Data Basics

Colby Community College

Definition

A **data matrix** is common way to organize data, especially if collecting data in a spreadsheet.

Example 1

	loan_amount	interest_rate	term	grade	state	total_income	homeownership
1	7500	7.34	36	A	MD	70000	rent
2	25000	9.43	60	B	OH	254000	mortgage
3	14500	6.08	36	A	MO	80000	mortgage
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	3000	7.96	36	A	CA	34000	rent

Definition

Each row is called a **case** or **observational unit**.

Definition

Each column is called a **variable**.

Note

It is very important to provide descriptions of all the variables in a data matrix. Be sure to include the units of measurement.

Data Set

	loan_amount	interest_rate	term	grade	state	total_income	homeownership
1	7500	7.34	36	A	MD	70000	rent
2	25000	9.43	60	B	OH	254000	mortgage
3	14500	6.08	36	A	MO	80000	mortgage
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	3000	7.96	36	A	CA	34000	rent

variable	description
loan_amount	Amount of the load received, in US dollars.
interest_rate	Interest rate on the loan, in an annual percentage.
term	The length of the loan, which is always a whole number of months.
grade	Loan grade, which takes values A through G and represents the quality of the loan and its likelihood of being repaid.
state	US state where the borrower resides.
total_income	Borrower's total income, including any second income, in US dollars.
homeownership	Indicates whether the person owns, owns but has a mortgage, or rents.

Definition

Numerical data consisting of numbers representing counts or measurements. Sometimes referred to as **Quantitative** data.

Note

Most of the data we will be considering in this course will be numerical.

Definition

Categorical data consisting of names or labels (not numbers). Sometimes referred to as **Qualitative** data.

Note

The names and labels in categorical data are sometimes coded with numbers. When a number is used as a name it is **not** numerical data.

Example 2

The ages (in years) of subjects enrolled in a clinical trail are?

Numerical data.

Example 3

The genders (male / female / non-binary) of subjects enrolled in a clinical trial are?

Categorical data.

Example 4

Identification numbers $1, 2, 3, \ldots, 25$ *are assigned randomly to 25 subjects in a clinical trail. The identification numbers are?*

Categorical Data.

Note

The numbers in Example 4 don't actually measure or count anything.

Definition

Discrete data result when the data values are numerical and the number of values is finite, or countable.

Example 5

Each of several physicians plans to count the number of physical examinations given during the next full week.

Example 6

Casino employees plan to roll a fair die until the number 5 is rolled, and they count the number of rolls required to get a 5. (It is possible, but unlikely, that a 5 could never be rolled.)

Definition

Continuous data result from infinitely many possible numerical values, where the collection of values is not countable.

Example 7

When a typical patient has blood drawn as part of a routine examination, the volume of blood drawn is between 0 mL and 50 mL.

Example 8

The unemployment rate for Los Angeles County is 4.69%.

Example 9

A grade school teacher measures the heights of his students.

Definition

The **nominal level of measurement** is characterized by data that consist of names, labels, or categories only. The data cannot be arranged in some meaningful order, such as low to high.

Example 10

A survey has responses of “yes”, “no”, and “undecided”

Example 11

For an item on a survey, respondents are given a choice of possible answers, and they are coded as follows:

1 is coded as “I agree”

2 is coded as “I disagree”

3 is coded as “I don’t care”

4 is coded as “I refuse to answer”

The numbers 1,2,3, and 4 don’t count or measure anything.

Definition

Data are at the **ordinal level of measurement** if they can be arranged in some order, but differences between data values either cannot be determined or are meaningless.

Example 12

A college professor assigns grades of A, B, C, D, or F. These grades can be arranged in a meaningful order, but grades are very individualized so the difference between two students grades cannot be calculated.

Example 13

A survey asks people what they felt their blood pressure was. The possible answers are “Low”, “Normal”, “High.” These responses can be arranged in a meaningful order, but the differences between “Low” and “High” doesn’t make sense.

