

Chapter 1

goal: Organize the way we think about data.

Def: "Individuals" are the objects described by a set of data

- They can be people, animals, or things

Def: A "variable" is a characteristic of an individual

- They can take different values for different individuals.

Case Study: Hypnosis & the immune system.

- 65 college students
- 33 are easily hypnotized
- 32 are not easily hypnotized

All student have their white blood cell count measured.

Students are randomly assigned a treatment from the following:

- group 1 — hypnotized
- group 2 — relaxed in sensory deprivation tank
- group 3 — no treatment

Then, white blood cell counts were remeasured

Results: The hypnotized group showed the largest increase in white blood cells.

Within that group, those who were easily hypnotized showed the greatest increase.

Who are the "individuals?"
What are the "variables?"

Def: A "categorical variable" places each individual into one of several groups or categories.

Ex: group, hypnotized easily or not easily

Def: A "quantitative variable" takes numerical values for which arithmetic operations make sense.

Ex: white blood cell count before, cell count after

Why is the group # a categorical variable?

Case Study: Body Mass Index and Heart Disease

In 1976 they began with 115,818 women between the age of 30 - 55.

Each woman's BMI was measured.

Each woman was asked how much she weighed at age 18.

This cohort of women was followed for 14 years, to see which ones would develop heart disease.

Who are the individuals?

What are the categorical variables?

What are the quantitative variables?

We would now like to explore the data that we have collected.

Def: The "distribution" of a variable tells us what values a variable can take and how often it takes each value.

- It can be a table, graph, or function

Ex: Class Makeup of the First Day
slides 1, 2, 3

A "pie chart" emphasizes proportions of a whole.

A "bar graph" is more flexible. It can be used even when we aren't looking at parts of a whole.

These are both graphs for representing the distribution of a categorical variable.

A "Histogram" is one type of graph that can be used for quantitative variables.

It is similar to a bar chart except that we must define each bar to represent a range of values for the quantitative variable.

Ex: Weight Data
slides 4, 5, 6

We would like to analyze the information when we look at the histogram.

Specifically we want to note the shape
center
spread

Shape

Def: A distribution is "symmetric" if the left and right sides are approximately mirror images.

Def: A distribution is "skewed to the right (left)" if the right (left) side extends much further than the other.

Def: A distribution is "unimodal (bimodal)" if it has one (two) peaks.

It's also possible to have a "multimodal" distribution.

Center

For now we will eyeball this as a point such that half the observations fall on each side.

Spread

For now we will describe the spread as the min-max observation.

Ex: Iowa Test Vocabulary
slide 7

Describe the shape, center, spread.

Ex: Foreign Born Residents
slide 8

Describe the shape, center, spread.

Ex: SAT
slide 9

Describe the shape, center, spread.

Why might there be multiple peaks?

Ans: Typically multiple peaks means that there are distinct groups within the population.

In this case the groups are Eastern and western states.

Def: An "outlier" is an observation that falls clearly outside the overall pattern.

Ex: Women Who've Never Married
slide 10

Describe the shape, center, spread.

7

Another type of graph to represent quantitative variables is a "stem plot."

Ex: Weight Data
—— slide 11, 12 ——

Advantages of a stem plot over a histogram:

- Data values are preserved.

Disadvantages:

- Very Yucky for large data sets.

Another type of data we might have is a "time series" which is the values of a variable at different times.

A graph for a time series is called a "time plot."

Ex: Class Makeup on the first day
—— slide 13 ——

Ex: Public vs. Private Tuition
—— slide 14 ——