Chapter 1: Stats Start Here

Chapter 2: Data

AN AVERAGE TEACHER MIGHT ASSIGN THREE PAGES OF HOMEWORK.
A MEAN TEACHER WOULD ASSIGN TEN PAGES.
OK
YOU SAID "MEAN" AND "AVERAGE" WERE THE SAME THING.
MATHEMATICALLY.
I JUST DON'T WANT TO BELIEVE THE AVERAGE PERSON IS MEAN.

Statistics

- ✓ Much of the world is **data driven** and **statistics** help us make sense of all of the data around us.
- ✓ Statistics is about how to think clearly with data. Being able to **communicate** clearly and concisely is an important skill that we will utilize and develop in this course.
- ✓ **Statistics** is all about variation… Data vary because we don't see everything and because what we do see and measure, we measure imperfectly.

What is ***statistics***? Not everyone knows what to expect when they enter this class.

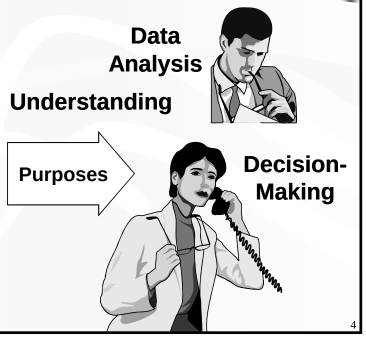
Statistics show up all around us...

- If you are a sports fan, you see stats on ESPN or in the paper.
- Ever see a commercial for a new drug that talks about side effects found in a double-blind study?
- Want to go to Vegas and win big? We'll learn all the odds in here (By the way, only the casino owners win big.)
- Seen all the political surveys? There's a method to how all of that polling is done.

What Is Statistics?

The science of data
Involves:

- ➢ *Collecting*
- ➢ *Classifying*
- ➢ *Summarizing*
- ➢ *Organizing*
- ➢ *Analyzing*
- ➢ *Interpreting*



Data Analysis

Understanding

Purposes

Decision-Making

The Three Steps to Statistics

- **There are three simple steps to doing Statistics right:**

Think first. Know where you're headed and why.

Show is about the mechanics of calculating statistics and making graphical displays, which are important (but are not the most important part of Statistics).

Tell what you've learned. You must explain your results so that someone else can understand your conclusions.

The Three Steps to Statistics

<u>Think</u>

Always use your head. Ask yourself where you are going and why.

<u>Show</u>

Always let people know what you are doing. This is the mechanics aspect of Statistics. Show your calculations, graphs, and displays of your data.

<u>Tell</u>

Always let people know what you learned. Until you've explained your results so that someone can understand your conclusion, the jobs not done!

This pie chart shows how much p
I ate while making this chart

Example

Sexual Discrimination Problem

Recently, a large company had to fire 10 employees because of the sluggish economy. Of these 10 employees, 5 were women. However, only 1/3 of the company's employees were women. This discrepancy has led the women who were fired to file a sexual discrimination lawsuit. Do they have a legitimate claim?

Example

1. What are the two options in this case?
   - ***they have a legitimate claim--the company fired a higher proportion of women on purpose***
   - ***they don't have a legitimate claim--this could have occurred by random chance***
2. Which of the two options can we actually assess?
   - ***not the first one--we cannot know what the boss was thinking***
   - ***however, we can estimate the probability of getting a result as surprising as this by random chance***

Definitions

- ***Statistics*** is the science of collecting, analyzing, and drawing conclusions from data. It is a way of reasoning, along with collection of tools and methods, designed to help us understand the world. Statistics is about variation.
- The ***population of interest*** is the entire collection of individuals or objects about which information is desired.
- When you study an entire population, it is called a ***census***.
- A ***sample*** is a subset of the population, selected for study in some prescribed manner.

What is Statistics Really About?

- Statistics (plural) are particular calculations made from data.
- Data (quantitative or qualitative/categorical) are values with a context.
- Statistics is about variation.
- All measurements are imperfect, since there is variation that we cannot see.
- Statistics helps us to understand the real, imperfect world in which we live.

Definitions

- ***Descriptive statistics*** is the branch of statistics that includes methods for summarizing data.
- ***Inferential statistics*** is the branch of statistics which involves generalizing about a population based on information from a sample of that population.
- ***Statistical inference*** is the process of drawing these generalizations.

