Ch1. Introduction to Data

1.1 Case Study

1.2 Data Basics

1.3 Sampling principles and strategies

1.1 Case Study:

Treating Chronic Fatigue Syndrome

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. 1997. *Cognitive behavior therapy for chronic fatigue syndrome: A randomized controlled trial.* The American Journal of Psychiatry 154:3.

Study design

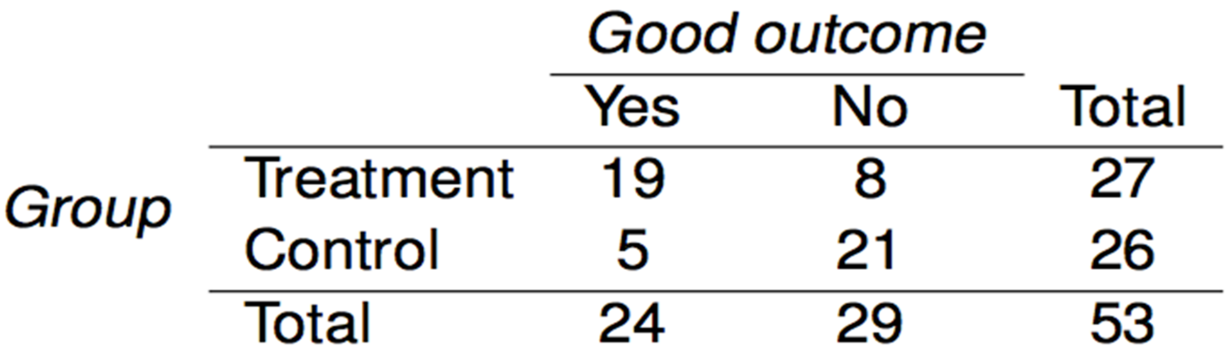
Patients randomly assigned to treatment and control groups, 30 patients in each group:

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.

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.



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

- Proportion with good outcomes in treatment group

  19/27 ≈ 0.70 → 70%

- Proportion with good outcomes in control group

  5/26 ≈ 0.19 → 19%

Do the data show a "real" difference between the groups?

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 use statistical tools to determine if the difference is so large that we should reject the notion that it was due to chance.

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

1.2 Data Basics

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:

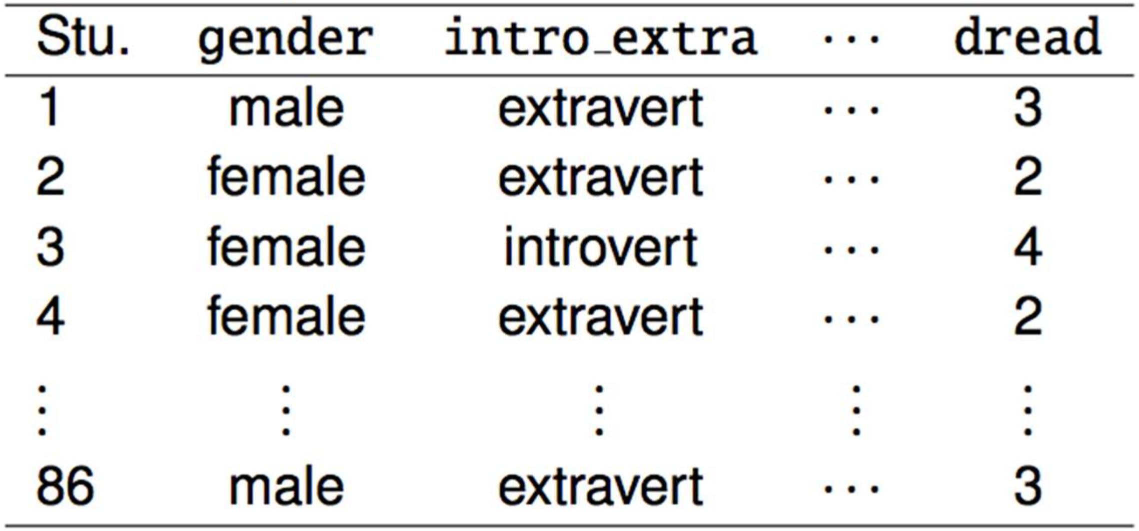
- **gender**: What is your gender?
- **intro_extra**: Introvert or extravert?
- **sleep**: How many hours do you sleep at night, on average?
- **bedtime**: What time do you usually go to bed?
- **countries**: How many countries have you visited?
- **dread**: On a scale of 1-5, how much do you dread being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

