Preliminary Chapter

 This chapter presents an overview of the course, highlighting the four themes of statistics: data production, data analysis, probability and inference. You are not expected to leave this chapter with a command of statistics but rather with a sense of what we will be studying this year.

Objectives:

* + Give a definition of “Statistics” and describe what statistics helps you do.
	+ Explain why anecdotal evidence is not considered valid statistical data.
	+ Identify the three main statistical designs for producing data
	+ Explain the difference between individuals and variables.
	+ Differentiate between categorical and quantitative data.
	+ Describe what is meant by exploratory data Analysis.
	+ Describe the four key questions you should ask about a data set.
	+ Explain how probability helps us decide if an observation can be attributed to chance.
	+ Explain how sample values are used to make inferences about a population.

Case Study: Can Magnets Help Reduce Pain?

Early research has shown that magnetic fields affected living tissue in humans. Lately, doctors have begun to use magnets to treat patients’ pain. Scientists wondered whether magnets would reduce the chronic pain suffered by polio users so they designed a study to find out.

 Fifty patients with polio who reported steady pain were recruited for the study. A doctor identified a painful site on each patient and asked him or her to rate the pain on a scale of 0 (mild pain) to 10 (severe pain). Then the doctor selected a sealed envelope containing a magnet from a box that contained both active and inactive magnets. That way, neither the doctor nor the patient knew which type of magnet was being used. The chosen magnet was applied to the site of the pain for 45 minutes. After “treatment” each patient was again asked to rate the pain level from 0 to 10.

 In all 29 patients were given active magnets and 21 patients were given inactive magnets. All but 1 of the patients rated their initial pain as an 8, 9 or 10. So, scientists decided to focus on patients final pain ratings. Here they are grouped by the type of magnet used:

 Active: 0, 4, 7, 0, 4, 2, 5, 5, 3, 2, 2, 2, 3, 4, 6, 4, 3, 0, 2, 0, 4, 4, 5, 9, 10, 10, 10, 10, 7

 Inactive: 4, 7, 5, 8, 8, 6, 8, 10, 10, 6, 10, 8, 10, 10, 10, 10, 9, 9, 10, 10, 9

What do the data tell us about whether the active magnets actually helped with the pain? In this chapter, you will investigate the tools that statisticians use for data production, exploratory data analysis, probability, and drawing conclusions from the data.

Activity: Water, Water Everywhere!

Bottled water is very popular these days; but can a person REALLY tell the difference between tap water and bottled water?

1. Go to your assigned station and pick up 3 cups of water, labeled A, B and C. Take them to your seat and wait for more instructions. DO NOT DRINK ANYTHING YET.
2. When told, drink all of the water in cup A, then all in cup B and then all of cup C. Write down which cup you think has the bottled water in it. DO NOT TALK TO ANYONE!

CUP \_\_\_\_\_\_\_\_\_\_\_\_ is bottled water.

1. When everyone is done, record your results on the board.
2. Let’s assume that no one can actually identify the bottled water. In that case, students would be guessing which cup contains the bottled water. What percent of students would you expect to guess correctly? Explain.
3. Do you believe that the students who correctly identified the bottled water actually could tell the difference? Lets perform a *simulation* before we answer this question. We’ll assume that everyone guessed which cup had the bottled water. Then each student would have a 1 in 3 chance of guessing correctly. Roll your die times (once for each person in the class) Let 1 or 2 represent a correct guess. Let 3 to 6 be an incorrect guess. Record the number of correct guesses.
4. On the number line on the board, make an X above the number of correct guesses in your simulations.
5. What kind of conclusions can we draw about the class’ ability to distinguish between bottled and tap water?

What is Statistics? Do cell phones cause brain cancer? How well do SAT scores predict college success? Should arthritis sufferers take Celebrex to ease their pain, or are the risks too great? What percent of U.S. children are overweight? How strong is the evidence for global warming? These are just a few of the questions that statistics can help answer. But what is statistics? And why should you study it?

Statistics:

 The science (and art!) of learning from data

The study of Statistics has four main themes:

* (Exploratory) Data Analysis (EDA)
* Producing Data
* Probability
* Statistical Inference

Chapters 1-4 will focus on exploratory data analysis:

 Data are numbers with a context

 Data is found everywhere, TV, newspapers, magazines, Internet

 In these chapters we will earn to describe data, picture it, and summarize it!

Producing Data: Chapter 5 will focus on producing data that has meaning for us. Data that is haphazardly collected will not help us answer our questions; rather it will lead us to more questions not answers.

Probability is the study of chance behavior. The laws of probability studied in chapters 6 – 9, will help us to decide how likely an outcome is.

Statistical Inference: Making decisions about a large group, the population, based on a smaller group, the sample

* Population: the set of all possible individuals
* Sample: a subgroup of the population

Data Production: Where do we get GOOD data?