Definitions

- A ***variable*** is any characteristic whose value may change from one individual to another.

  ex: height, hair color, brand of car, GPA

Hair color

Time until a light bulb burns out

Definitions

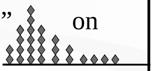
- ***Data*** can be numbers, record names, or other labels.
- Not all data represented by numbers are numerical data (e.g., 1 = male, 2 = female).
- Data are useless without their **context (see p. 8)**…
- Data are results from making observations on one or more variables. It is important to remember that data is not just a set of numbers, but **a set of numbers with a context.**

  ex: the numbers {78, 82, 83} have no meaning by themselves, but when told that they are students' test scores, they become meaningful.

Definitions

- A ***distribution*** shows the values a variable can take and how often it takes those values.

  ex: a generic dotplot with "*variable*" on the *x*-axis



- A ***univariate*** data set consists of observations of a **single variable.**

  ex: number of pencils, weight of backpack
- A ***bivariate*** data set consists of observations of **2 variables** for each member of the sample.

  ex: height and weight of students, GPA and SAT of students

Definitions

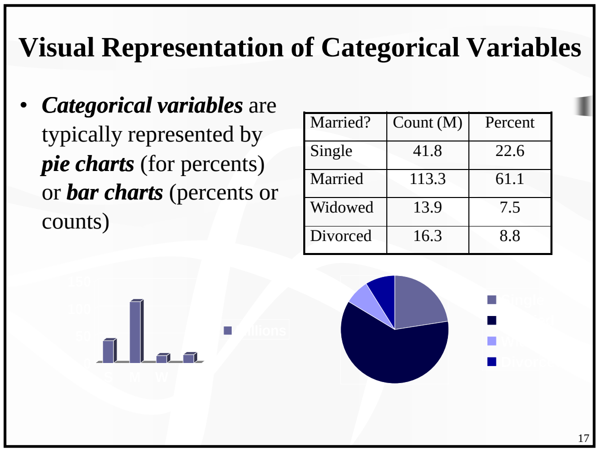
- A variable is ***categorical*** (or ***qualitative***) if the possible responses fall into categories.

  ex: brand of car, hair color (usually words)
- A variable is **numerical** (or ***quantitative***) if the possible responses are numerical in nature.

  ex: height, AP score (usually numbers)
- ***NOTE:*** **One way to tell the difference is to consider the question: "*Would it make sense to find the average of this variable?*" If you can, it's numerical. If you can't, it's categorical.**
- It is important to note that the word ***variable*** here is ***NOT the same*** as in Algebra. A ***variable*** is simply a characteristic of an individual that changes from case to case.

Types of Variables

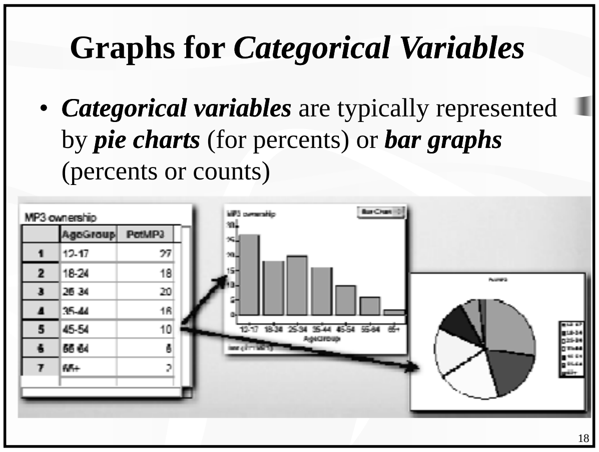
- ***Categorical variables*** record which group or category an individual belongs to (measure qualities or characteristics)
  - *What color is your hair?*
  - *What year are you in school?*
  - *What city do you live in?*
  - *It **does NOT** make sense to average the results*
- ***Quantitative variables*** take on numeric values
  - *How tall is a person?*
  - *What score did a person get on the SAT?*
  - *How many desks are in a room?*
  - *It **DOES** make sense to average the results*