Data matrix

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

variable

↓

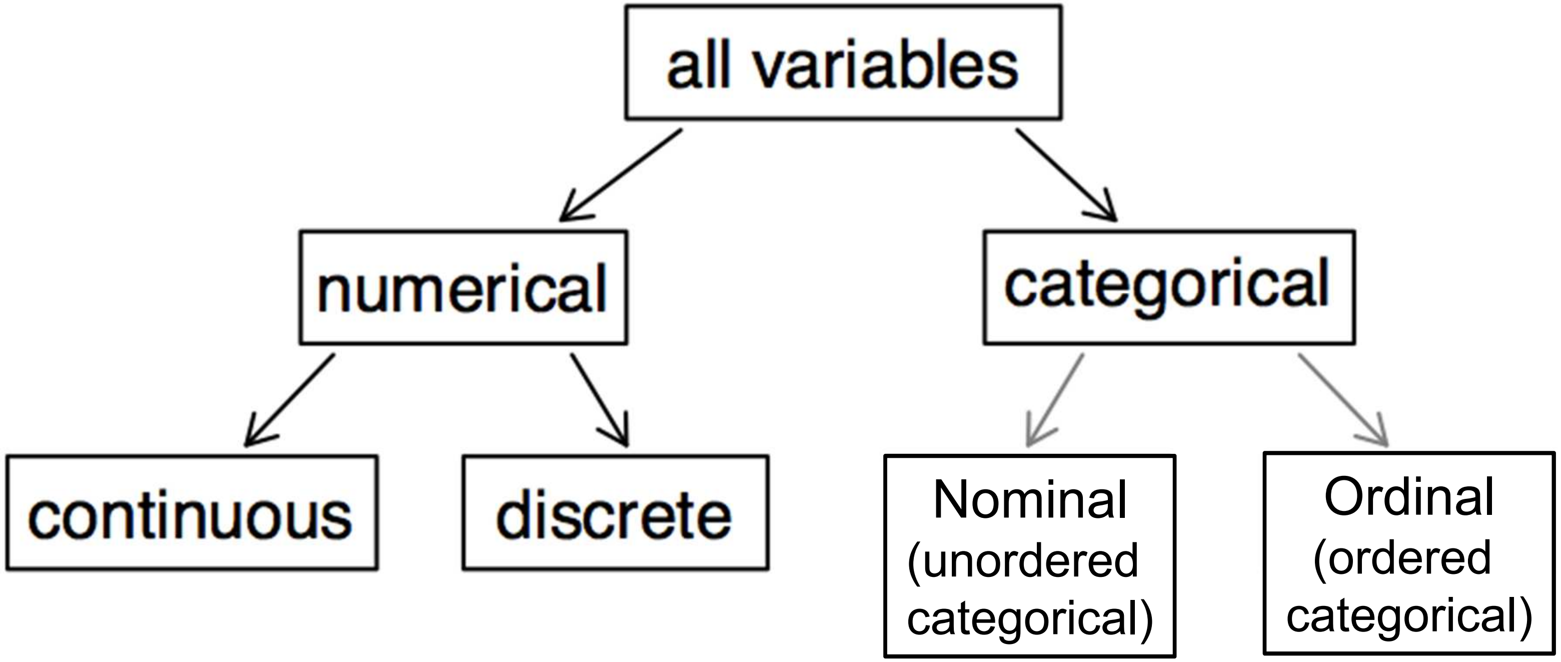


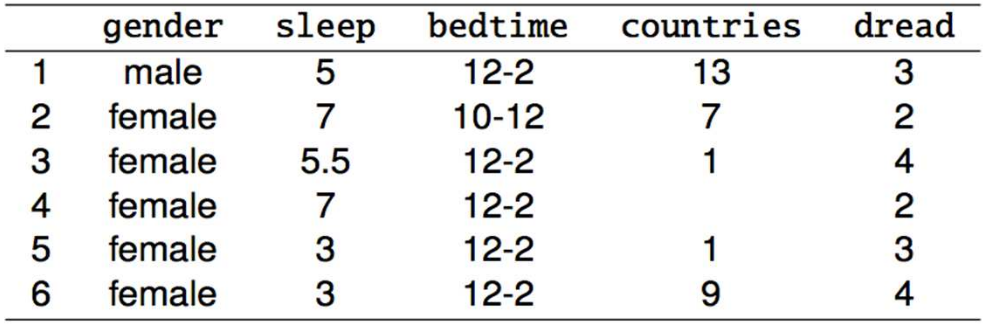
Stu.	gender	intro_extra	···	dread
1	male	extravert	···	3
2	female	extravert	···	2
3	female	introvert	···	4
