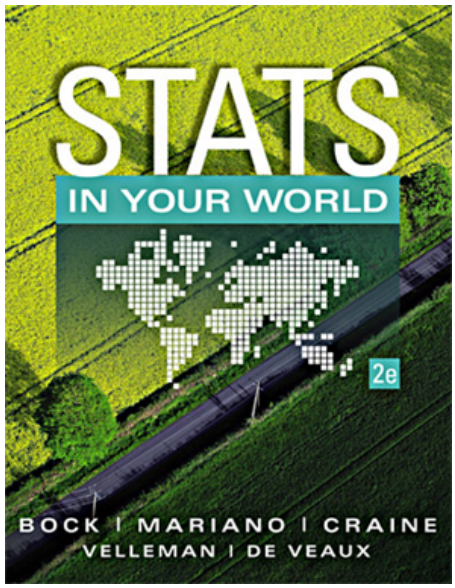
STRATISTICS

Section 1

Chapter 1:

Stats Starts Here



Sep 21-9:21 AM

What Is (Are?) Statistics?

Statistics (the discipline) is a way of reasoning, using a collection of tools and methods, designed to help understand the world.

Statistics (plural noun) are particular calculations made from data. For example, the mean and the median are statistics.

Data are values with a context.

Sep 1-2:16 PM

What is Statistics (the discipline) Really About?

Statistics is about variation.

People have different opinions about important issues. It can be important to see how their answers (and justifications) vary.

When we take measurements in an experiment, we expect individuals to be slightly different. How much difference is simply due to random variation? And when is a difference "large enough" that we believe something other than random variation is at work?

Sep 1-2:16 PM

Think, Show, Tell

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.

Sep 1-2:16 PM

What Are Data?

Data can be numbers, record names, or other labels.

Not all data represented by numbers are numerical data (i.e. 1 = male, 2 = female).

Other examples: Zip code, Phone number

Data are useless without their context!

Sep 1-2:16 PM

Data

- Try to arrange the following data in a way that makes sense:

Nashville Bad Blood

Chris G.
Katherine H.
Samuel P.

Chicago Illinois 312 Monique D.

N
N
N
Y

Canada 902

Orange County

Kansas Garbage

16.99
11.99
10.99
15.98

B00000I5Y6
B000068ZVQ
B000002BK9
B000001OAA

Let Go

Boston Massachusetts 413

Ohio 440

Sep 1-2:16 PM

The "W's"

To provide context we need the W's

- Who
- What (and in what units)
- When
- Where
- Why (if possible)
- and hoW

of the data.

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

Sep 1-2:16 PM

Data Tables

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

Katharine H.	Ohio	10.99	440	Nashville	N	B00000I5Y6	Kansas
Samuel P.	Illinois	16.99	312	Orange County	Y	B000002BK9	Boston
Chris G.	Massachusetts	15.98	413	Bad Blood	N	B000068ZVQ	Chicago
Monique D.	Canada	11.99	902	Let Go	N	B000001OAA	Garbage

Sep 1-2:16 PM

Who (and hoW)

The Who of the data tells us the individual *cases* for which (or whom) we have collected data.

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.

***The hoW (methodology used to create a study or experiment) will be examined in future chapters.*

Sep 1-2:16 PM

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.

In this course, we will discuss data collected from many different individuals. For example,

- Students in our school
- The 50 fastest roller coasters
- Godzilla movies

Question: Do we study every *case* that fits in the category of items we are interested in?

Take a sample of population

Sep 1-2:16 PM

What and Why

Variables (the what) are characteristics recorded about each individual.

The variables should have a name that identify **What** has been measured.

To understand variables, you must think about**What** you want to know.

Sep 1-2:16 PM

What and Why (cont.)

Some variables have units that tell how each value has been measured and tell the scale of the measurement. Variables with units are often *quantitative*.

Examples:

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

Sep 1-2:16 PM

What and Why (cont.)

A *categorical variable* names categories ***and*** answers questions about how cases fall into those categories.

Categorical examples: gender, race, ethnicity

A *quantitative variable* is a measured variable (with units) that answers questions about the quantity of what is being measured.

Quantitative examples: income ($), height (inches), weight (pounds)

Sep 1-2:16 PM

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.

Can you think of a way that a measured variable (given in units) could be categorical?

Sep 1-2:16 PM

Counts Count

We usually use a *frequency table* (which tells how many cases fit in each classification) for categorical variables.

For example, here is a frequency table for the question "What shipping method was chosen?"

Freq

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

Sep 1-2:16 PM

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 **What** is teens.

The **Who** is months.

The *units* are the *number of teenage customers*.

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

Sep 1-2:16 PM

Quantities

When the values you want to study are quantities, you are examining a quantitative variable.

For example, the question "How many AP tests will you take?" is a quantitative variable.

Student	# of AP Exams
Doug	2
Carl	4
Erica	7
Deanna	8
Jason	1
Corey	0
Michael	3
Paul	4
Adam	5

Sep 1-2:16 PM

Quantities (as categories)

When there are quantities you use to divide the study into groups, you are using a it as a categorical variable.

For example, the question "What is the distribution of AP scores earned in AP Calculus?" uses numerical data as a categorical variable.

AP Exam Score	# of AP Exams
5	6
4	5
3	7
2	3
1	3

Categorical Quantitative

Sep 1-2:16 PM

Quantities

Another example of a quantitative variable is the question:

"How far did the catapult throw the soup can?".

Catapult	Distance (cm)
A	67
B	55
C	79
D	112
E	13
F	98
G	137
H	67
I	178

Sep 1-2:16 PM

Identifying Identifiers

Identifier variables are categorical variables with exactly one individual in each category.

Examples: Social Security Number, ISBN, FedEx Tracking Number

Other examples:

Don't analyze identifier variables.

Be careful not to consider all variables with one case per category, like year or name, as identifier variables.

The Why will help you decide how to treat identifier variables.

Sep 1-2:16 PM

Where and When

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 Hickman, Rock Bridge, or Battle, may mean something different than values collected at Tolton.

Sep 1-2:16 PM

How

How the data are collected can make the difference between insight and nonsense.

Example:

Results from voluntary surveys on the Internet are often useless.

The first step of any data analysis should be to examine the W's—this is a key part of the **Think** step of any analysis.

Often times, it is not practical to study all members of a *population*. Many times a *sample* (or a smaller portion of the population) is taken to observe and collect data from.

Sep 1-2:16 PM

What Can Go Wrong?

Don't label a variable as categorical or quantitative without thinking about the question you want it to answer.

Just because your variable's values are numbers, don't assume that it's quantitative (i.e., think about jersey numbers or zip codes).

ABS: Always be skeptical—don't take data for granted or make assumptions.

Sep 1-2:16 PM

What have we learned?

- Statistics is about data
- It helps us understand the world.
- We seek to describe how much data varies.
- 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.

Sep 1-2:16 PM

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.
- Sometimes we treat a variable as categorical or quantitative depending on what we want to learn from it, which means that some variables can't be pigeonholed as one type or another.

Sep 1-2:16 PM