OpenIntro Statistics (4th Edition)

David Diez
Mine Çetinkaya-Rundel
Christopher D Barr
https://www.openintro.org/book/os/

Chapter 1. Intro to data

- Data: observations
- Statistics: study of how best to collect, analyze and draw conclusions from data

1.1 Stent study

- Classic challenge in stats: evaluating efficacy of a medical treatment
- Question: Does the use of stents reduce the risk of stroke?
- 451 at-risk patients
    - Treatment group: medical management plus stent
    - Control group: medical management only
- Patient outcome: “stroke” or “no event” at the end of a time period

Summary statistic:

- Single number summarizing a large amount of data

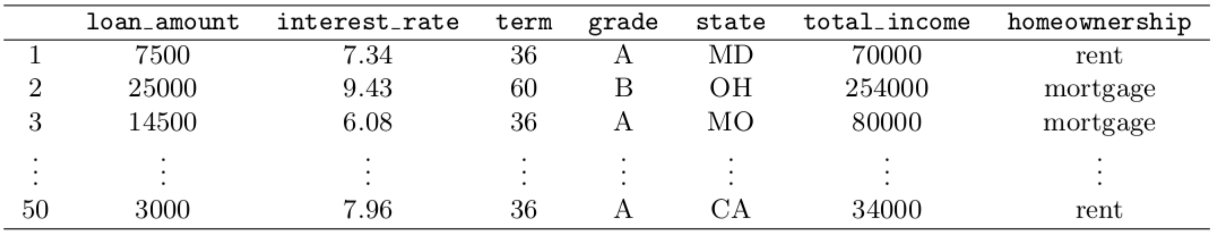
Proportion who had a stroke in the treatment (stent) group: $45/224 = 0.20 = 20\%$.

Proportion who had a stroke in the control group: $28/227 = 0.12 = 12\%$.

- Consistent with or contrary to what was expected?
- Is this a “real” difference between groups? Is the difference so large that we should reject the notion that it was due to chance?
- Statistical tests can be performed for analysis
- Be careful not to generalize results of a single study to all patients and all stents

1.2 Data basics

- Data matrix: common way to organize data



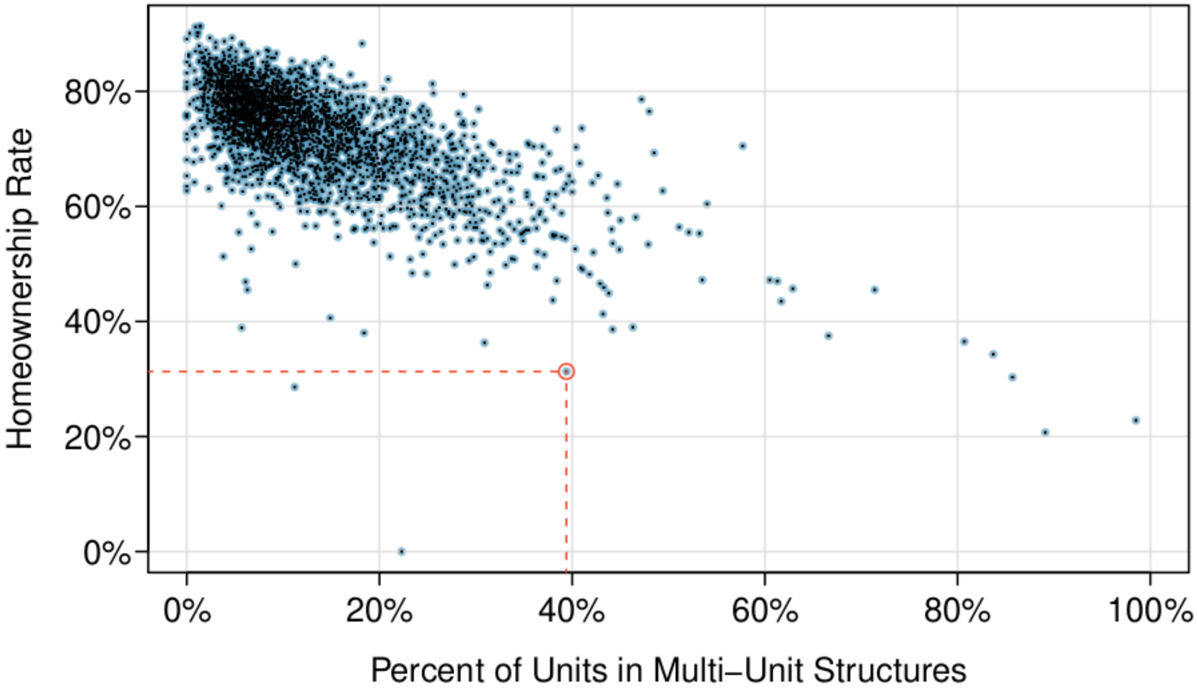
	loan_amount	interest_rate	term	grade	state	total_income	homeownership
1	7500	7.34	36	A	MD	70000	rent
2	25000	9.43	60	B	OH	254000	mortgage
3	14500	6.08	36	A	MO	80000	mortgage
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	3000	7.96	36	A	CA	34000	rent

