VIII. STATISTICS

Part I

IN THIS CHAPTER:

- An introduction to descriptive statistics

- Measures of central tendency: mean, median, and mode

- Measures of spread, dispersion, and variability: range, variance, and standard deviation

- Descriptive measures only tell us about the sample being studied; they are not representative of a wider population

IN THIS AND THE FOLLOWING TWO SUBCHAPTERS, WE WILL DISCUSS THE ROLE OF STATISTICS IN CROSS-CULTURAL RESEARCH.

The information provided here will not substitute for formal instruction in statistical methods, but it will hopefully help you understand how statistics are used to answer cross-cultural questions, as well as how to interpret some basic statistics when you encounter them in cross-cultural studies.

Statistics is a broad subject encompassing the collection, analysis, and interpretation of qualitative and quantitative data. Statistics can be **univariate** (focused on one variable), **bivariate** (focused on two variables), or **multivariate** (focused on three or more variables). There are two overarching types of statistics we will cover in our statistics sections. The first is **descriptive statistics**.

THE ROLE OF DESCRIPTIVE STATISTICS IS TO OBTAIN **SUMMARY INFORMATION** ABOUT YOUR DATA. THERE ARE SEVERAL MEASURES THAT PROVIDE THIS KIND OF INFORMATION.

MEASURES OF CENTRAL TENDENCY

These measures use one number to convey the center of a distribution of scores.

MEAN:

The mean, or **average**, is calculated by adding all of the scores for a group of cases and dividing that number by the total number of cases. The mean is useful in determining central tendency when a group of scores is **evenly distributed**. When a group of scores is **skewed**, or has some extreme scores on one end, the mean can be a misleading measure of central tendency. Imagine what the mean income would be if one person in a community earned a million dollars a year and 99 others earned no more than fifty-thousand.

MEDIAN:

The **median** is found by putting a group of scores in rank-order and identifying the **middle score**. This score can be a valuable measure of central tendency when a group of scores is skewed, or has some extreme scores on one end that will affect the mean. The **median is unaffected by extreme scores on either end**. However, the median can be misleading if there are very few scores in the center of the distribution.

MODE:

The **mode** is the score that **occurs most frequently** in a group of cases. If there are many different scores, the mode is not a very useful number. However, if we group scores together that fall within certain ranges (such as in a bar graph), the modal group can give us useful information about central tendency.

Mode can also be used to more accurately describe distributions in which there are an equal number of scores on both the low end and the high end of the distribution, with fewer scores in the middle (this is called a **bimodal distribution**).

MEASURES OF SPREAD, DISPERSION, AND VARIABILITY

These measures help us to understand how the scores are spread out.

RANGE:

This is the simplest measure of variability. **Range** describes the **lowest and the highest scores** in a group of cases. The range can misrepresent the amount of variability in a group of cases if the cases are mostly clustered in the middle with a few scores at the high and low ends.

A better way to describe the amount of variability within a group of scores is to determine the distance of every score from the center and then calculate an average distance from the center. This is accomplished by calculating the **variance** and **standard deviation**.

VARIANCE:

Variance is found by subtracting the mean from every score and squaring the difference. The squared differences are then summed and the total is divided by the number of cases. The differences are squared so that the negative numbers that resulted from the first calculation (subtracting the means from every score) will not cancel out the positive numbers when summed.

STANDARD DEVIATION:

