

Lecture 1

Chapter 1: Introduction to data

OpenIntro Statistics, 4th Edition

Slides developed by Mine Çetinkaya-Rundel of OpenIntro.

The slides may be copied, edited, and/or shared via the CC BY-SA license.

Some images may be included under fair use guidelines (educational purposes).

Modifications by Abe Kazemzadeh, available on Github fork.

Course Objectives

- Introduction to statistics
 - Experiment design
 - Probability
 - Distributions of random variables
 - Statistical tests
 - Regression models
- Introduction to R
 - Basics of the language
 - Notebook-like usage
 - Manipulating and viewing data
 - External tools (e.g. Markdown, L^AT_EX, Github)

Introduction

Why I Like This Class

- Statistics basics are good to refresh
 - There are subtle nuances
 - It's very useful/relevant to many fields
 - Statistics is one perspective on how to find the truth, a formalization of the scientific method
- R is a fun programming language
 - We'll be using it for specific purposes (statistics) so you only need to learn the relevant parts.
 - RStudio helps a lot in learning R
 - One tricky part: base R vs. Tidyverse
- Online classes are relatively new and have a lot of promise

Approach: Leverage Book, Instructor, and Class

The book gives a good overview of statistics, many examples, and problem sets.

The professor will help interpret the book, create R tutorials, and connect the material to other classes, history, and professional experience in programming, data science, and machine learning.

The students will contribute to the class by asking questions, discussing, sharing relevant experience, and by doing a final project.

3

Policies and Schedule

See syllabus and Canvas.

4

Introductions

- What is your name (including any nicknames or pronunciation help)?
- What program are you in?
- Any experience with R or stats?
- Places: where you were born, where you grew up, where you currently live, and one place you'd like to visit?

5

Case study

Treating Chronic Fatigue Syndrome

- Objective: Evaluate the effectiveness of cognitive-behavior therapy for chronic fatigue syndrome.
- Participant pool: 142 patients who were recruited from referrals by primary care physicians and consultants to a hospital clinic specializing in chronic fatigue syndrome.
- Actual participants: Only 60 of the 142 referred patients entered the study. Some were excluded because they didn't meet the diagnostic criteria, some had other health issues, and some refused to be a part of the study.

Deale et. al. *Cognitive behavior therapy for chronic fatigue syndrome: A randomized controlled trial*. The American Journal of Psychiatry 154.3 (1997).

6

Results

The table below shows the distribution of patients with good outcomes at 6-month follow-up. Note that 7 patients dropped out of the study: 3 from the treatment and 4 from the control group.

Group	Good outcome		
	Yes	No	Total
Treatment	19	8	27
Control	5	21	26
Total	24	29	53

- Proportion with good outcomes in treatment group:

$$19/27 \approx 0.70 \rightarrow 70\%$$

- Proportion with good outcomes in control group:

$$5/26 \approx 0.19 \rightarrow 19\%$$

Study design

- Patients randomly assigned to treatment and control groups, 30 patients in each group:
 - *Treatment*: Cognitive behavior therapy – collaborative, educative, and with a behavioral emphasis. Patients were shown on how activity could be increased steadily and safely without exacerbating symptoms.
 - *Control*: Relaxation – No advice was given about how activity could be increased. Instead progressive muscle relaxation, visualization, and rapid relaxation skills were taught.

7

Understanding the results

Do the data show a “real” difference between the groups?

- Suppose you flip a coin 100 times. While the chance a coin lands heads in any given coin flip is 50%, we probably won't observe exactly 50 heads. This type of fluctuation is part of almost any type of data generating process.
- The observed difference between the two groups ($70 - 19 = 51\%$) may be real, or may be due to natural variation.
- Since the difference is quite large, it is more believable that the difference is real.
- We need statistical tools to determine if the difference is so large that we should reject the notion that it was due to chance.

8

9

Generalizing the results

Are the results of this study generalizable to all patients with chronic fatigue syndrome?

These patients had specific characteristics and volunteered to be a part of this study, therefore they may not be representative of all patients with chronic fatigue syndrome. While we cannot immediately generalize the results to all patients, this first study is encouraging. The method works for patients with some narrow set of characteristics, and that gives hope that it will work, at least to some degree, with other patients.

10

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: Do you consider yourself introverted or extraverted?
- 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?

