HOCHSCHULE
DER MEDIEN

Data Basics

Prof. Dr. Jan Kirenz
HdM Stuttgart

Observations, variables, and data matrices

Six observations from the loan50 dataset

variable

	loan_amount	interest_rate	term	grade	state	total_income	homeownership
1	22,000	10.90	60	B	NJ	59,000	rent
2	6,000	9.92	36	B	CA	60,000	rent
3	25,000	26.30	36	E	SC	75,000	mortgage
4	6,000	9.92	36	B	CA	75,000	rent
5	25,000	9.43	60	B	OH	254,000	mortgage
6	6,400	9.92	36	B	IN	67,000	mortgage

observation

value

- The data in this table represents a **data frame**
- Each **row** is a unique case (observational unit),
- Each **column** is a variable
- Each **cell** is a single value

Source: Çetinkaya-Rundel & Hardin (2021);

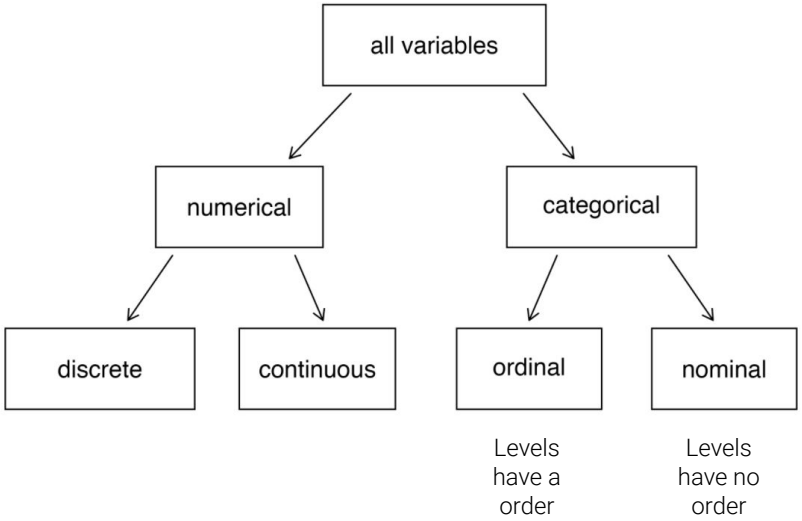
Variables descriptions for the loan50 dataset.

Variable	Description
`loan_amount`	Amount of the loan received, in US dollars.
`interest_rate`	Interest rate on the loan, in an annual percentage.
`term`	The length of the loan, which is always set as a whole number of months.
`grade`	Loan grade, which takes a 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.

Source: Çetinkaya-Rundel & Hardin (2021);

Types of variables

Breakdown of variables into their respective types



Six observations and six variables from the county dataset

name	state	pop2017	pop_change	unemployment_rate	median_edu
Autauga County	Alabama	55,504	1.48	3.86	some_college
Baldwin County	Alabama	212,628	9.19	3.99	some_college
Barbour County	Alabama	25,270	-6.22	5.90	hs_diploma
Bibb County	Alabama	22,668	0.73	4.39	hs_diploma
Blount County	Alabama	58,013	0.68	4.02	hs_diploma
Bullock County	Alabama	10,309	-2.28	4.93	hs_diploma

3,142 counties in the United States

What is the difference between the variables:

- unemployment_rate
- pop2017
- state
- median_edu

Variable	Description
`name`	Name of county.
`state`	Name of state.
`pop2000`	Population in 2000.
`pop2010`	Population in 2010.
`pop2017`	Population in 2017.
`pop_change`	Population change from 2010 to 2017 (in percent).
`poverty`	Percent of population in poverty in 2017.
`homeownership`	Homeownership rate, 2006-2010.
`multi_unit`	Multi-unit rate: percent of housing units that are in multi-unit structures, 2006-2010.
`unemployment_rate`	Unemployment rate in 2017.
`metro`	Whether the county contains a metropolitan area, taking one of the values yes or no.
`median_edu`	Median education level (2013-2017), taking one of the values below_hs, hs_diploma, some_college, or bachelors.
`per_capita_income`	Per capita (per person) income (2013-2017).
`median_hh_income`	Median household income.
`smoking_ban`	Describes the type of county-level smoking ban in place in 2010, taking one of the values none, partial, or comprehensive.

Example

Data were collected about students in a statistics course. Three variables were recorded for each student:

1. number of siblings,
2. student height, and
3. whether the student had previously taken a statistics course.