4	female	extravert	···	2
⋮	⋮	⋮	⋮	⋮
86	male	extravert	···	3

←

observation

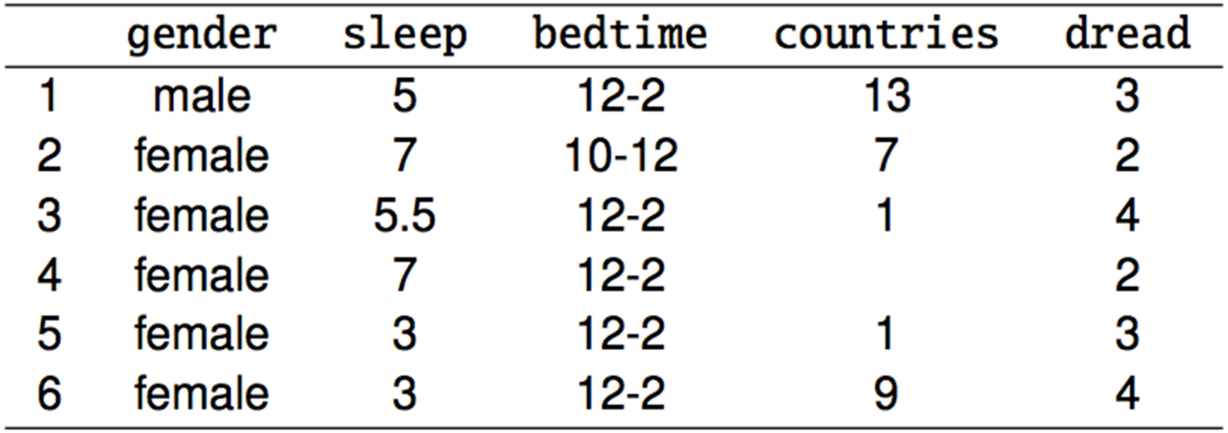
Types of variables


all variables
numerical
categorical
continuous
discrete
Nominal (unordered categorical)
Ordinal (ordered categorical)