11

Data basics

Data matrix

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

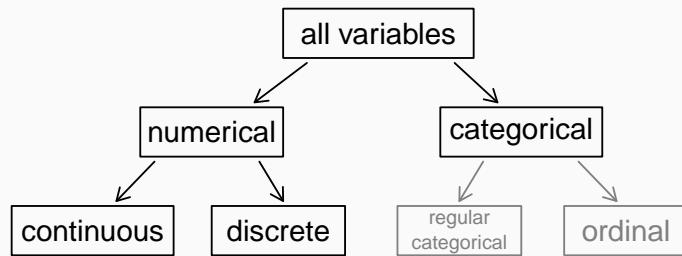
variable
↓

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

12

Types of variables



13

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:
- sleep:
- bedtime:
- countries:
- dread:

14

Practice

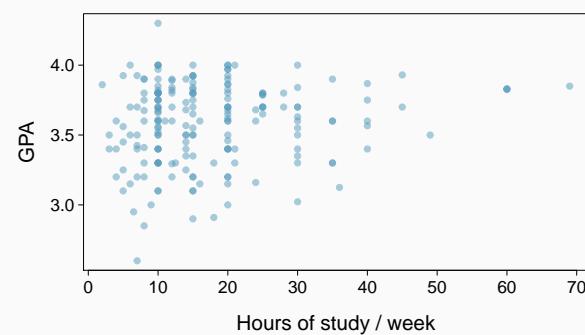
What type of variable is a telephone area code?

- (a) numerical, continuous
- (b) numerical, discrete
- (c) categorical
- (d) categorical, ordinal

15

Relationships among variables

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



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

16

Explanatory and response variables

- To identify the explanatory variable in a pair of variables, identify which of the two is suspected of affecting the other:

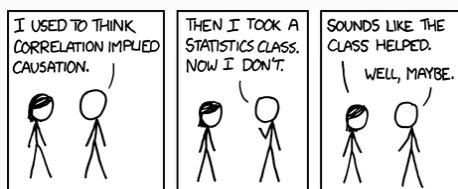
explanatory variable $\xrightarrow{\text{might affect}}$ response variable

- Labeling variables as explanatory and response does not guarantee the relationship between the two is actually causal, even if there is an association identified between the two variables. We use these labels only to keep track of which variable we suspect affects the other.

17

Association vs. causation

- When two variables show some connection with one another, they are called *associated* variables.
 - Associated variables can also be called *dependent* variables and vice-versa.
- If two variables are not associated, i.e. there is no evident connection between the two, then they are said to be *independent*.
- In general, association does not imply causation, and causation can only be inferred from a randomized experiment.



<http://xkcd.com/552/>

19

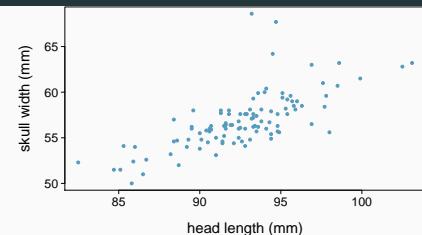
Two primary types of data collection

- Observational studies:** Collect data in a way that does not directly interfere with how the data arise (e.g. surveys).
 - Can provide evidence of a naturally occurring association between variables, but they cannot by themselves show a causal connection.
- Experiment:** Researchers randomly assign subjects to various treatments in order to establish causal connections between the explanatory and response variables.

18

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.

20

Sampling principles and strategies

Anecdotal evidence and early smoking research

- Anti-smoking research started in the 1930s and 1940s when cigarette smoking became increasingly popular. While some smokers seemed to be sensitive to cigarette smoke, others were completely unaffected.
- Anti-smoking research was faced with resistance based on *anecdotal evidence* such as “My uncle smokes three packs a day and he’s in perfectly good health”, evidence based on a limited sample size that might not be representative of the population.
- It was concluded that “smoking is a complex human behavior, by its nature difficult to study, confounded by human variability.”
- In time researchers were able to examine larger samples of cases (smokers), and trends showing that smoking has negative health impacts became much clearer.

22

Populations and samples

PHYS ED | AUGUST 29, 2012, 12:01 AM 21 Comments
Finding Your Ideal Running Form
By GRETCHEN REYNOLDS



<http://well.blogs.nytimes.com/2012/08/29/>

finding-your-ideal-running-form

Research question: Can people become better, more efficient runners on their own, merely by running?