Classify each of the variables as

- continuous numerical
- discrete numerical
- nominal categorical
- ordinal categorical

What are the possible levels of the variables?

Practice - Start

Classroom survey

A survey was conducted on students in an introductory statistics course. Below are a few of the questions on the survey, and the corresponding variables the data from the responses were stored in:

- **gender**: What is your gender?
- **intro_extra**: Are you an introvert or an extrovert?
- **sleep**: How many hours do you sleep at night, on average?
- **bedtime**: What time do you usually go to bed?
- **countries**: How many countries have you visited?
- **dread**: On a scale of 1-5, how much do you dread being here?

Data matrix

Data collected on students in a statistics class on a variety of variables:

variable ↓

Stu.	gender	intro_extra	···	dread
1	male	extravert	···	3
2	female	extravert	···	2
3	female	introvert	···	4
4	female	extravert	···	2
⋮	⋮	⋮	⋮	⋮
86	male	extravert	···	3

← *observation*

Types of variables (cont.)

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- gender:

Types of variables (cont.)

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- gender: *categorical, nominal*

Types of variables (cont.)

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- gender: *categorical, nominal*
- sleep:

Types of variables (cont.)

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- gender: *categorical, nominal*
- sleep: *numerical, continuous*

Types of variables (cont.)

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- gender: *categorical, nominal*
- sleep: *numerical, continuous*
- bedtime:

Types of variables (cont.)

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- gender: *categorical, nominal*
- sleep: *numerical, continuous*
- bedtime: *categorical, ordinal*

Types of variables (cont.)

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- gender: *categorical, nominal*
- sleep: *numerical, continuous*
- bedtime: *categorical, ordinal*
- countries:

Types of variables (cont.)

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- gender: *categorical, nominal*
- sleep: *numerical, continuous*
- bedtime: *categorical, ordinal*
- countries: numerical, discrete

Types of variables (cont.)

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- gender: *categorical, nominal*
- sleep: *numerical, continuous*
- bedtime: *categorical, ordinal*
- countries: *numerical, discrete*
- dread:

Types of variables (cont.)

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- gender: *categorical, nominal*
- sleep: *numerical, continuous*
- bedtime: *categorical, ordinal*
- countries: *numerical, discrete*
- dread: *categorical, ordinal - could also be used as numerical*

Practice - End

Relationships between variables

Many analyses are motivated by looking for a relationship between two or more variables.

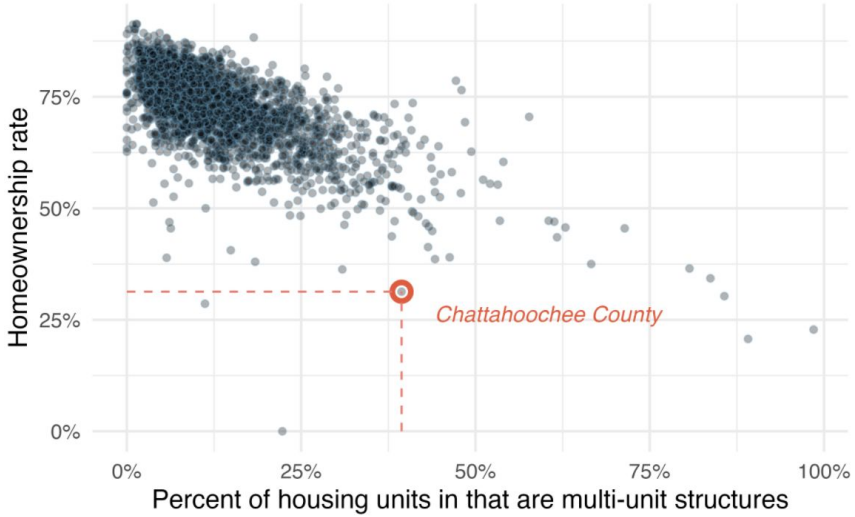
Relationships between variables

- When two variables show some connection with one another, they are called **associated variables**.
- If two variables are not associated, then they are said to be **independent**