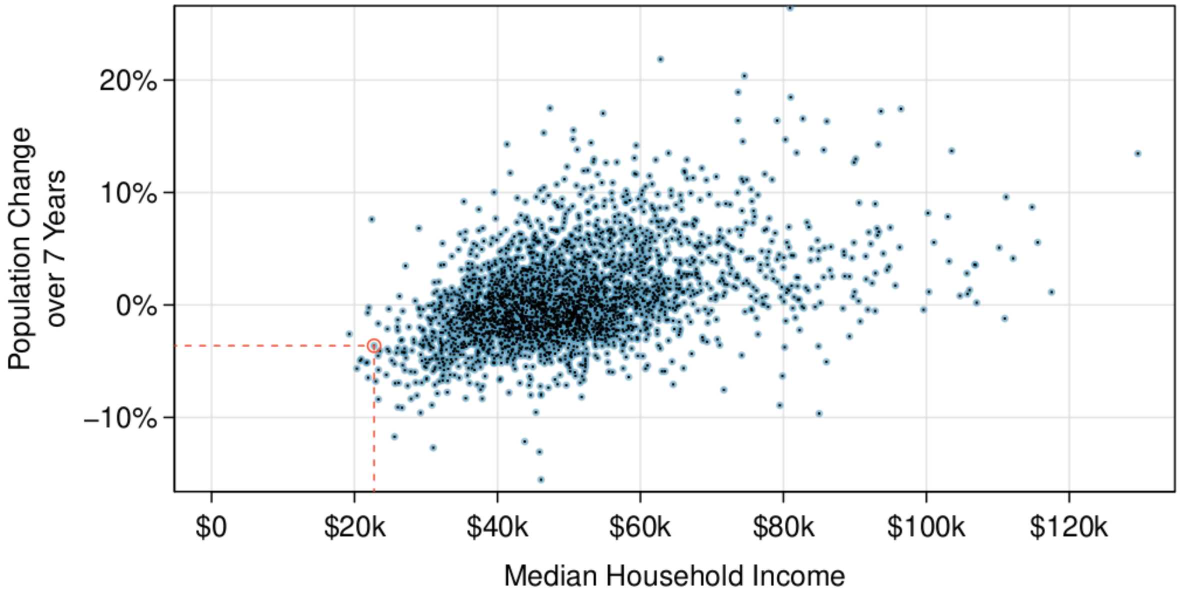
Figure 1.3: Four rows from the loan50 data matrix.

- Case / observational unit: unique case (row)
- Variables: characteristics (columns)
  - Numerical: discrete (count), continuous
  - Categorical: with levels
  - Hybrid: ordinal (with ordering) - will be treated as nominal (unordered) categorical (for simplicity)

Relationship between variables:

- Summary statistics and graphs (scatterplots) can help
- Associated: two variables show some connection (dependent)
  - Negative vs. positive association
- Independent (not associated): no evident relationship
- **Explanatory** variable may causally affect **response** variable (hypothesized relationship)
- Association IS NOT causation





Data collection methods:

- Observational study: observe data that arise
- Experiment: investigate possibility of a causal connection
  - Randomized experiment: individuals in a sample are randomly assigned to a group (treatment vs. control: drug vs. placebo)