 It is tempting to draw conclusions from our own experiences and then apply our conclusion to all teens. Unfortunately, this is not the best way to go. Consider the following example:

 **Example 1:** A plane crash kills 300 people. The same day a car accident in your town kills 1 person. You therefore conclude that flying is not safe. What is wrong with this thinking?

 **Example 2:**If you visit the National Center for Health Statistics Website: [www.cdc.gov/nchs](http://www.cdc.gov/nchs), you will learn several things.

* Accidents are the most common cause of death for US Citizens ages 20 – 24. Homicide is second followed by suicide. AIDS is 7th behind heart disease and cancer.
* Data also shows that it is dangerous to be a young man; the overall death rate for men ages 20-24 is **3** times that for a women.

Can we conclude that young men ages 20-24 should lock themselves in their homes and refuse to come out if they wish to reach age 25?

 **Example 3:**Suppose you want to find out if HHS students prefer McDonald’s burgers or Burger King Burgers. You decide to hang out in McDonald’s since the dining room is nicer. You ask 50 teens who look friendly which burger they prefer. What is wrong with the data you have collected?

Available Data:

 Data that were produced in the past for some other purpose but that may help answer a present question.

**Three main designs for collecting data:**

* **Survey select a sample/population, ask questions and record responses**
* **Observational Study** observe individuals and measure the variable(s) of interest- no attempt is made to influence responses
* **Experiment** Apply a treatment to individuals and measure the responses

**When collecting data be careful where the data comes from as seen in the next example.**

 **Example 4:** Ann Landers once conducted a survey by asking her readers “If you had to do it over again, would you have children?” A few weeks later, her column had the headline: “70% of parents say kids are not worth it!” Indeed nearly 70% of the 10,000 parents who responded claimed they would not have had children if they could make the choice again. Do you believe that 70% of all parents really wish they never had children? Explain.

Observational Study vs. Experiment

An observational study does not attempt to change conditions or impose a treatment. Data is collected from interviews, reviewing previous records, or watching and recording.

An experiment imposes a treatment of some type and measures the response to this treatment.

 **Example 5:** Should women take hormones after menopause? In 1992, several medical organizations answered this question with a resounding yes. Women who took hormones seemed to reduce their risk of heart attack by 35 to 50%. The risks associated with taking hormones seemed small when compared to the benefits. This evidence came from a number of studies that simply compared women who were taking hormones with those who were not. What type of data collection method was used? What are some problems with drawing the conclusions based on this?

Cause and effect relationships can only be determined by a well-designed experiment!

Data Analysis: Making Sense of Data

The first step in understanding data is to hear what the data say, to “let the statistics speak for themselves.” But numbers speak clearly only when we help them speak by organizing, displaying, summarizing, and asking questions. That's ***data analysis.***

**Individuals**: are objects described by a set of data; they may be people but can be animals, plants, plots of land, or other objects

**Variables:**  any characteristic of an individual that is being measured, can take on different values for different individuals

**4 Key Questions to ask about EVERY set of data**

1. Who are the individuals? Identify the individuals and how many there are.
2. What are the variables and their units of measure (if applicable)?
3. Why was the data collected? Is there a specific question to be answered?
Will a generalization be made for a population?
4. When, where, how, and by whom were the data produced? Is it available data or new data? How old is the data? An observational study or an experiment? Who funded the study? Any Possible sources of conflict?

**Two Types of Variables:**

1. Quantitative: numerical data in which mathematical operations make sense

 Examples: height, weight, test scores

2. Categorical: places an individual into one of several categories

 Examples: eye color, hair color, zip code

**Example 6:** Here is a small part of a data set that describes public education in the United States:



Answer the four key questions about these data.

A variable is called a variable since it takes on various values depending on the situation. The pattern of variation is called the distribution.

Distribution: The distribution of a variable tells us what values the variable takes and how often it takes those values.

Statistical tools and ideas can help you examine data in order to describe their main features. This examination is sometimes called ***exploratory data analysis.*** (We prefer data analysis.) Like an explorer crossing unknown lands, we first simply describe what we see. Each example we meet will have some background information to help us, but our emphasis is on examining the data. Here are two basic strategies that help us organize our exploration of a set of data:

• Begin by examining each variable by itself. Then move on to study relationships among the variables.

• Begin with a graph or graphs. Then add numerical summaries of specific aspects of the data.

**Describing Categorical variables**

**Example 7:** Each year, the National Highway and Traffic Safety Administration (NHTSA) conducts an observational study on seat belt use. The table below shows the percent of front-seat passengers who were observed to be wearing their seat belts in each region of the United States in 1998 and 2003.**5**



1. What do these data tell us about seat belt usage by front-seat passengers?
2. Sketch a bar graph of 2003.
3. Sketch a side-by-side bar graph for the data.