Visual Representation of Categorical Variables

- ***Categorical variables*** are typically represented by ***pie charts*** (for percents) or ***bar charts*** (percents or counts)

Married?	Count (M)	Percent
Single	41.8	22.6
Married	113.3	61.1
Widowed	13.9	7.5
Divorced	16.3	8.8



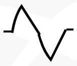
Graphs for *Categorical Variables*

- ***Categorical variables*** are typically represented by ***pie charts*** (for percents) or ***bar graphs*** (percents or counts)

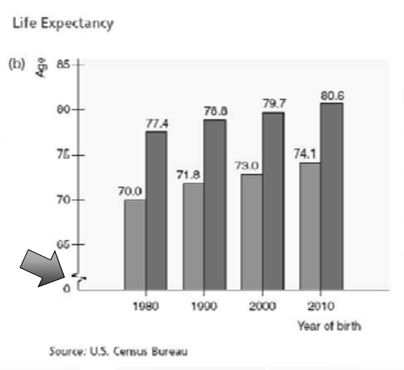
MP3 ownership

	AgeGroup	PctMP3
1	12-17	27
2	18-24	18
3	25-34	20
4	35-44	16
5	45-54	10
6	55-64	6
7	65+	2

Graphs for *Categorical Variables*

- Whenever you use a change of scale in a graphic, use a squiggle on the changed axis.
- A ***squiggle***: 

Bar Graph With a Change of Scale



Source: U.S. Census Bureau

Definitions

- Numerical data is ***discrete*** if the possible values are isolated points on the number line.

  ex: shoe size, number of birthdays
- Numerical data is ***continuous*** if the possible values form an entire interval on the number line.

  ex: foot length, age
- ***NOTE:*** **In general, you measure continuous variables and count discrete variables.**

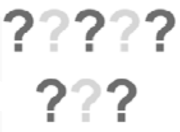
Example:

- For each of the following variables, determine if they are categorical or numerical. If its numerical, determine if it is continuous or discrete:

– length of a pen	◆ **numerical, continuous**
– type of pen	◆ **categorical**
– number of pens in a box	◆ **numerical, discrete**
– color of pants	◆ **categorical**
– number of pockets	◆ **numerical, discrete**
– length of inseam	◆ **numerical, continuous**
– subject of book	◆ **categorical**
– number of pages	◆ **numerical, discrete**
– area of a page	◆ **numerical, continuous**

The "W's"

- To provide context within a problem, we need to always consider the W's
  - **W**ho
  - **W**hat (and in what units)
  - **W**hen
  - **W**here
  - **W**hy (if possible)
  - and ho**W** of the data.



- ***Note:*** **the answers to "who" and "what" are essential.**

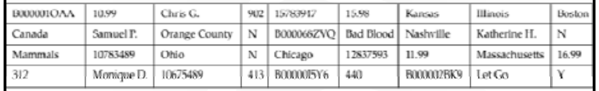
Individuals and Variables

When you meet a new set of data, ask yourself the following questions:

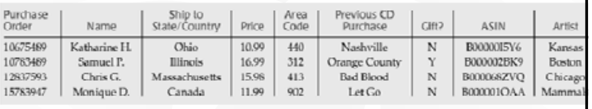
1. ***Who?*** What **individuals** do the data describe? **How many** individuals appear in the data?
2. ***What?*** How many **variables** are there? What are the **exact definitions** of these variables? In what **units** is each variable recorded?
3. ***Why?*** What is the reason the data were gathered? Do we hope to answer some specific questions? Do we want to draw conclusions about individuals other than the ones we actually have data for?

Read Chapter 2 Case: Amazon.com, p. 7-8

- Make sense with these data



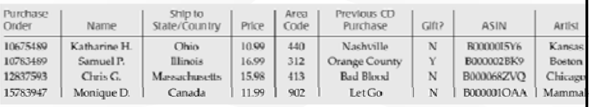
B000001OAA	10.99	Chris G.	902	15783947	15.98	Kansas	Illinois	Boston
Canada	Samuel P.	Orange County	N	B000066ZVQ	Bad Blood	Nashville	Katherine H.	N
Mammals	10783489	Ohio	N	Chicago	12837593	11.99	Massachusetts	16.99
312	Monique D.	10675489	413	B00000I5Y6	440	B000002BK9	Let Go	Y