Definition

Data are at the **interval level of measurement** if they can be arranged in order, and differences between data values can be found and are meaningful. *Data at this level do not have a natural zero starting point at which none of the quantity is present.*

Example 14

Body temperatures of 98.2°F and 98.9°F are examples of data at this interval level of measurement. The values are ordered, and we can calculate their difference. There is no natural starting point. (The value 0°F is an arbitrary choice and doesn't represent the complete absence of heat.)

Example 15

The years 1492 and 1776 can be arranged in order, and the difference of 284 years is meaningful. But, time doesn't not have a natural starting point that represents "no time."

Definition

Data are at the **ratio level of measurement** if they can be arranged in order, differences can be found and are meaningful, and there is a natural starting point which indicates that none of the quantity is present.

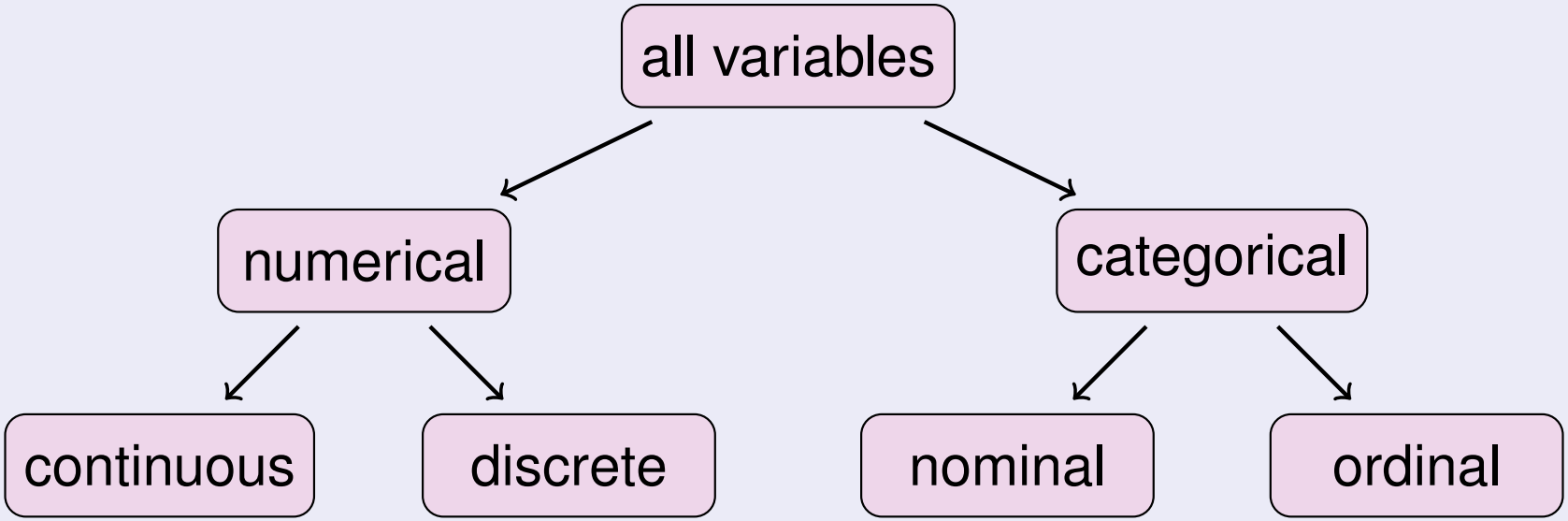
Example 16

Heights of 180cm and 90cm for a high school student and a preschool student are at the ratio level of measurement. (The starting point is 0cm and 180cm is twice as tall as 90cm.)

Example 17

The times of 50 minutes and 100 minutes for a math class are at the ratio level of measurement. (The starting point is 0 minutes and 100 minutes is twice as long as 50 minutes.)

Classification of Variables



Note

For simplicities sake, we will often treat ordinal data as nominal data.

Note

If your data consists of real numbers, then it is usually continuous.