Population of interest:

Sample: Group of adult women who recently joined a running group

Population to which results can be generalized:

21

Census

- Wouldn’t it be better to just include everyone and “sample” the entire population?
 - This is called a *census*.
- There are problems with taking a census:
 - It can be difficult to complete a census: there always seem to be some individuals who are hard to locate or hard to measure. *And these difficult-to-find people may have certain characteristics that distinguish them from the rest of the population.*
 - Populations rarely stand still. Even if you could take a census, the population changes constantly, so it’s never possible to get a perfect measure.
 - Taking a census may be more complex than sampling.

23

Illegal Immigrants Reluctant To Fill Out Census Form

by PETER O'DOWD

March 31, 2010 4:00 AM

from KJZZ

Listen to the Story  Morning Edition 3 min 48 sec + Playlist 

There is an effort underway to make sure Hispanics are accurately counted in the 2010 Census. Phoenix has some of the country's "hardest-to-count" districts. Some Latinos, especially illegal residents, fear that participating in the count will expose them to immigration raids or government harassment.

<http://www.npr.org/templates/story/story.php?storyId=125380052>

24

Sampling bias

- **Non-response:** If only a small fraction of the randomly sampled people choose to respond to a survey, the sample may no longer be representative of the population.
- **Voluntary response:** Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the issue. Such a sample will also not be representative of the population.

Quick vote

Do you get paid sick days at your job?

Yes No

What job?

Quick vote

Do you get paid sick days at your job?

Read Related Articles

Response	Percentage	Total Votes
Yes	63%	20056
No	21%	6816
What job?	15%	4885

Total votes: 31757
This is not a scientific poll

cnn.com, Jan 14, 2012

- **Convenience sample:** Individuals who are easily accessible are more likely to be included in the sample.

Exploratory analysis to inference

- Sampling is natural.
- Think about sampling something you are cooking - you taste (examine) a small part of what you're cooking to get an idea about the dish as a whole.
- When you taste a spoonful of soup and decide the spoonful you tasted isn't salty enough, that's *exploratory analysis*.
- If you generalize and conclude that your entire soup needs salt, that's an *inference*.
- For your inference to be valid, the spoonful you tasted (the sample) needs to be *representative* of the entire pot (the population).
 - If your spoonful comes only from the surface and the salt is collected at the bottom of the pot, what you tasted is probably not representative of the whole pot.
 - If you first stir the soup thoroughly before you taste, your spoonful will more likely be representative of the whole pot.

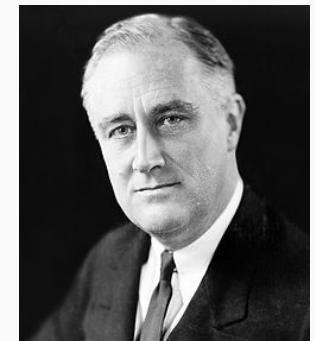
25

Sampling bias example: Landon vs. FDR

A historical example of a biased sample yielding misleading results:



In 1936, Landon sought the Republican presidential nomination opposing the re-election of FDR.

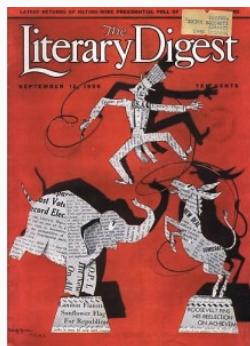


26

27

The Literary Digest Poll

- The Literary Digest polled about 10 million Americans, and got responses from about 2.4 million.
- The poll showed that Landon would likely be the overwhelming winner and FDR would get only 43% of the votes.
- Election result: FDR won, with 62% of the votes.
- The magazine was completely discredited because of the poll, and was soon discontinued.



28

The Literary Digest Poll – what went wrong?

- The magazine had surveyed
 - its own readers,
 - registered automobile owners, and
 - registered telephone users.
- These groups had incomes well above the national average of the day (remember, this is Great Depression era) which resulted in lists of voters far more likely to support Republicans than a truly *typical* voter of the time, i.e. the sample was not representative of the American population at the time.

29

Large samples are preferable, but...