Purchase Order	Name	Ship to State/Country	Price	Area Code	Previous CD Purchase	Gift?	ASIN	Artist
10675489	Katharine H.	Ohio	10.99	440	Nashville	N	B00000I5Y6	Kansas
10783489	Samuel P.	Illinois	16.99	312	Orange County	Y	B000002BK9	Boston
12837593	Chris G.	Massachusetts	15.98	413	Bad Blood	N	B000068ZVQ	Chicago
15783947	Monique D.	Canada	11.99	902	Let Go	N	B000001OAA	Mammal

Data Tables

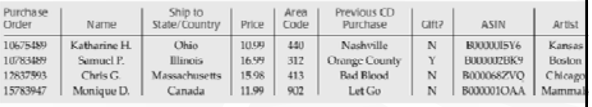
- The following **data table** clearly shows the context of the data presented:



Purchase Order	Name	Ship to State/Country	Price	Area Code	Previous CD Purchase	Gift?	ASIN	Artist
10675489	Katharine H.	Ohio	10.99	440	Nashville	N	B00000I5Y6	Kansas
10783489	Samuel P.	Illinois	16.99	312	Orange County	Y	B000002BK9	Boston
12837593	Chris G.	Massachusetts	15.98	413	Bad Blood	N	B000068ZVQ	Chicago
15783947	Monique D.	Canada	11.99	902	Let Go	N	B000001OAA	Mammal

- Notice that this data table tells us the ***What*** (column—***<u>what</u>*** has been recorded) and ***Who*** (row) the variables are about for these data.
- Are the data for each column categorical or numerical?

Data Tables



Purchase Order	Name	Ship to State/Country	Price	Area Code	Previous CD Purchase	Gift?	ASIN	Artist
10675489	Katharine H.	Ohio	10.99	440	Nashville	N	B00000I5Y6	Kansas
10783489	Samuel P.	Illinois	16.99	312	Orange County	Y	B000002BK9	Boston
12837593	Chris G.	Massachusetts	15.98	413	Bad Blood	N	B000068ZVQ	Chicago
15783947	Monique D.	Canada	11.99	902	Let Go	N	B000001OAA	Mammal

- Column—***What*** has been recorded
- Row— ***Who*** the variables are about
- ***<u>Remember</u>***: Even if people are involved, they may not be the ***Who*** of the data. For example, the ***Who*** here are the **<u>purchase orders</u>** (not the people who made the purchases)

Who

- The ***Who*** of the data tells us the individual **cases** for which (or whom) we have collected data. The rows of a data table correspond to individual cases about ***Whom*** (or about which—if they're not people) we record some characteristics.
  - ♦ Individuals who answer a survey are called **respondents**.
  - ♦ People on whom we experiment are called **subjects** or **participants**.
  - ♦ Animals, plants, and inanimate subjects are called **experimental units**.

Who (cont.)

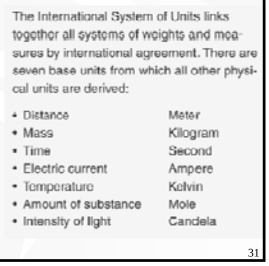
- Sometimes people just refer to data values as **observations** and are not clear about the ***Who***.
  - ♦ But we need to know the ***Who*** of the data so we can learn what the data say.

What and Why

- ***Variables*** are characteristics recorded about each individual. These are usually shown as the columns of a data table, and they should have a name that identify ***What*** has been measured.
- To understand variables, you must *Think* about what you want to know.

What and Why (cont.)

- Some variables have ***units*** that tell how each value has been measured and tell the scale of the measurement. ***Units*** tell us how each value has been measured.



The International System of Units links together all systems of weights and measures by international agreement. There are seven base units from which all other physical units are derived:

• Distance	Meter
• Mass	Kilogram
• Time	Second
• Electric current	Ampere
• Temperature	Kelvin
• Amount of substance	Mole
• Intensity of light	Candela

What and Why (cont.)

- A ***categorical*** (or ***qualitative***) variable names categories and answers questions about how cases fall into those categories.
  - *Examples*: sex, race, ethnicity
- A ***numerical*** (or ***quantitative***) variable is a measured variable (with units) that answers questions about the quantity of what is being measured.
  - *Examples*: income ($), height (inches), weight (pounds)

What and Why (cont.)