Note

If your data consists of only integers, then it is usually discrete.

Data Set

The following slides will use the `county` data set, which contains data on all 3142 counties in the United States.

variable	description
name	County names.
state	State names.
pop2000	Population in 2000.
pop2010	Population in 2010.
pop2017	Population in 2017.
pop_change	Population change from 2010 to 2017.
poverty	Percent of population in poverty in 2017.
homeownership	Home ownership rate, 2006-2010.
multi_unit	Percent of housing units in multi-unit structures, 2006-2010.
unemployment_rate	Unemployment rate in 2017.
metro	Whether the county contains a metropolitan area.
median_edu	Median education level (2013-2017).
per_capita_income	Per capita (per person) income (2013-2017).
median_hh_income	Median household income.
smoking_ban	Describes whether the type of county-level smoking ban in place in 2010, taking one of the values ‘‘none’’, ‘‘partial’’, or ‘‘comprehensive’’.

Relationships Between Variables

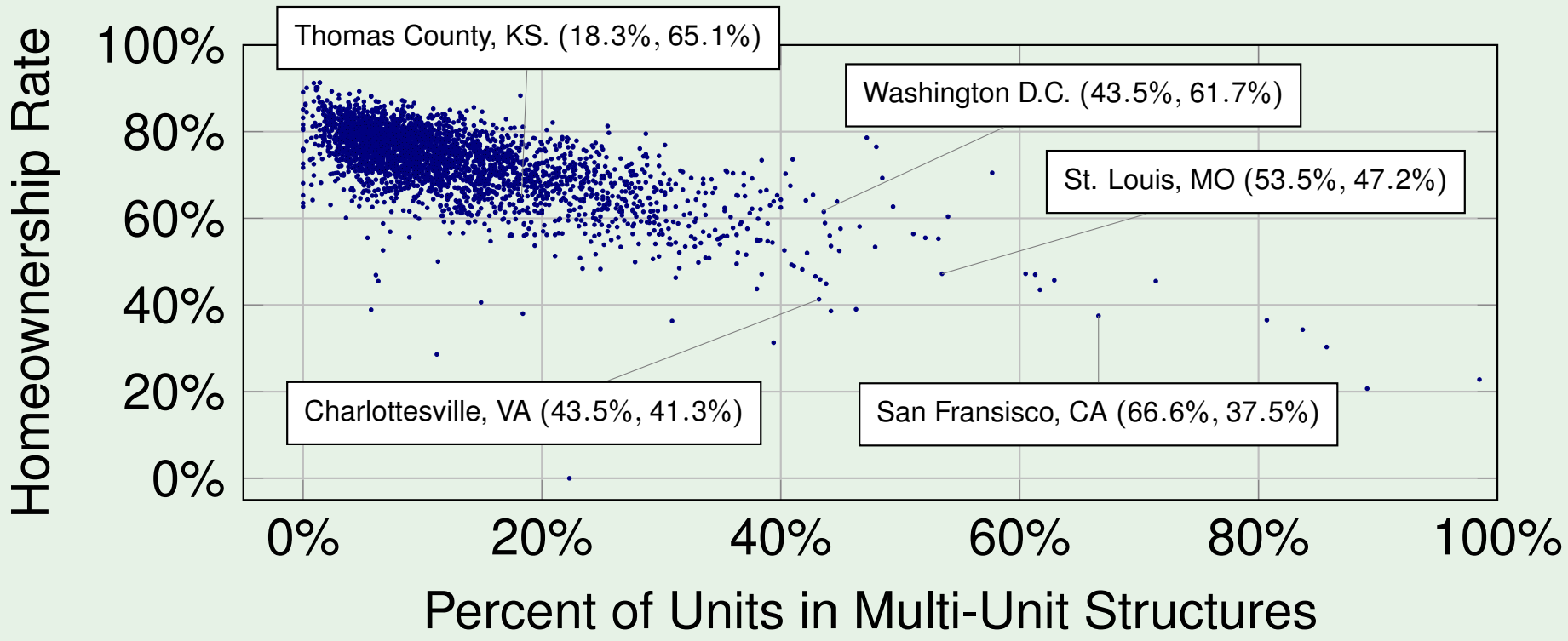
Often, researchers will want to study the relationship between variables.