1.3 Sampling

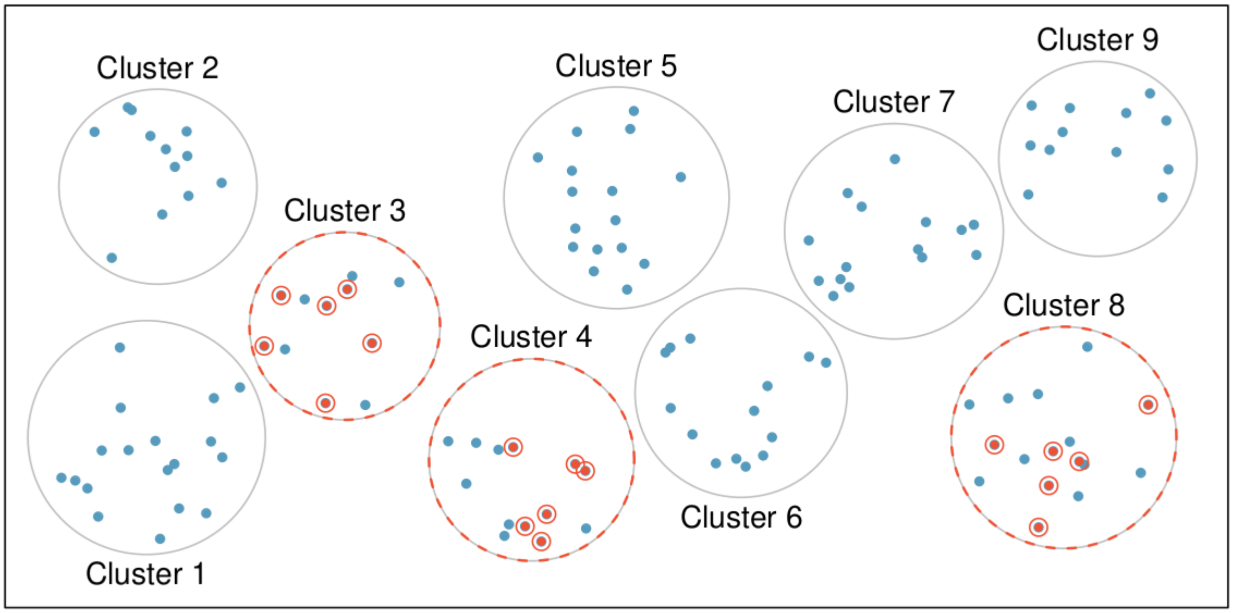
- Sample: subset of target population (hopefully representative)
- Anecdotal evidence: data collected in a haphazard fashion, may represent unusual cases
- Random selection to reduce the risk of picking a biased sample
- Common downfalls:
  - Non-response bias when non-response rate is high
  - Convenience sample (stopping people walking for survey)
- Confounding variable: correlated with both explanatory and response variables
  - Also called lurking variable, confounding factor, confounder
- Almost all statistical methods are based on notion of implied randomness