- The questions we ask a variable (the *Why* of our analysis) shape what we think about and how we treat the variable.

What and Why (cont.)

- ***Example:*** In a student evaluation of instruction at a large university, one question asks students to evaluate the statement "***The instructor was generally interested in teaching"*** on the following scale:

  1 = Disagree Strongly 2 = Disagree 3 = Neutral
  4 = Agree 5 = Agree Strongly
- ***Question:*** Is interest in teaching categorical or quantitative?
- Variables like ***interest in teaching*** are often called ***ordinal variables.***
  - With an ordinal variable, look at the *Why* of the study to decide whether to treat it as categorical or quantitative.

Example, #7, p. 17

To each description of data, identify WHO and WHAT were investigated and the population of interest.

• **Who**	➢ 2,500 cars
• **What**	➢ Distance from the bicycle to the passing car (in inches)
• **Population of interest**	➢ All cars passing bicyclists

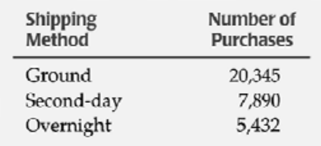
Example, #8, p. 17

To each description of data, identify WHO and WHAT were investigated and the population of interest.

• **Who**	➢ 30 similar companies
• **What**	➢ 401 (k) employee participation rates (in percent)
• **Population of interest**	➢ All similar companies

Counts Count

- When we count the cases in each category of a categorical variable, **<u>the counts are not the data</u>**, but something we summarize about the data.
  - The ***<u>category labels</u>*** are the ***What***, and
  - the individuals counted are the ***Who***.

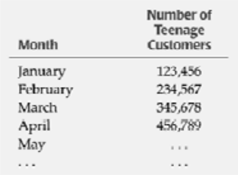


Shipping Method	Number of Purchases
Ground	20,345
Second-day	7,890
Overnight	5,432

37

Counts Count (cont.)

- When we focus on the amount of something, we use counts differently. For example, Amazon might track the growth in the number of teenage customers each month to forecast CD sales (the ***Why***).
  - The ***<u>What</u>*** is ***teens***, the ***<u>Who</u>*** is ***months***, and the ***<u>units</u>*** are ***number of teenage customers***.



Month	Number of Teenage Customers
January	123,456
February	234,567
March	345,678
April	456,789
May	. . .
. . .	. . .

38

Identifying Identifiers

- ***Identifier variables*** are categorical variables with **<u>exactly one individual</u>** in each category.
  - ***Examples***: Social Security Number, ISBN, FedEx Tracking Number
- Don’t be tempted to analyze identifier variables.
- Be careful not to consider all variables with one case per category, like *year*, as identifier variables.
  - The *Why* will help you decide how to treat identifier variables.

39

Where, When, and How

- We need the *Who*, *What*, and *Why* to analyze data. But, the more we know, the more we understand.
- ***When*** and ***Where*** give us some nice information about the context.
  - ***Example***: Values recorded at a large public university may mean something different than similar values recorded at a small private college.

40

Where, When, and How (cont.)

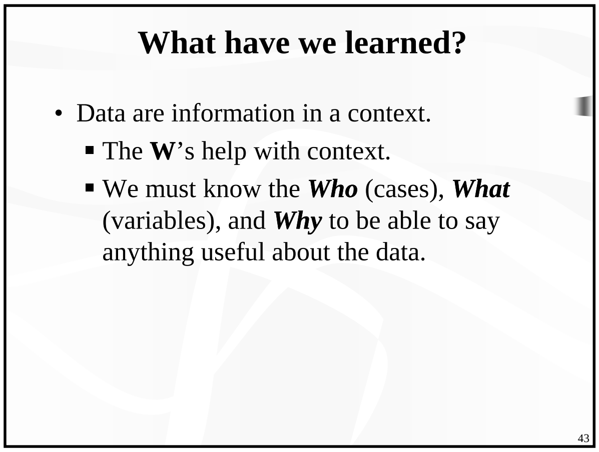
- ***How*** the data are collected can make the difference between insight and nonsense.
  - ***Example***: results from Internet surveys are often useless
- **<u>The first step of any data analysis should be to examine the W’s</u>**—this is a key part of the *Think* step of any analysis.
- And, make sure that you know the *Why*, *Who*, and *What* before you proceed with your analysis.