- If homeownership is lower than the national average in one county, will the percent of multi-unit structures in that county tend to be above or below the national average?
- Does a higher than average increase in county population tend to correspond to counties with higher or lower median household incomes?
- How useful a predictor is median education level for the median household income for US counties?

Definition

A **scatterplot** is a plot of paired (x, y) numerical data with a horizontal x-axis and a vertical y-axis. The horizontal axis is used for the first variable (x), and the vertical axis for the second variable (y).

Example 18



Definition

When two variables show some connection with one another, they are called **associated** variables. Sometimes referred to as **dependent**.

Definition

If there is a downward trend in the scatter plot, the variables are said to be **negatively associated**.

Definition

If there is an upward trend in the scatter plot, the variables are said to be **positive associated**.

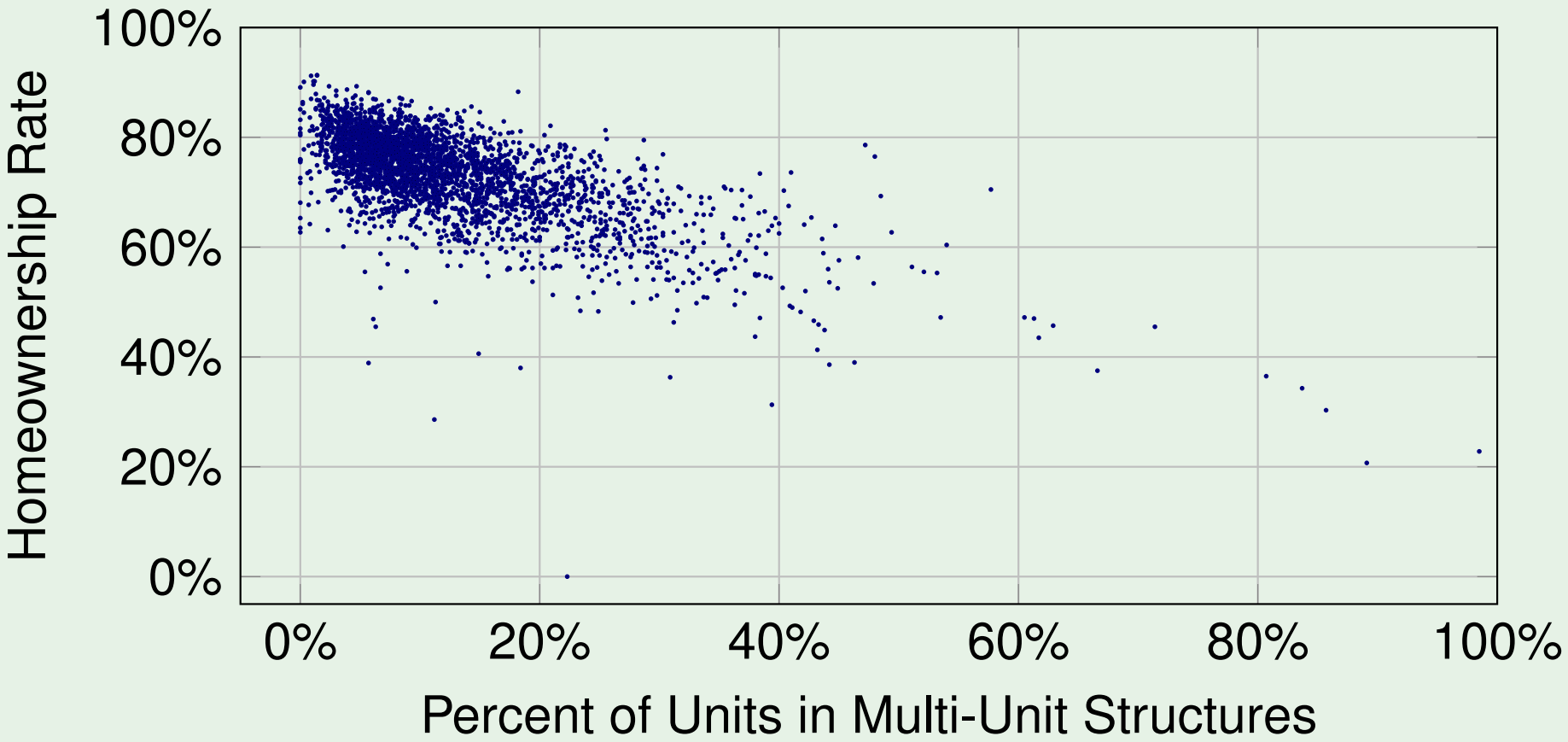
Definition

If variables are not associated, then they are said to be **independent**.

Note

It is impossible to be both associated and independent.