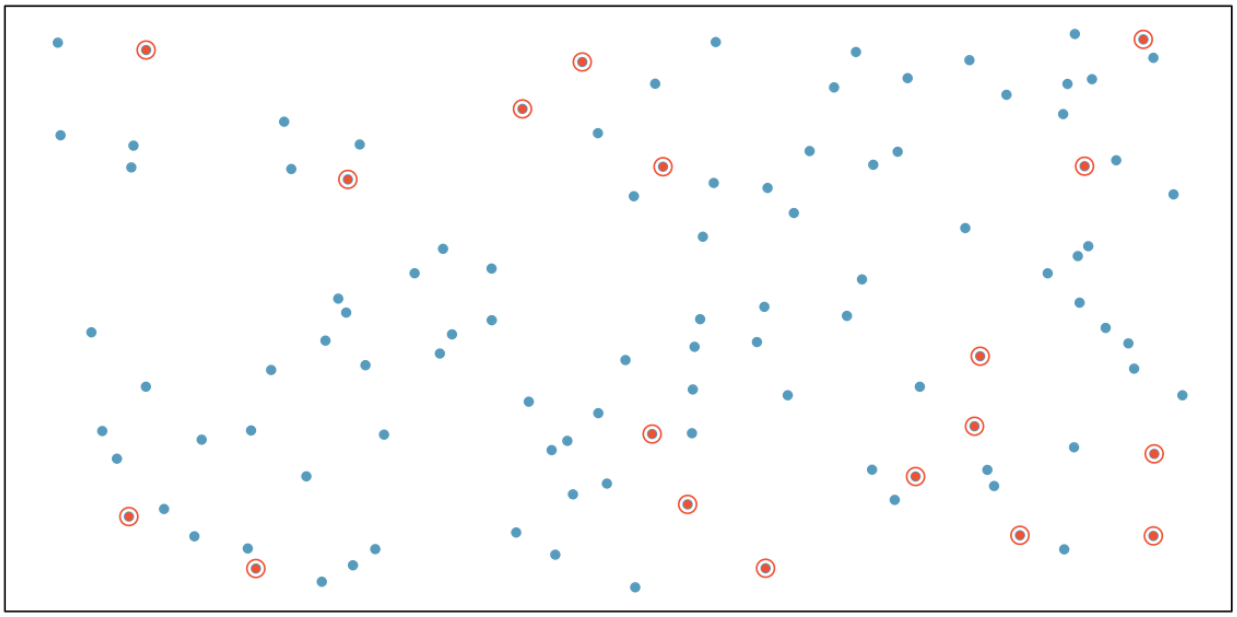
Random sampling methods

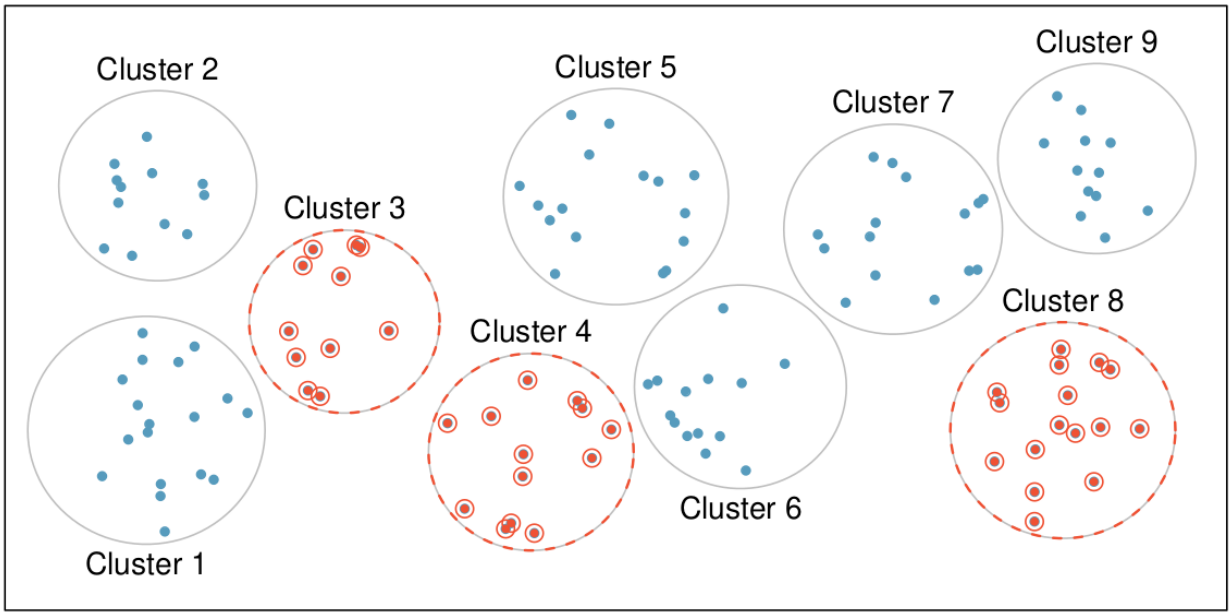
1. **Simple random sampling**: each case in the population has an equal chance of being included
2. **Stratified sampling**: population is first divided into groups (strata; similar cases are grouped together), then a second sampling method usually simple random sampling is done within each stratum