41

What Can Go Wrong?

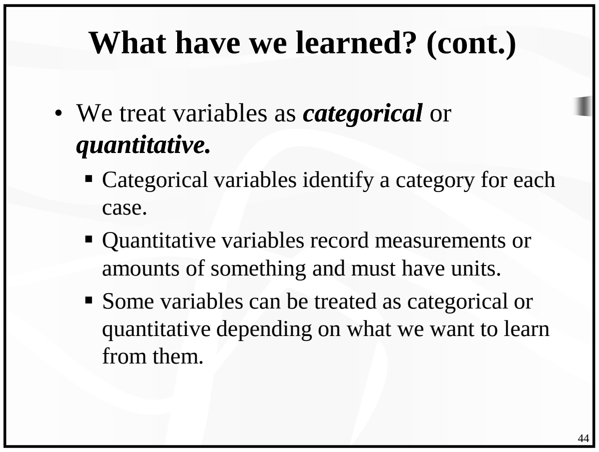
- ***Don’t label a variable as categorical or quantitative without thinking about the question you want it to answer.*** The same variable can sometimes take on different roles.
- ***Just because your variable’s values are numbers, don’t assume that it’s quantitative.*** Categories are often given numerical labels. Don’t let that fool you into thinking they have quantitative meaning. Look at the context.
- ***Always be skeptical—don’t take data for granted!***

42



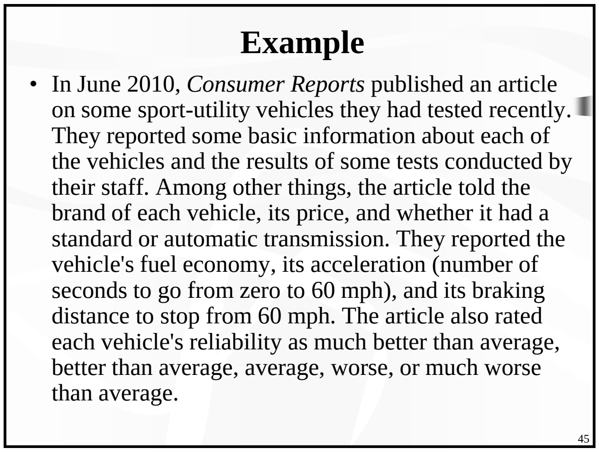
What have we learned?

- Data are information in a context.
  - The **W**’s help with context.
  - We must know the ***Who*** (cases), ***What*** (variables), and ***Why*** to be able to say anything useful about the data.



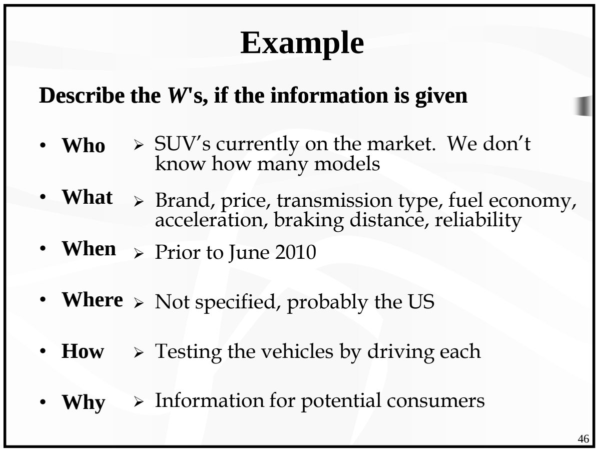
What have we learned? (cont.)

- We treat variables as ***categorical*** or ***quantitative.***
  - Categorical variables identify a category for each case.
  - Quantitative variables record measurements or amounts of something and must have units.
  - Some variables can be treated as categorical or quantitative depending on what we want to learn from them.