**Describing Quantitative variables**

**Example 8:** The number of goals scored by the U.S. women's soccer team in 34 games played during the 2004 season is shown below:

3 0 2 7 8 2 4 3 5 1 1 4 5 3 1 1 3 3 3 2 1 2 2 2 4 3 5 6 1 5 5 1 1 5

What do these data tell us about the performance of the U.S. women's team in 2004?

A ***dotplot*** of the data is shown below. What can we see from this graph?



Making a statistical graph is not an end in itself. After all, a computer or graphing calculator can make graphs faster than we can. The purpose of a graph is to help us understand the data. After you (or your calculator) make a graph, always ask, “What do I see?”

**Exploring relationships between Variables**

A common goal of statistics is to determine what, if any, is the relationship between two variables. For example, we may wish to know how a student’s GPA is related to their score on the ACT. Sometimes the relationship isn’t as clear cut as we would like to be as seen in the next example.

**Example 9:** Air travelers would like their flights to arrive on time. Airlines collect data about on-time arrivals and report them to the Department of Transportation. Here are one month's data for flights from several western cities for two airlines:



What percent of flights were late for each airline?

This isn't the whole story, however. For each flight (individual), we have data on two categorical variables: the airline and whether or not the flight was late. Let's add data on a third categorical variable, departure city.**7** The following table summarizes the results.



Find the percent of flights that were late for each airline **by city.**

This example represents the concept of a lurking variable.

**Lurking variable: A variable in the background that affects the relationship between 2 variables- usually the statistician is unaware of the lurking variable**

**Probability: What are the chances?**

 Consider tossing a single coin. The result is a matter of chance. It can't be predicted in advance, because the result will vary if you toss the coin repeatedly. But therea is still a regular pattern in the results, a pattern that becomes clear only after many tosses. This remarkable fact is the basis for the idea of ***probability.***

**Example 10:** When you toss a coin, there are only two possible outcomes, heads or tails. Figure P.5 shows the results of tossing a coin 1000 times. For each number of tosses from 1 to 1000, we have plotted the proportion of those tosses that gave a head. The first toss was a head, so the proportion of heads starts at 1. The second toss was a tail, reducing the proportion of heads to 0.5 after two tosses. The next three tosses gave a tail followed by two heads, so the proportion of heads after five tosses is 3/5, or 0.6.



What happens to the proportion of heads **in the long run?**

Key concept of Probability: Chance behavior is unpredictable in the short run but has a regular and predictable pattern in the long run.

**Example 11:** How can probability help us determine whether students can distinguish bottled water from tap water? Suppose that in Mrs. Newberry’s class, 13 out of 21 students made correct identifications. If we assume that the students in his class *cannot* tell bottled water from tap water, then each one is basically guessing, with a 1-in-3 chance of being correct. So we'd expect about one-third of his 21 students, that is, about 7 students, to guess correctly. How likely is it that as many as 13 of his 21 students would guess correctly?



So what do we conclude?

**Statistical Inference: Drawing conclusions from Data**

As the previous example shows, probability allows us to decide whether an observed outcome is too unlikely to be due to chance variation. Too many students were able to identify which of their three cups contained a different type of water for us to believe that they were guessing. In effect, we tested the claim that the students were guessing. This is our first encounter with ***statistical inference.*** Notice the important role that probability played in leading us to a conclusion.

Statistical inference allows us to use the results of properly designed experiments, sample surveys, and other observational studies to draw conclusions that go beyond the data themselves. Whether we are testing a claim, as in the bottled versus tap water Activity, or computing an estimate, as in the Gallup survey, we rely on probability to help us answer research questions with a known degree of confidence. Unfortunately, we cannot be *certain* that our conclusions are correct. The following example shows you why.

**Example 12:** Most women who reach middle age have regular mammograms to detect breast cancer. Do mammograms really reduce the risk of dying of breast cancer? To seek answers, doctors rely on “randomized clinical trials” that compare different ways of screening for breast cancer. We will see later that data from randomized comparative experiments are the gold standard. The conclusion from 13 such trials is that mammograms reduce the risk of death in women aged 50 to 64 years by 26%.

On average, then, women who have regular mammograms are less likely to die of breast cancer. Of course, the results are different for different women. Some women who have mammograms every year die of breast cancer, and some who never have mammograms live to 100 and die when they crash their motorcycles. In spite of this individual variation, the results of the 13 clinical trials provide convincing evidence that women who have mammograms are less likely to die from breast cancer. That's because probability tells us that the large difference in death rates between women who had regular mammograms and those who didn't was unlikely to have occurred by chance. Can we be *sure* that mammograms reduce risk on the average? No, we can't be sure. **Because variation is everywhere, we cannot be certain about our conclusions.** However, statistics helps us better understand variation so that we can make reasonable conclusions.