	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

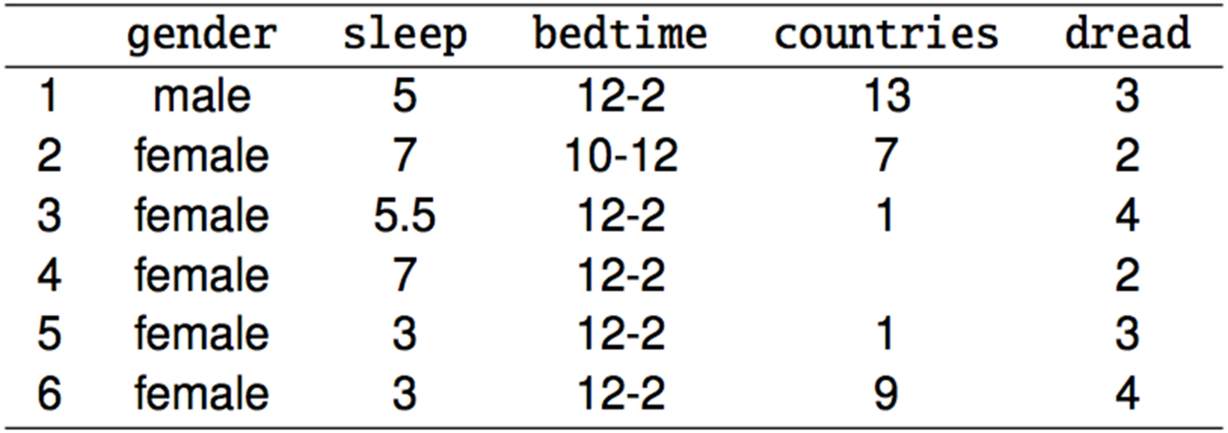
Types of variables (cont.)



	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:

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

Practice

What type of variable is a telephone area code?

(a) numerical, continuous

(b) numerical, discrete

(c) categorical

(d) categorical, ordinal

Relationships among variables

What is the relationship between two or more variables?

For example,

(1) If *homeownership* is lower than the national average in one country, will the percent of *multi-unit structures* in that country tend to be above or below the national average?

(2) Does a higher than average increase in county *population* tend to correspond to countries with higher or lower median *household incomes*?

Relationships among variables(cont.)