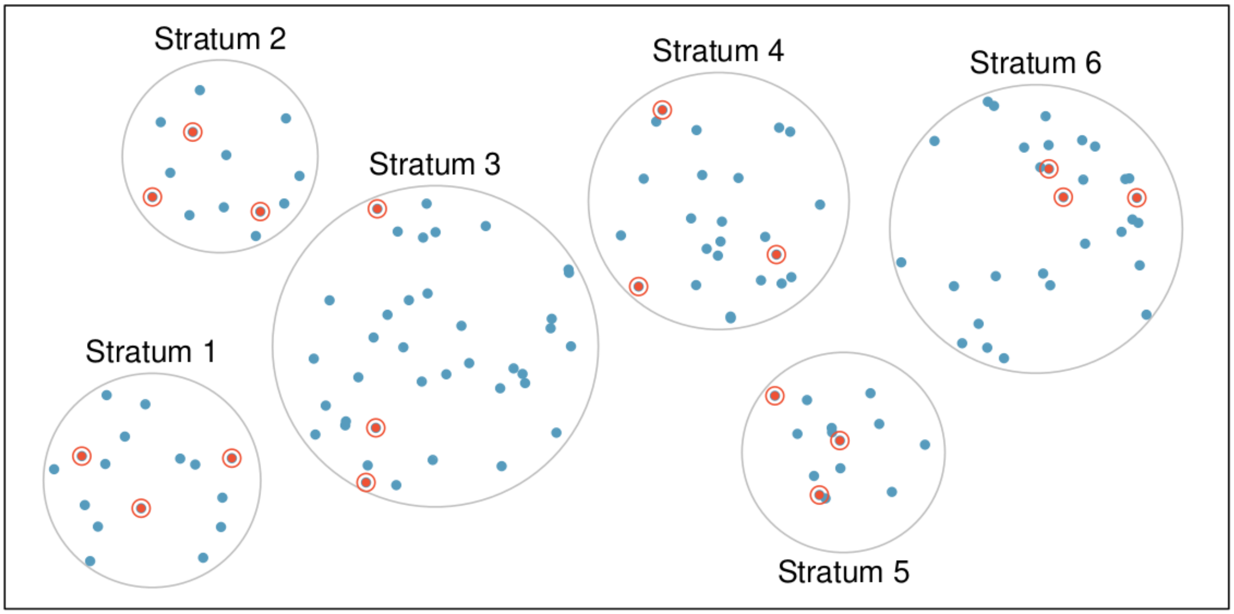
3. **Cluster sample**: population is first broken up into many groups (clusters) then a fixed number of clusters are sampled and all observations from each of the chosen clusters are included in the sample
4. **Multistage sample**: like cluster sample but a random sample within each selected cluster are collected

Advantages of 3,4 are that data collection cost can be reduced.
Disadvantages are that analysis can be more complicated.









1.4 Experiments

- Experiments: studies where researchers assign treatments to cases
- Randomized experiment: when the treatment assignment includes randomization
- Experimental design principles:
  - Controlling: researchers do their best to control any other differences in the groups
  - Randomization: of sample into treatment vs. control groups

    - Replication: the more cases the more accurately we can estimate the effect of a treatment; in a single study, this is done by collecting a sufficiently large sample; otherwise an entire study may be done to replicate an earlier finding
    - Blocking: e.g., split patients in the study into low- vs. high-risk blocks, then randomly assign patients within each block to control vs. treatment group

Reducing bias

- Blind: researchers keep the patients uninformed about their treatment
- Placebo: fake treatment given to control group e.g., sugar pill that is made to look like the actual treatment pill
- Placebo effect: a placebo results in a slight but real improvement in patients
- Double-blind: doctors or researchers who interact with patients are unaware of who is or is not receiving the treatment
- Difficult ethics of a sham surgery

Chapter exercises

1.35 Pet names.

a) Observational study.
b) Of names shared between cats and dogs, Lucy is the most common dog name. Luna is the most common cat name. There could be other more common names not on this plot.
c) Lily, Oliver.
d) Relationship between proportion of dog vs. cats in common names could be possibly positively associated. This means that certain names are very common among dogs and are also common among cats.

1.36 Stressed out, Part II.

a) Experiment.
b) This may be used to conclude a causal relationship but the study could use a better placebo.

1.37 Chia seeds and weight loss.

a) Experiment.
b) Experimental treatment: 25 g of chia seeds twice a day; control: placebo
c) Yes, blocking was used on sex of the participant.
d) ~~The description does not say whether blinding was used in the study; probably not.~~ Yes, single blind because patients were blinded.
e) It is not clear whether we can make a causal statement without knowing if there were confounding factors and concerns (how the samples were chosen, age of the subjects, etc.). While the study found no significant difference between the groups, other studies may have found a significant effect of chia seeds, so it should be further evaluated.

