

## DATA: QUALITATIVE (CATEGORICAL) and QUANTITATIVE (NUMERICAL)

Data is obtained by measuring or observing characteristics of individuals of interest. We call these characteristics *variables*.

A **variable** is a characteristic that varies from one person or thing to another. (Example: a person's age, height, number of children, number of cars, eye color, favorite type of music, etc.)

Typically, a variable can describe either a quantitative or a qualitative characteristic of an individual.

A **quantitative** (numerical) variable is a variable that takes on numerical values. (A quantitative variable has natural units of measurement. The units of measurement tell us how much of something we have or how far apart two values are.)

Examples: age, height, weight, number of children, number of cars, exam score, distance one travels to get to school, number of coins in a person's pocket, etc.

A **qualitative** (categorical) variable is a variable whose values are descriptive, not numerical.

Examples: gender, race, eye color, favorite type of music, political party, birth month, handedness (left, right, ambidextrous), school district, etc.

### Comment #1:

Sometimes, numerical labels are used for qualitative (categorical) data.

Example: When we categorize individuals by eye color, we may denote blue eyes by 1, brown eyes by 2, green eyes by 3, and hazel eyes by 4. (This kind of data is sometimes called "coded data.") The use of numerical labels for categorical data does not make the data quantitative.

When numerical labels are used for qualitative (categorical) data, it will often make no sense to do arithmetic to the data (for example, taking an average).

Take a sample of 20 adults and suppose 4 of them have blue eyes, 7 have brown eyes, 2 have green eyes, and 7 have hazel eyes. The coded data would look like this:

1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 4, 4, 4, 4, 4, 4

We can average these numbers and get  $52/20 = 2.6$ , but this result has no meaningful interpretation.

Remember, data has context. If you are asked to analyze some data, make sure you are given the context - ask "What do the numbers in the data set represent?" When we realize the context of the numbers above is eye color, then we understand that taking an average of those numbers isn't meaningful. (Another example: area codes.)

### Comment #2:

Some variables can be considered quantitative (numerical) or qualitative (categorical), depending on *why* you are looking at it.

Example: Amazon.com is constantly monitoring and evolving their web site to best serve their customers and maximize their sales performance. To make changes to the site, they experiment, collecting data and analyzing what works best. Suppose Amazon asks for your *age* in years. That seems quantitative, and would be if they want to know the average age of those customers who visit their site after 3am. But suppose they want to decide which CD to offer you in a special deal. Then thinking of your age in one of the categories child, teen, adult, or senior might be more useful.

If it isn't clear whether a variable is categorical or quantitative, think about *why* you are looking at it and what you want it to tell you.

Comment #3:

A typical course evaluation survey asks, "How valuable do you think this course will be to you?" 1 = worthless; 2 = slightly; 3 = middling; 4 = reasonably; 5 = invaluable. Is the variable *educational value* categorical or quantitative? It depends on what we want it to tell us.

A teacher might just count the number of students who gave each response for her course, treating *educational value* as a categorical variable.

When the teacher wants to see whether the course is improving, for example, she might treat the student responses as the *amount* of perceived value - in effect, treating the variable as quantitative. But, what are the units of measurement? There is certainly an *order* of perceived worth; higher numbers indicate higher perceived worth. A course that averages 4.5 seems more valuable than one that averages 2, but we should be careful about treating *educational value* as purely quantitative. To treat it as purely quantitative, we have to imagine that it has "educational value units" or some similar arbitrary construction. Because there are no natural units, we should be cautious.

Variables like this that report order without natural units are often called **ordinal variables**.

Comment #4:

Some categorical variables are simply identifiers. My grade book has a lot of data in it, including things like student ID #, major, Test 1 score, Test 2 score, etc. The variable *student ID #* is categorical, but there is exactly one individual in every category. The variable is simply playing the role of an identifier, and analyzing it is not informative.



- (c) Babies. Medical researchers at a large city hospital investigating the impact of prenatal care on newborn health collected data from 882 births during 1998-2000. They kept track of the mother's age, the number of weeks the pregnancy lasted, the type of birth (cesarean, induced, natural), the level of prenatal care the mother had (none, minimal, adequate), the birth weight and sex of the baby, and whether the baby exhibited health problems (none, minor, major).

### QUANTITATIVE DATA: DISCRETE AND CONTINUOUS

Data consisting of quantitative (numerical) variables can be further divided into two groups: *discrete* and *continuous*.

If the set of all possible values the variable might take on, when pictured on the number line, consists only of isolated points, then the numerical data is **discrete**. There are gaps between consecutive possible values because discrete variables assume only isolated values.

If the set of all values the variable might take on, when pictured on the number line, consists of an interval of numbers (or several intervals of numbers), possibly unbounded, then the numerical data is **continuous**.

Note that discrete quantitative data involves *counting*. Continuous quantitative data involves *measurement*.

### Example

Suppose the fire department mandates that all firefighters must weigh between 150 and 250 pounds. The weight of a fire fighter would be an example of a continuous variable, since a fire fighter's weight could take on any value between 150 and 250 pounds.

Suppose we flip a coin and count the number of heads. The number of heads could be any integer value greater than or equal to 0. However, it could not be *any* number greater than or equal to 0. We could not, for example, get 2.5 heads. Therefore, the number of heads must be a discrete variable.

2. Circle "discrete" or "continuous" for the variables described below. Remember that discrete variables arise from counting, and continuous variables arise from measurement.

(a) The number of heads obtained when two coins are tossed together.

Discrete (arise from counting) or Continuous (arise from measurement)

(b) The total score obtained when two dice are thrown.

Discrete (arise from counting) or Continuous (arise from measurement)

(c) The weights of year 7 pupils at a nearby school

Discrete (arise from counting) or Continuous (arise from measurement)

(d) The time it takes you to get to school each day

Discrete (arise from counting) or Continuous (arise from measurement)

- (e) The number of touchdowns scored by the Jonesboro high school football team last season

Discrete (arise from counting) or Continuous (arise from measurement)

- (f) The heights of British basketball players aged 20 or over.

Discrete (arise from counting) or Continuous (arise from measurement)

- (g) The time it takes to run 100 meters.

Discrete (arise from counting) or Continuous (arise from measurement)

- (h) The number of matches in a box marked "average contents 50."

Discrete (arise from counting) or Continuous (arise from measurement)

- (i) Correct answers out of 10 questions on a mental arithmetic test

Discrete (arise from counting) or Continuous (arise from measurement)