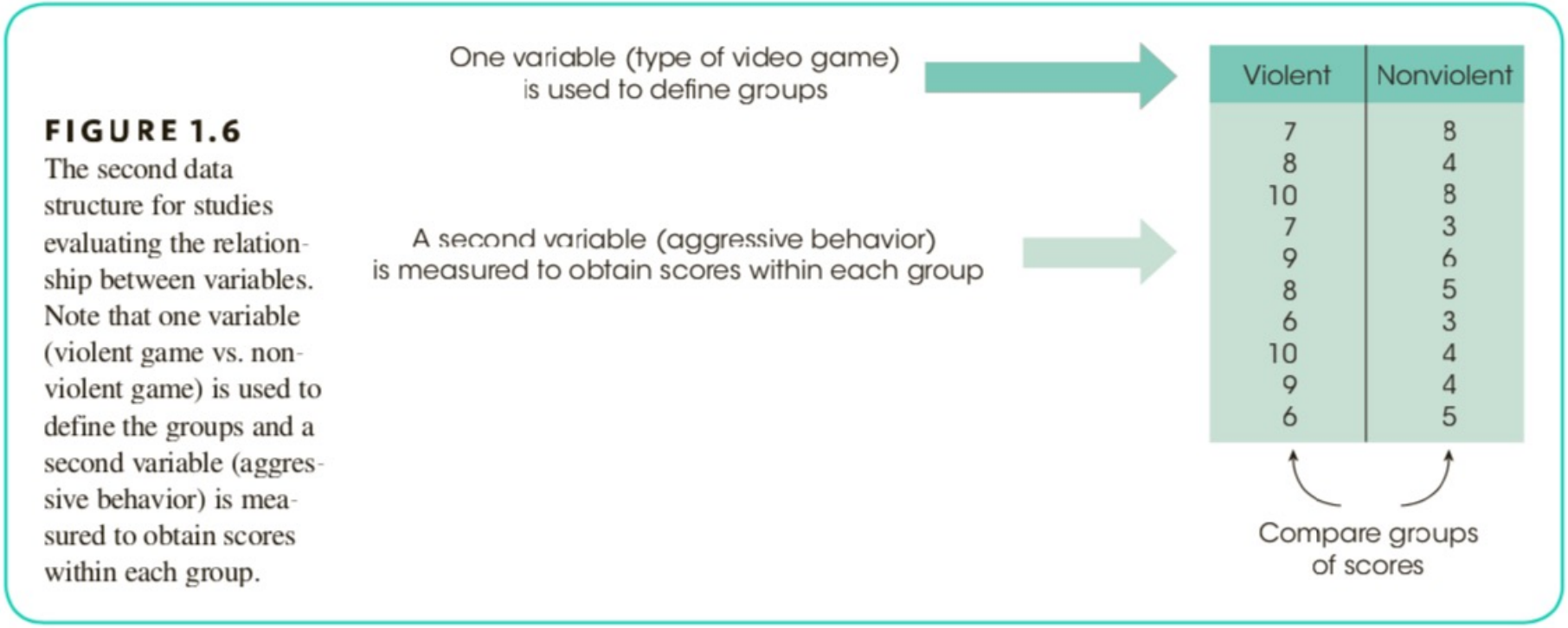


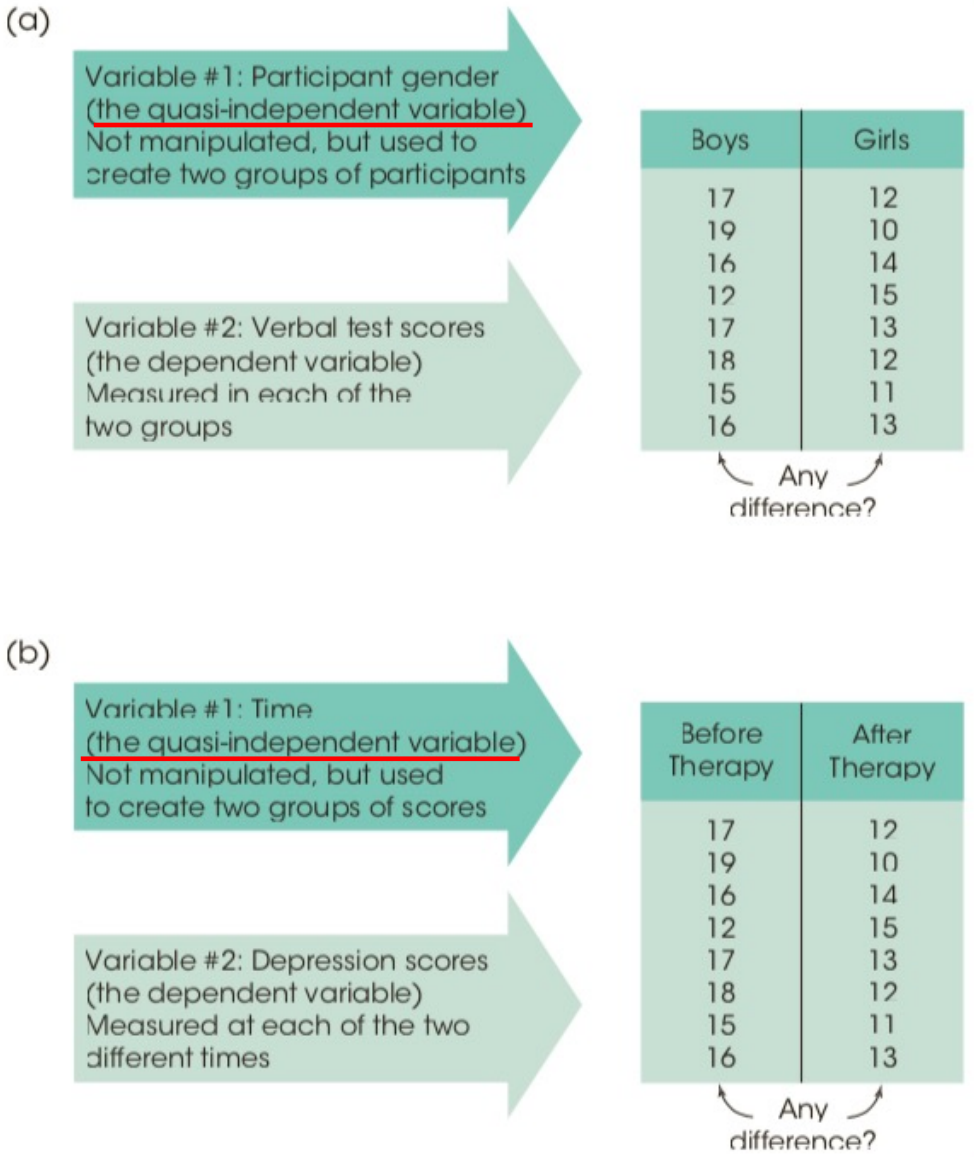

**FIGURE 1.6**
The second data structure for studies evaluating the relationship between variables. Note that one variable (violent game vs. nonviolent game) is used to define the groups and a second variable (aggressive behavior) is measured to obtain scores within each group.

# Terminology in the Experimental Method

- The **independent variable** is the variable that is manipulated by the researcher. In behavioral research, the independent variable usually consists of the two (or more) treatment conditions to which subjects are exposed. The independent variable is manipulated *prior* to observing the dependent variable.
- The **dependent variable** is the one that is observed to assess the effect of the treatment.
- Individuals in a **control condition** do not receive the experimental treatment. Instead, they either receive no treatment or they receive a neutral, placebo treatment. The purpose of a control condition is to provide a baseline for comparison with the experimental condition.
- Individuals in the **experimental condition** do receive the experimental treatment.

# Example of nonexperimental methods

**FIGURE 1.7**
Two examples of nonexperimental studies that involve comparing two groups of scores. In (a), a participant variable (gender) is used to create groups, and then the dependent variable (verbal score) is measured in each group. In (b), time is the variable used to define the two groups, and the dependent variable (depression) is measured at each of the two times.



# Terminology in Nonexperimental Research

- In a nonexperimental study, the “independent variable” that is used to create the different groups of scores is often called the **quasi-independent variable**.

# Try R!

- Open Source project: *The R Project for Statistical Computing*, led by Ross Ihaka and Robert Gentleman
- http://cran.r-project.org.
- Download RStudio: a convenient graphic interface, download from http://posit.co/products/open-source/rstudio.

# Today's summary

- **Why statistics**
- **What it does**
- **Terms:**
  1) variable, parameter, data, statistic
  2) construct and operational definition
  3) descriptive vs inferential statistics
  4) four scales
  5) three data structures
  6) experimental vs non-experimental