Scatterplot

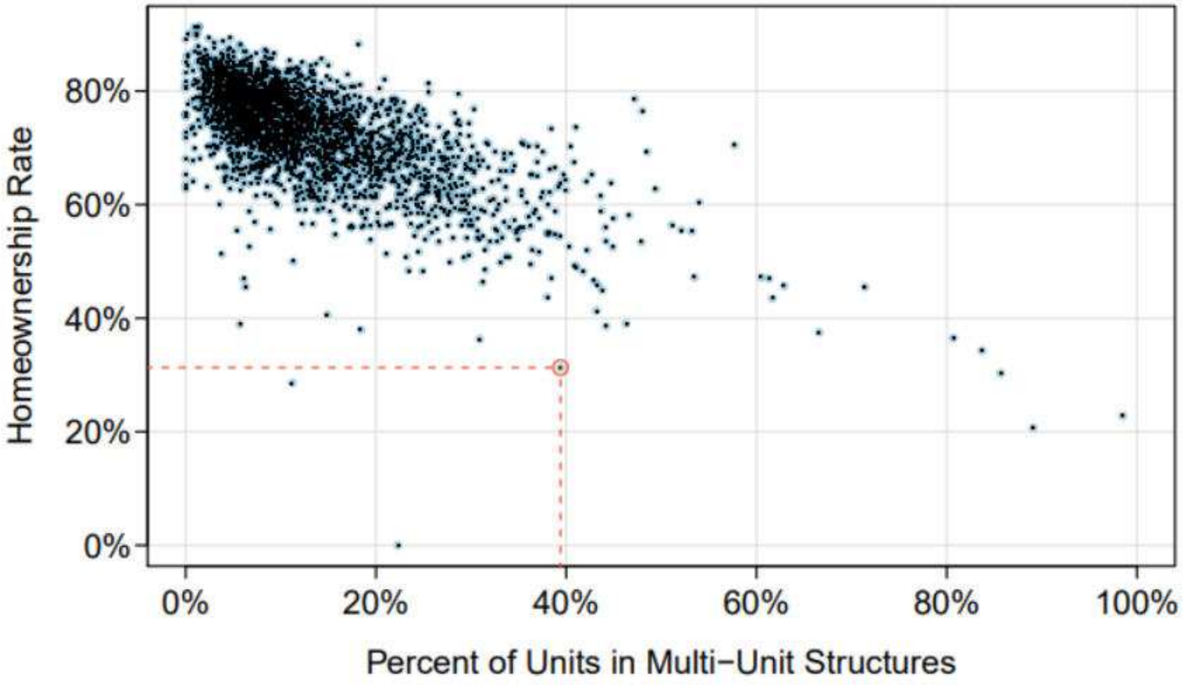


Figure 1.8: A scatterplot of homeownership versus the percent of units that are in multi-unit structures for US counties. The highlighted dot represents Chattahoochee County, Georgia, which has a multi-unit rate of 39.4% and a homeownership rate of 31.3%.

Relationships among variables(cont.)