Example

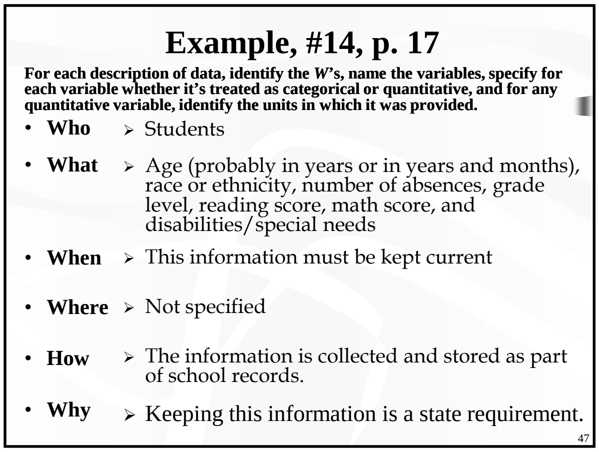
- In June 2010, *Consumer Reports* published an article on some sport-utility vehicles they had tested recently. They reported some basic information about each of the vehicles and the results of some tests conducted by their staff. Among other things, the article told the brand of each vehicle, its price, and whether it had a standard or automatic transmission. They reported the vehicle's fuel economy, its acceleration (number of seconds to go from zero to 60 mph), and its braking distance to stop from 60 mph. The article also rated each vehicle's reliability as much better than average, better than average, average, worse, or much worse than average.



Example

Describe the *W*'s, if the information is given

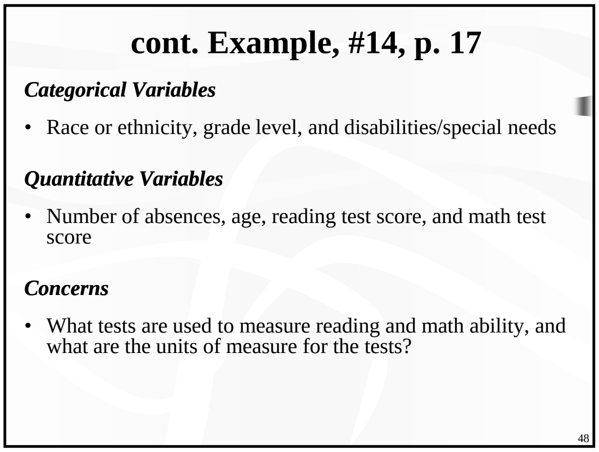
- **Who** ➢ SUV’s currently on the market. We don’t know how many models
- **What** ➢ Brand, price, transmission type, fuel economy, acceleration, braking distance, reliability
- **When** ➢ Prior to June 2010
- **Where** ➢ Not specified, probably the US
- **How** ➢ Testing the vehicles by driving each
- **Why** ➢ Information for potential consumers



Example, #14, p. 17

For each description of data, identify the *W*’s, name the variables, specify for each variable whether it’s treated as categorical or quantitative, and for any quantitative variable, identify the units in which it was provided.

- **Who** ➢ Students
- **What** ➢ Age (probably in years or in years and months), race or ethnicity, number of absences, grade level, reading score, math score, and disabilities/special needs
- **When** ➢ This information must be kept current
- **Where** ➢ Not specified
- **How** ➢ The information is collected and stored as part of school records.
- **Why** ➢ Keeping this information is a state requirement.



cont. Example, #14, p. 17

Categorical Variables

- Race or ethnicity, grade level, and disabilities/special needs

Quantitative Variables

- Number of absences, age, reading test score, and math test score

Concerns

- What tests are used to measure reading and math ability, and what are the units of measure for the tests?

Example

- List the variables. Indicate whether each variable is categorical or quantitative. **If the variable is quantitative, tell the units.**

➢ ***Categorical:*** Brand, transmission type, reliability

➢ ***Quantitative:*** Price ($), fuel economy (mpg), acceleration (second), braking distance (possibly feet)

The Worth of Data

- Collecting data is an extremely important part of any statistical study. If the data is not properly collected, it is worthless and we shouldn't use it to draw any conclusions. In general, there are 2 methods for collecting data: an ***observational study*** and an ***experiment***.