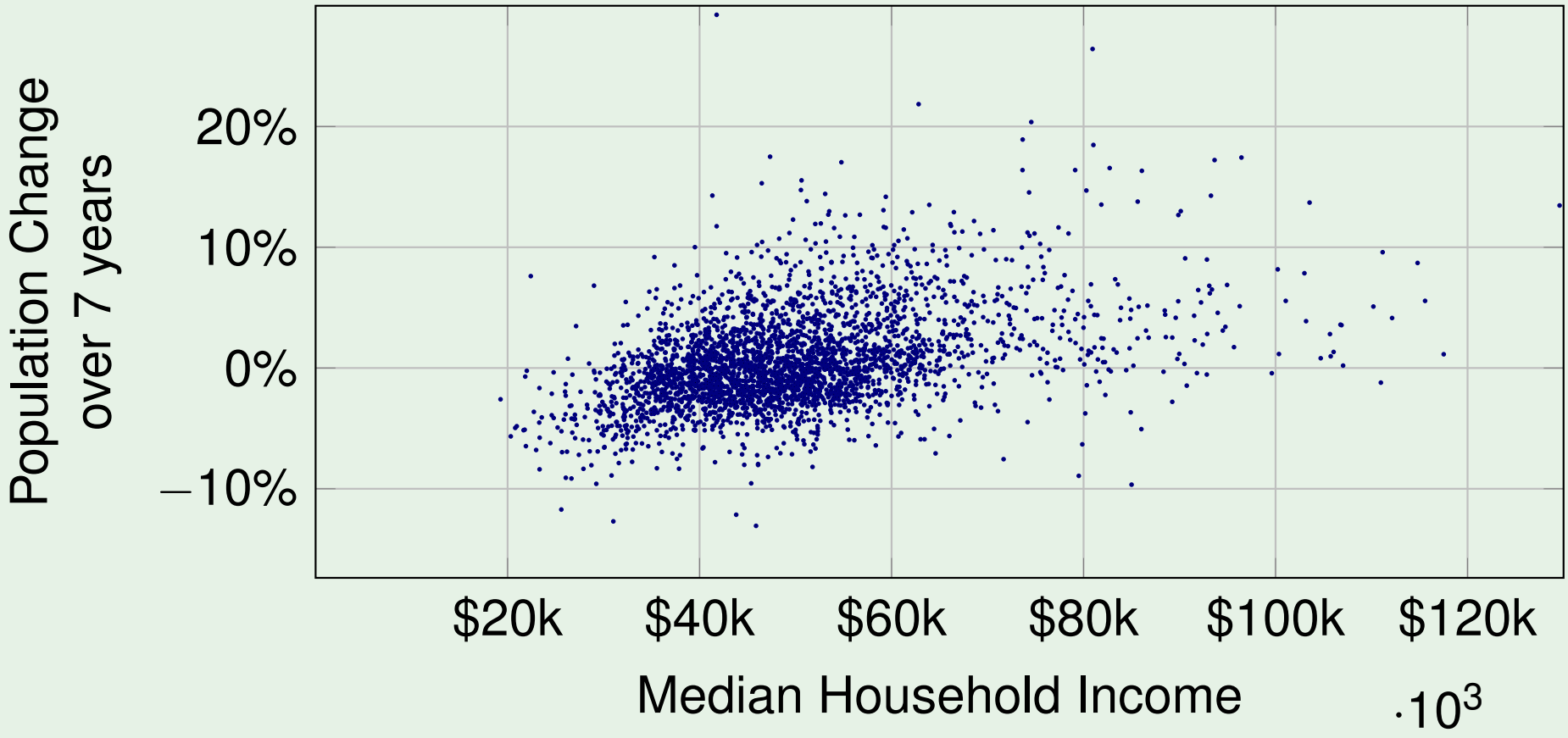
Example 19



Are the Multi-Unit Structure Rate and the Homeownership Rate associated?

Yes, they are negatively associated.

Example 20



Are the Median Household Income and the Population Change associated?

Yes, they are positively associated.

Example 21

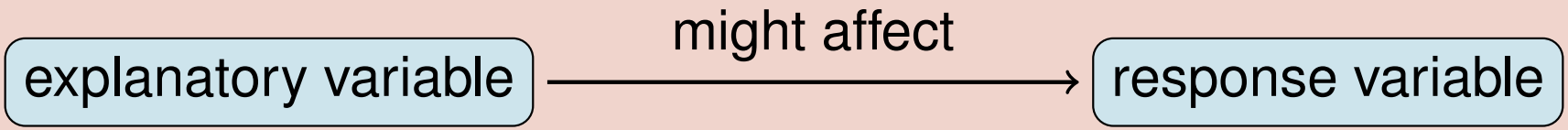
Consider the question:

If there is an increase in the median household income in a county, does this drive an increase in its population?

A higher median income is likely one of the causes of an increased population.

Definition

When we suspect one variable might causally affect another, we label the first variable the **explanatory variable** and the second the **response variable**.



Note

Labeling a variable like this does **nothing** to guarantee that a causal relationship exists!

Definition

In an **experiment**, we apply some treatment and then proceed to observe its effects on the individuals.

Definition

The individuals in an experiment are called **subjects**.

Definition

A **Placebo** is a treatment that has no medicinal effect. (Such as a sugar pill or saline injection.)

Definition

The group that receives a placebo is called the **control group**.

Definition

The group that receives a treatment is called the **treatment group**.

Definition

Replication is the repetition of an experiment on more than one individual. This means larger sample sizes are often needed.

Definition

Blinding is used when the subject doesn't know if they are receiving a placebo or a real treatment.

Definition

Double blinding is when both the patients and the researchers are unaware of who is getting the placebo and who is getting the treatment.

Definition

Randomization is used when individuals are assigned to different groups through a process of random selection.

Example 22

In 1954, an experiment was designed to test the effectiveness of the vaccine developed by Jonas Salk in preventing polio, which had killed or paralyzed thousands of children.

By random selection $401,974$ children were assigned to two groups:

- $200,745$ children were given injections of the Salk vaccine.
- $201,229$ children were given placebo injections that contained no drug.

Among the children given the Salk vaccine, 33 later developed paralytic polio, and among the children given a placebo, 115 later developed paralytic polio.

Note

In general, causation can only be inferred from a randomized experiment.

Definition

In an **observational study**, we observe and measure specific characteristics, but we don't attempt to *modify* the subjects.

Note

Experiments are preferable to observational studies. But there may be cost, time, or ethical concerns that prohibit an experiment.

Example 23

Suppose we want to determine if listening to music while driving increases the chance of being in an collision.

- **Observational study:** We gather police reports about collisions and use them to determine if the person was listening to music or not.
- **Experiment:** We randomly assign subjects to either listen to music while driving or listen to nothing. We then count how many collisions each subject is involved in.