Relationships between variables

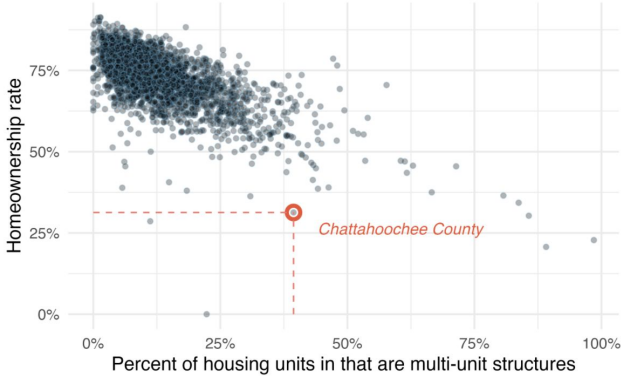
- Examining **summary statistics** (like the mean) can provide numerical insights about the specifics of each of these questions.
- **Scatterplots** are one type of graph used to study the relationship between two numerical variables

Scatterplot of US county dataset



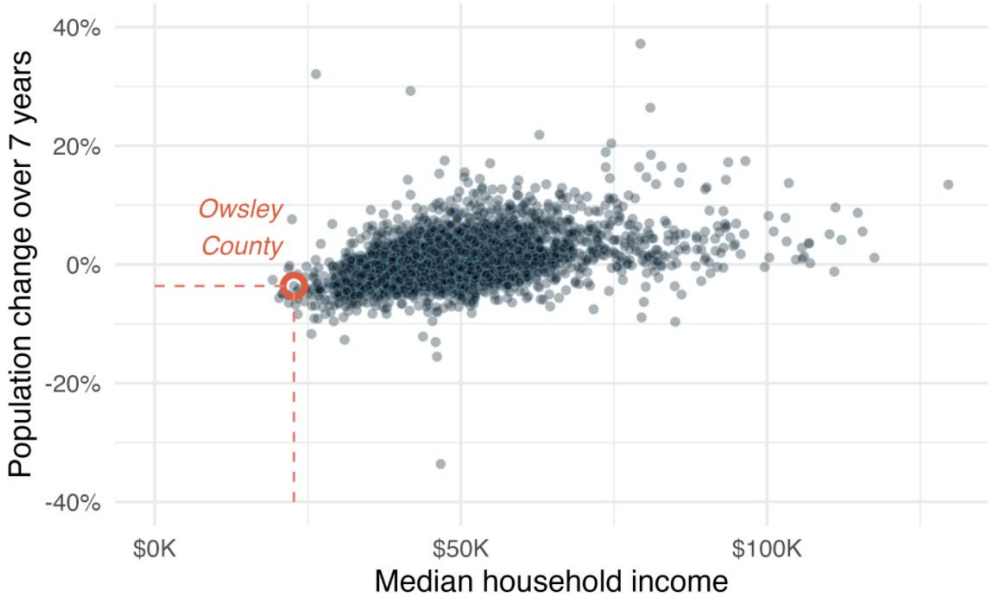
If homeownership in one county is lower than the national average, will the percent of housing units that are in multi-unit structures in that county tend to be above or below the national average?

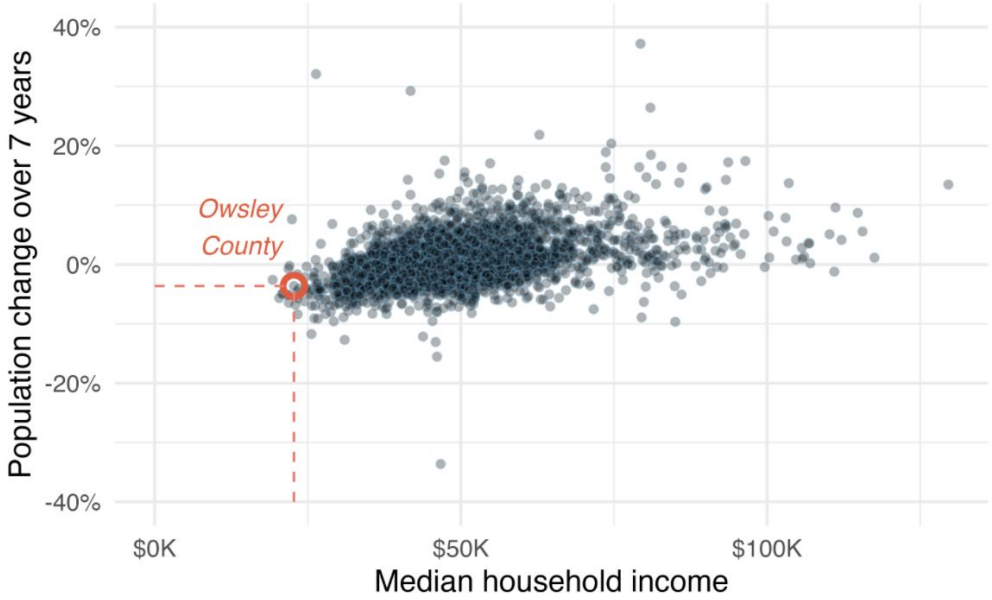
The scatterplot suggests a relationship between the two variables