Definitions

- An ***observational study*** investigates characteristics of a sample in order to draw conclusions about a population.
  - *ex*: What is the average height of students at SDHS?
  - *ex*: Do girls have higher GPA's than boys at SDHS?
  - *ex*: Is there an correlation between GPA and SAT scores at SDHS?
- An ***experiment*** investigates how a response variable behaves when the researcher manipulates one or more factors (or explanatory variables). The purpose is usually to determine if changes in the explanatory variable *cause* changes in the response variable (what you are measuring).
  - *ex*: Does caffeine affect pulse rates?
  - *ex*: Does your seat location affect your grade? (Could be both?)
  - *ex*: Is your sense of taste affected by your sense of smell?

Definitions

- ***Note:*** In an ***observational study***, we **CANNOT** conclude that changes in the explanatory variable ***cause*** changes in the response variable because of the presence of confounding variables.
- A **confounding variable** is one that is related both to the explanatory variable and to the response variable.

  ex: "Increase in drowning deaths linked to rise in ice cream sales"
  - What is the explanatory variable?
  - What is the response variable?
  - What is the confounding variable?
  - What can we conclude?

Example

"Increase in drowning deaths linked to rise in ice cream sales"

- What is the explanatory variable?

  Ice cream sales
- What is the response variable?

  Drowning deaths
- What is the confounding variable?

  Temperature
- What can we conclude?

  We ***cannot*** conclude that an increase in ice cream sales ***causes*** an increase in drowning deaths, BUT we *can* predict that a month with high ice cream sales will also have high drowning deaths.

Example

- ***Study Links Mothers' Pesticide Exposure to ADHD in Children*** **(Article)**
- What were the explanatory and response variables?
  - ♦ **Explanatory = Pesticide and Response = ADHD**
- Was this an observational study or an experiment? How do you know?
  - ♦ **Observational study--researchers didn't manipulate the explanatory variable; they didn't intentionally expose mothers to various levels of pesticide.**
- What can we conclude?
  - ♦ **We can conclude that there is a correlation between a mother's exposure to pesticide and having children with ADHD. In other words, the higher levels of exposure to pesticide, the more likely the childe will have ADHD.**

Example

- ***Study Links Mothers' Pesticide Exposure to ADHD in Children*** **(Article)**
- Can we conclude that pesticide causes ADHD? Why or why not?
  - **No, there are possible confounding variables. For example, a mother exposed to pesticide may also garden, couldn't the excess work in gardening be the cause? Pesticide exposure could mean alternative diets, couldn't diet be the cause?**
- How could we prove that pesticide causes ADHD?
  - **We would need to do an experiment, randomly selecting some mothers and exposing them to various levels of pesticide (from none to a high dosage). Then, the effects of possible confounding variables will be spread equally among the sets of mothers. The only difference will be the pesticide levels.**

Expectations

- **My goal is for you to understand Statistics and how it plays a role in your current and future life.**
- **My hope is you will be successful in the course and earn a passing score on the AP Exam.**
- **My expectation is that you will put forth the effort necessary to be successful in a college-level course.**

Let´s Talk Stats...

- A claim has been made that students who study while listening to Mozart perform better on exams than students who listen to other music or no music at all.
- How could we test this claim?
  - How would you gather, analyze, and use data to study this?

The AP Exam

- **I will be preparing you for the Advanced Placement Statistics Exam taking place on Friday, May 9, 2014.**
  - **3 hours -- 40 Multiple Choice Questions, 5 Free-Response Questions, 1 "Investigative Task"**
- **You CAN be successful on this exam IF you put forth the effort ALL YEAR LONG.**
  - **I will provide you with LOTS of preparation materials as well as insight from the grading of the exam.**
  - **BUT I need you to provide the effort...**

Assignment

Chapter 1 Chapter 2	**<u>Lesson</u>**: Data	**<u>Read</u>**: Chapter 1 Chapter 2	**<u>Problems</u>**: 1 – 21 (odd)

Recall

Data are values along with their context. Data can be numbers or labels.

In order to determine the context of data, consider the "*W*'s"

- **Who –** *the cases (about whom the data was collected). People are referred to as **respondents, subjects, or participants, while objects are referred to as experimental units.***
- **What (and in what units) –** *the variables recorded about each individual.*
- **When –** *when the data was collected.*
- **Where –** *where the data was collected.*
- **How –** *how the data was collected.*
- **Why –** *why the data was collected. This can determine whether a variable is treated as **categorical** or **quantitative.***