- The Literary Digest election poll was based on a sample size of 2.4 million, which is huge, but since the sample was *biased*, the sample did not yield an accurate prediction.
- Back to the soup analogy: If the soup is not well stirred, it doesn't matter how large a spoon you have, it will still not taste right. If the soup is well stirred, a small spoon will suffice to test the soup.

30

Practice

A school district is considering whether it will no longer allow high school students to park at school after two recent accidents where students were severely injured. As a first step, they survey parents by mail, asking them whether or not the parents would object to this policy change. Of 6,000 surveys that go out, 1,200 are returned. Of these 1,200 surveys that were completed, 960 agreed with the policy change and 240 disagreed. Which of the following statements are true?

- I. Some of the mailings may have never reached the parents.
 - II. The school district has strong support from parents to move forward with the policy approval.
 - III. It is possible that majority of the parents of high school students disagree with the policy change.
 - IV. The survey results are unlikely to be biased because all parents were mailed a survey.
- (a) Only I (b) I and II (c) I and III (d) III and IV (e) Only IV

31

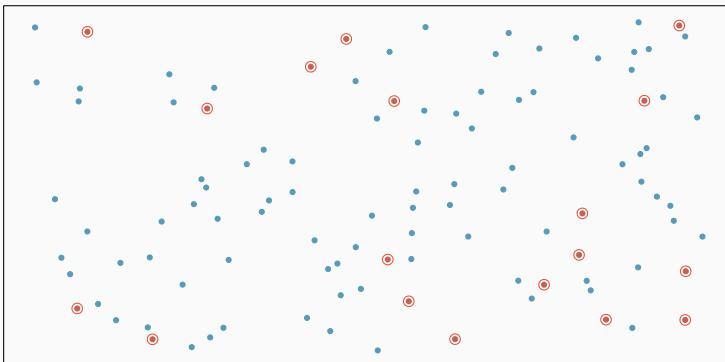
Observational studies

- Researchers collect data in a way that does not directly interfere with how the data arise.
- Results of an observational study can generally be used to establish an association between the explanatory and response variables.

32

Simple random sample

Randomly select cases from the population, where there is no implied connection between the points that are selected.



34

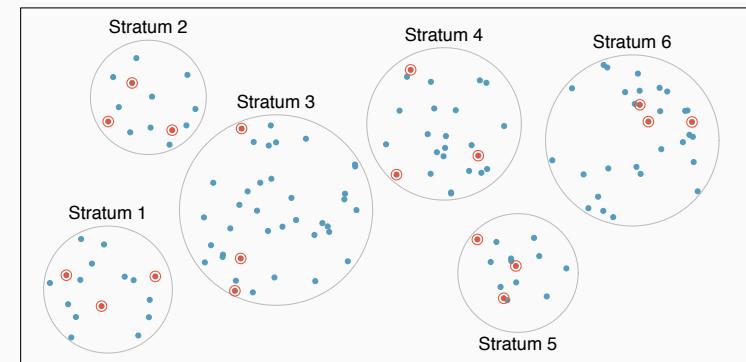
Obtaining good samples

- Almost all statistical methods are based on the notion of implied randomness.
- If observational data are not collected in a random framework from a population, these statistical methods – the estimates and errors associated with the estimates – are not reliable.
- Most commonly used random sampling techniques are *simple*, *stratified*, and *cluster* sampling.

33

Stratified sample

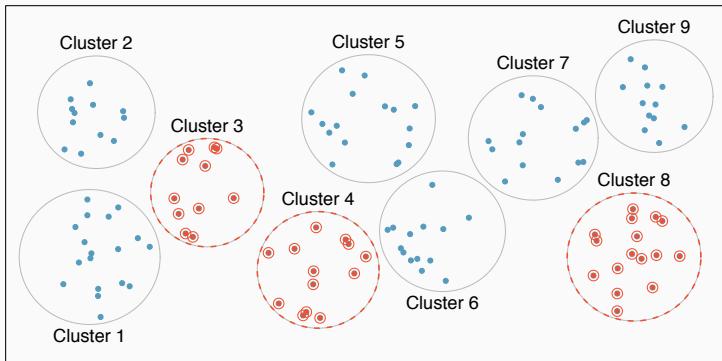
Strata are made up of similar observations. We take a simple random sample from each stratum.