Chattahoochee County, Georgia
39.4% multi-unit structures
31.3% homeownership rate

Because there is a downward trend – counties with more housing units that are in multi-unit structures are associated with lower homeownership – these variables are said to be **negatively associated**


40%
20%
0%
-20%
-40%
Population change over 7 years
Owsley
County
$0K
$50K
$100K
Median household income



A scatterplot showing population change against median household income.

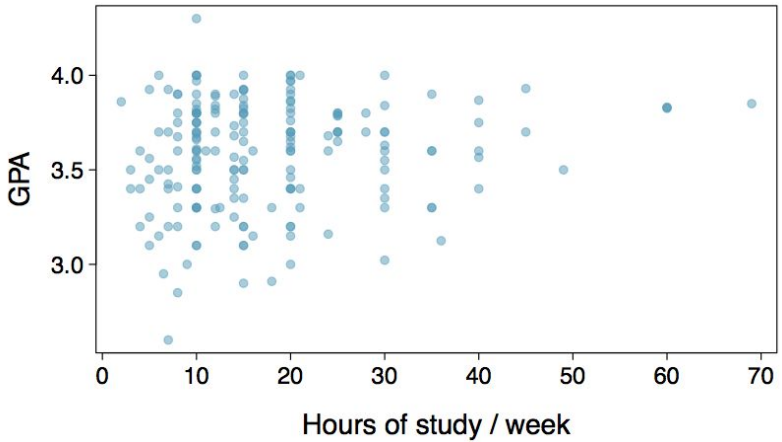
Owsley County of Kentucky, is highlighted, which lost 3.63% of its population from 2010 to 2017 and had median household income of $22,736.

A **positive association** is shown in the relationship between the median_hh_income and pop_change variables, where counties with higher median household income tend to have higher rates of population growth.

Practice - Start

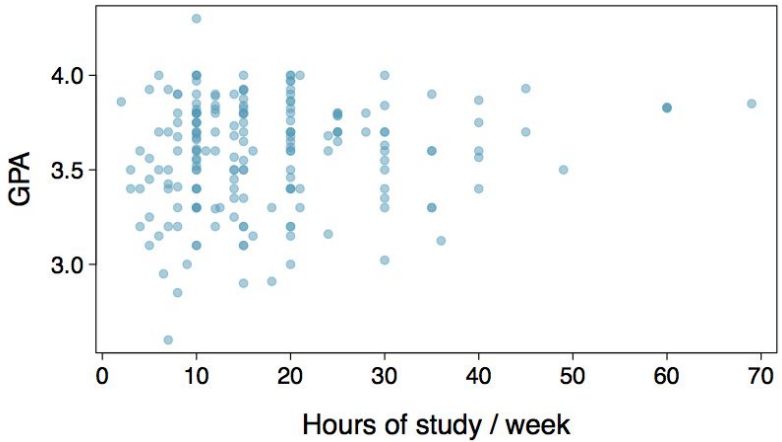
Relationships among variables

Does there appear to be a relationship between the hours of study per week and the GPA of a student?



Relationships among variables

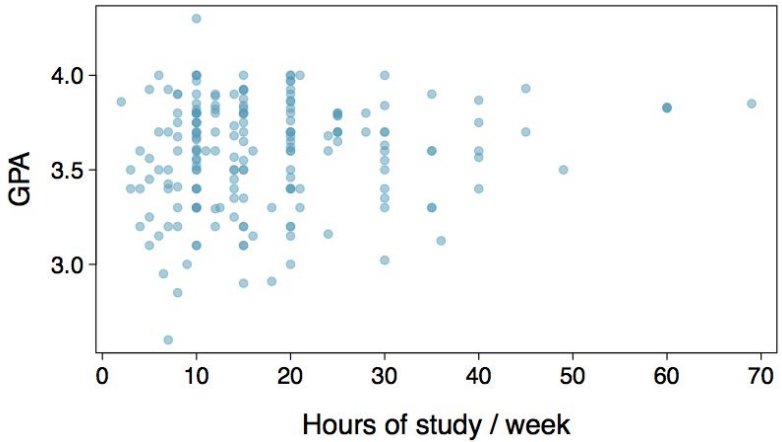
Does there appear to be a relationship between the hours of study per week and the GPA of a student?