1.38 City council survey.

a) Randomly sample 200 households from the city.

    Simple random sampling. Pro: statistically straightforward to analyze. Con: some neighborhoods many have much more households than others and they may be unintentionally over-represented. May be expensive to survey so many.

b) Divide the city into 20 neighborhoods, and sample 10 households from each neighborhood.

    Stratified sampling (but not sure if random). Pro: each neighborhood is represented equally. Con: statistically complicated to analyze (nested stats required). May be expensive to survey so many.

c) Divide the city into 20 neighborhoods, randomly sample 3 neighborhoods, and then sample all households from those 3 neighborhoods.

    Cluster sampling. Pro: might be easier to administer surveys, if distributing in person. Con: statistically complicated to analyze. The 3 selected neighborhoods may not be representative of the whole population. If some of the neighborhoods contain a lot of households, might be expensive to survey so many.

d) Divide the city into 20 neighborhoods, randomly sample 8 neighborhoods, and then randomly sample 50 households from those neighborhoods.

    Multistage sample. Pro: many different types of neighborhoods are potentially represented. Con: still a lot of surveys to distribute, analyze (400!). Might be complicated to analyze.

e) Sample the 200 households closest to the city council offices.

    Convenience sample. Pro: convenient. Con: not random/representative of whole population.

1.39 Flawed reasoning.

a) Non-response bias. Make sure the response rate is close to 100%.
b) Convenience sample. May not be representative of the whole population; the researchers should try to track down the previously sampled population better or at least examine the change of address factor more deeply.
c) ~~Anecdotal evidence.~~ Observational study. We don’t know if those particular jogging patients have anything special about them like having excellent health and also the fact that they do not have joint problems may contribute to their regular jogging habit.

1.40 Income and education in US counties.

a) Explanatory variable: percent with bachelor’s degree; response variable: per capita income.
b) Higher the percent of the county’s population is with bachelor’s degree, the higher its per capita income. The two variables have a positive relationship. The slope of the relationship seems flatter below 40% with bachelor’s degree, and then appears to be slightly higher beyond 40% although data become sparser in the latter region.
c) No. Association does not equal causation; cannot conclude that having a bachelor’s increases one’s income but the two appear to be highly correlated at the county level.

1.41 Eat better, feel better?

a) Experiment (Randomized controlled experiment).
b) Explanatory variable: treatment group (categorical with 3 levels); Response variable: psychological well-being score on a nightly survey.
c) Results of the study may be generalizable to University of Otago students, if the sample did not suffer from convenience sampling or non-response bias. More research is needed to see if they generalize to other populations. **No, because participants were volunteers.**
d) Results of the study may be used to establish causal relationships, given that the study had sufficiently large sample sizes and care was taken to minimize possibility for bias or confounding factors. **Yes, because it was an experiment.**
e) The results of this study provide **evidence** that… (rest is fine).

1.42 Screens, teens, and psychological well-being.

a) Observational study.
b) Explanatory variable: screen time.
c) Response variable: psychological well-being score.
d) Results may or may not be generalizable to the population, depending on how study participants were recruited. Were they randomly sampled? If they were, because it was a "nationally representative large-scale data" from three countries, this may well be representative to the population.
e) No, because a controlled experiment was not done. But results may show something about the relationship between variables.

1.43 Stanford Open Policing.

a) Variables collected on each traffic stop: county, state, driver's race, year, whether car was stopped, whether car was searched, whether driver was arrested.
b) Categorical variables: county, state, driver's race, whether car was stopped, searched, driver arrested (whether: boolean) - all are not ordinal; Numerical variable: year (discrete).
c) Explanatory variable: driver's race; Response variable: vehicle search rate (**whether the car was searched or not**).

1.44 Space launches.

a) Variables collected on each launch: whether it was private/state/startup, year, failure/success.
b) Categorical: whether private/state/startup (not ordinal), failure/success (not ordinal); Numerical: year (discrete).
c) Explanatory variable: launching agency; Response variable: success rate.