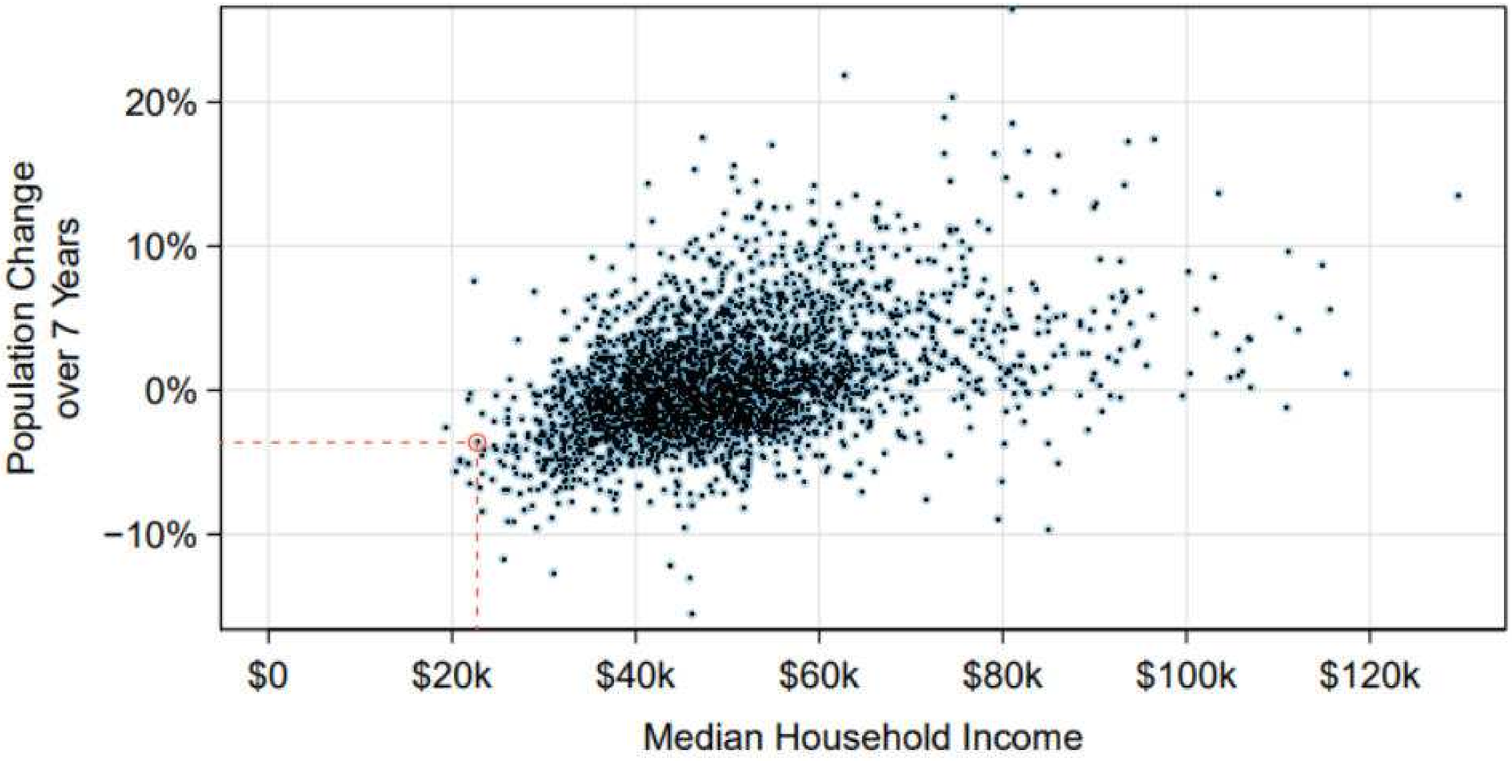
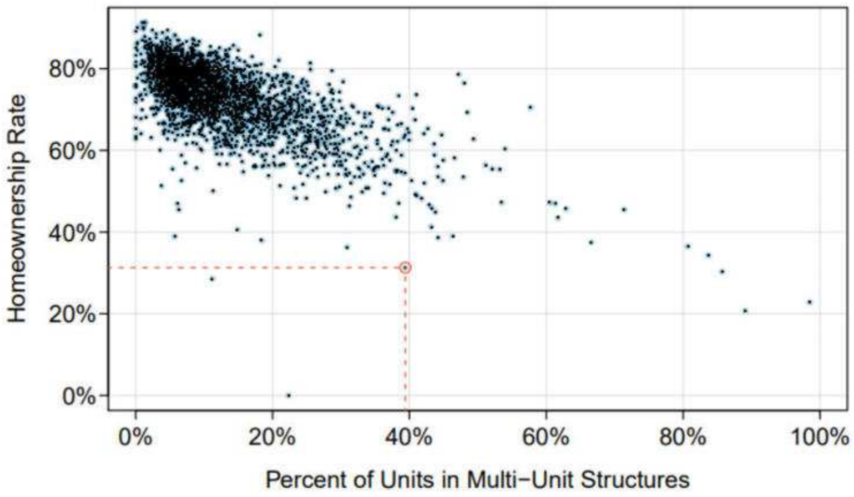
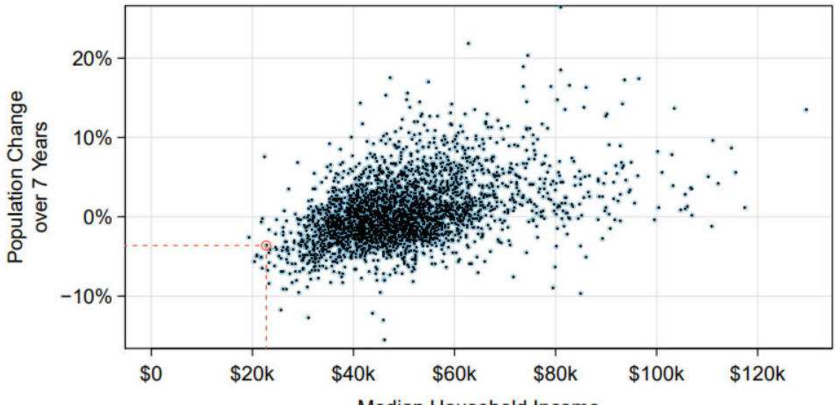


Figure 1.9: A scatterplot showing `pop_change` against `median_hh_income`. Owsley County of Kentucky, is highlighted, which lost 3.63% of its population from 2010 to 2017 and had median household income of $22,736.