Standard deviation is calculated by taking the **square root of the variance**, which transforms the variance value to match the original scale. Because the variance and standard deviation use all scores, not just the lowest and highest as the range does, they give a better picture of the degree to which the scores spread out from the center.
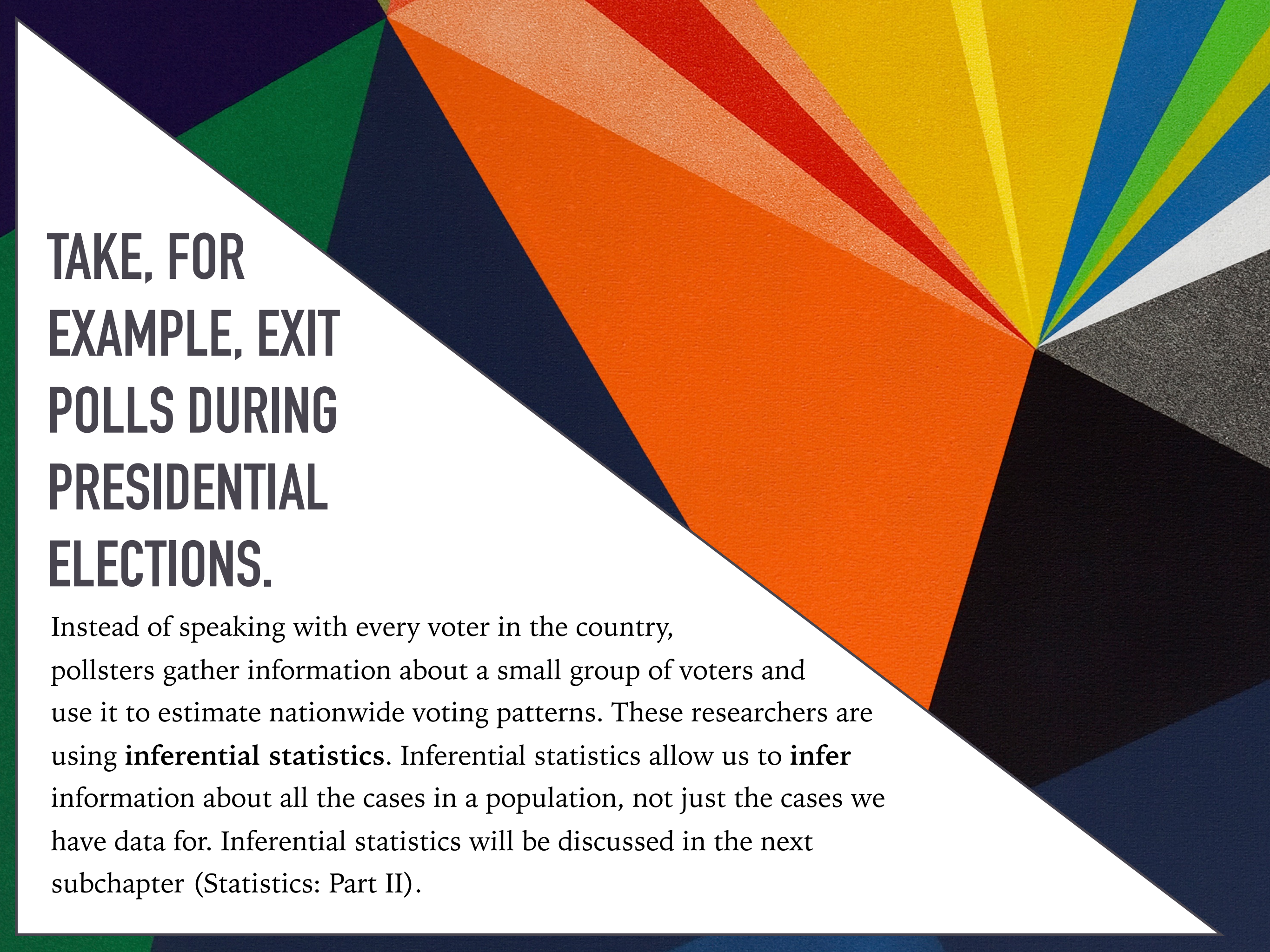
- **Depending on your data**, you may be interested in one of these measures more than others.

  Often, a combination of these measures— such as the mean and standard deviation— is the best option to paint a quick picture of your data.

A KEY FEATURE OF DESCRIPTIVE STATISTICS IS THAT THEY DESCRIBE ONLY THE CASES FOR WHICH YOU HAVE DATA.

If your population is a 5th grade classroom, you could obtain every student's test scores and present descriptive statistics on the class as a whole. However, it is frequently the case that researchers do not gather information about the entire population, but rather focus on a **sample of cases** with the goal of generalizing their findings to a wider population.

TAKE, FOR EXAMPLE, EXIT POLLS DURING PRESIDENTIAL ELECTIONS.

Instead of speaking with every voter in the country, pollsters gather information about a small group of voters and use it to estimate nationwide voting patterns. These researchers are using **inferential statistics**. Inferential statistics allow us to **infer** information about all the cases in a population, not just the cases we have data for. Inferential statistics will be discussed in the next subchapter (Statistics: Part II).

SUMMARY

- **Descriptive statistics** are primarily designed to summarize your collected data.

- **Measures of central tendency** use one number to convey the center of a distribution.

    - **Means** are most affected by extreme scores and skewed distributions

    - **Medians** are unaffected by extreme scores but are misleading if there are few cases in the middle of the range

    - **Modes** are useful for nominal categories and when scores are grouped

- **Measures of dispersion** help us understand the degree to which scores are spread out across a distribution.

- Most researchers use both a measure of central tendency and a measure of dispersion to describe their data.

CITATION AND TERMS OF USE

Using this course: The material in this course is intended for educational purposes only--either by individuals for personal use or for classroom use by instructors with appropriate attribution. For any commercial use or other uses please contact HRAF. (Email: <hraf@yale.edu>).

This chapter is from: Carol R. Ember. 2016. Introducing Cross-Cultural Research. Human Relations Area Files. <http://hraf@yale.edu/ccc/>