Can you spot anything unusual about any of the data points?

Relationships among variables

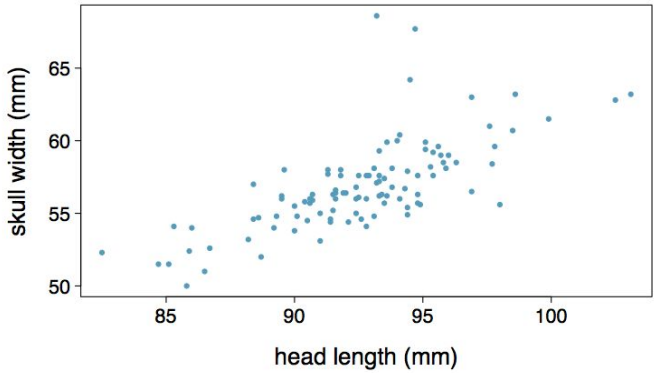
Does there appear to be a relationship between the hours of study per week and the GPA of a student?



Can you spot anything unusual about any of the data points?
There is one student with *GPA* > 4.0, this is likely a data error.

Practice

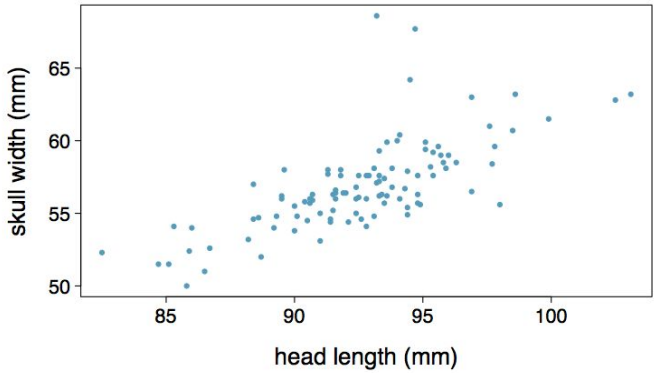
Based on the scatterplot on the right, which of the following statements is correct about the head and skull lengths of possums?



(a) There is no relationship between head length and skull width,
i.e. the variables are independent.
(b) Head length and skull width are positively associated.
(c) Skull width and head length are negatively associated.
(d) A longer head causes the skull to be wider.
(e) A wider skull causes the head to be longer.

Practice

Based on the scatterplot on the right, which of the following statements is correct about the head and skull lengths of possums?



(a) There is no relationship between head length and skull width,
i.e. the variables are independent.

(b) *Head length and skull width are positively associated.*

(c) Skull width and head length are negatively associated.

(d) A longer head causes the skull to be wider.

(e) A wider skull causes the head to be longer.

Practice - End

Explanatory and response variables

Explanatory and response variables

When we ask questions about the relationship between two variables, we sometimes also want to determine if the change in one variable causes a change in the other.

explanatory variable → might affect → response variable

Example

Consider the following question about the county dataset:

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

What is explanatory and what is the response variable?

- In this question, we are asking whether one variable affects another.
- If this is our underlying belief, then median household income is the **explanatory variable**
- and the population change is the **response variable** in the hypothesized relations

Observational studies and experiments

There are two primary types of data collection: experiments and observational studies.

Experiments

- When we want to evaluate the effect of particular traits, treatments, or conditions, we conduct an **experiment**.
- When individuals are randomly assigned to a group, the experiment is called a **randomized experiment**.

Experiments

- We typically use **placebos** (fake treatment) to ensure roughly equal conditions

Observational study

- We perform an **observational study** when we collect data in a way that does not directly interfere with how the data arise.
- We may collect information via surveys, review medical or company records
- Or follow a **cohort** of many similar individuals over time

Terms you need to know

associated
case
categorical
cohort
continuous
data
data frame
dependent
discrete
experiment
explanatory variable
independent
level
negative association
nominal
numerical
observational study
observational unit
ordinal
placebo
positive association
randomized experiment
response variable
summary statistic
variable

Resources

The slides are based on the excellent book “Introduction to Modern Statistics” by Mine Çetinkaya-Rundel and Johanna Hardin.

The online version of the book can be **accessed for free**:

https://openintro-ims.netlify.app/index.html