35

Cluster sample

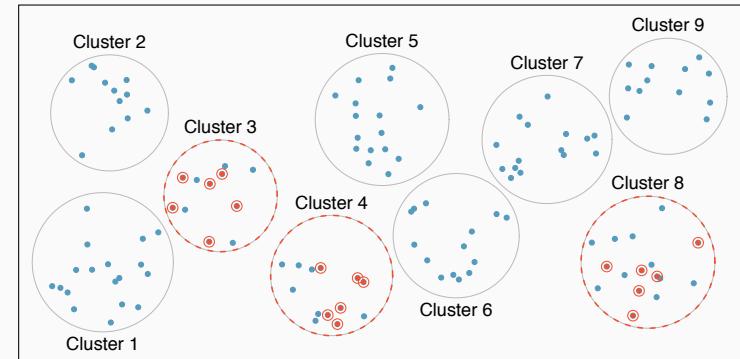
Clusters are usually not made up of homogeneous observations. We take a simple random sample of clusters, and then sample all observations in that cluster. Usually preferred for economical reasons.



36

Multistage sample

Clusters are usually not made up of homogeneous observations. We take a simple random sample of clusters, and then take a simple random sample of observations from the sampled clusters.



37

Practice

A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes, some with only apartments. Which approach would likely be the least effective?

- (a) Simple random sampling
- (b) Cluster sampling
- (c) Stratified sampling
- (d) Blocked sampling

38

Experiments

Principles of experimental design

1. **Control:** Control for the (potential) effect of variables other than the ones directly being studied.
2. **Randomize:** Randomly assign subjects to treatments, and randomly sample from the population whenever possible.
3. **Replicate:** Within a study, replicate by collecting a sufficiently large sample. Or replicate the entire study.
4. **Block:** If there are variables that are known or suspected to affect the response variable, first group subjects into *blocks* based on these variables, and then randomize cases within each block to treatment groups.

39

Practice

A study is designed to test the effect of light level and noise level on exam performance of students. The researcher also believes that light and noise levels might have different effects on males and females, so wants to make sure both genders are equally represented in each group. Which of the below is correct?

- (a) There are 3 explanatory variables (light, noise, gender) and 1 response variable (exam performance)
- (b) There are 2 explanatory variables (light and noise), 1 blocking variable (gender), and 1 response variable (exam performance)
- (c) There is 1 explanatory variable (gender) and 3 response variables (light, noise, exam performance)
- (d) There are 2 blocking variables (light and noise), 1 explanatory variable (gender), and 1 response variable (exam performance)

41

More on blocking



- We would like to design an experiment to investigate if energy gels makes you run faster:
 - Treatment: energy gel
 - Control: no energy gel
- It is suspected that energy gels might affect pro and amateur athletes differently, therefore we block for pro status:
 - Divide the sample to pro and amateur
 - Randomly assign pro athletes to treatment and control groups
 - Randomly assign amateur athletes to treatment and control groups
 - Pro/amateur status is equally represented in the resulting treatment and control groups

40

Why is this important? Can you think of other variables to block for?

Difference between blocking and explanatory variables

- Factors are conditions we can impose on the experimental units.
- Blocking variables are characteristics that the experimental units come with, that we would like to control for.
- Blocking is like stratifying, except used in experimental settings when randomly assigning, as opposed to when sampling.

42

More experimental design terminology...

- **Placebo**: fake treatment, often used as the control group for medical studies
- **Placebo effect**: experimental units showing improvement simply because they believe they are receiving a special treatment
- **Blinding**: when experimental units do not know whether they are in the control or treatment group
- **Double-blind**: when both the experimental units and the researchers who interact with the patients do not know who is in the control and who is in the treatment group

43

Practice

What is the main difference between observational studies and experiments?

- Experiments take place in a lab while observational studies do not need to.
- In an observational study we only look at what happened in the past.
- Most experiments use random assignment while observational studies do not.
- Observational studies are completely useless since no causal inference can be made based on their findings.

44

Random assignment vs. random sampling

ideal experiment	Random assignment	No random assignment	most observational studies
Random sampling	Causal conclusion, generalized to the whole population.	No causal conclusion, correlation statement generalized to the whole population.	Generalizability
No random sampling	Causal conclusion, only for the sample.	No causal conclusion, correlation statement only for the sample.	No generalizability
most experiments	Causation	Correlation	bad observational studies

45