Associated vs. independent

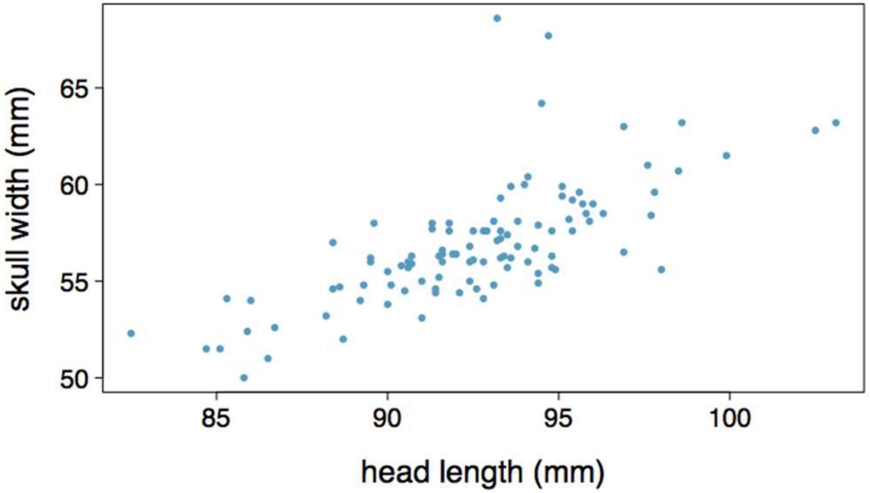
- When two variables show some connection with one another, they are called *associated* variables.
  - Associated variables can also be called *dependent* variables and vice-versa.
  - Positive association vs negative association
- If two variables are not associated, i.e. there is no evident connection between the two, then they are said to be *independent*.





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.

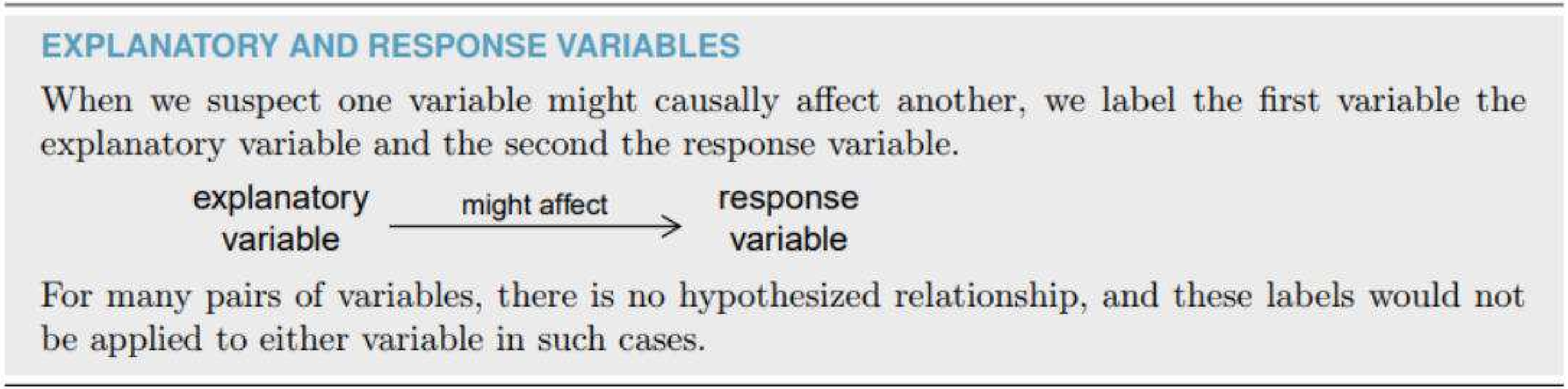
(b) Head length and skull width are positively associated.

(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.

Explanatotory and response variable



EXPLANATORY AND RESPONSE VARIABLES

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

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

For many pairs of variables, there is no hypothesized relationship, and these labels would not be applied to either variable in such cases.

1.3 Sampling principles and strategies

Populations and Samples

PHYS ED | AUGUST 29, 2012, 12:01 AM | 21 Comments

Finding Your Ideal Running Form

By GRETCHEN REYNOLDS

David De Lossy/Getty Images

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: All people

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

Population to which results can be generalized: Adult women, if the data are randomly sampled

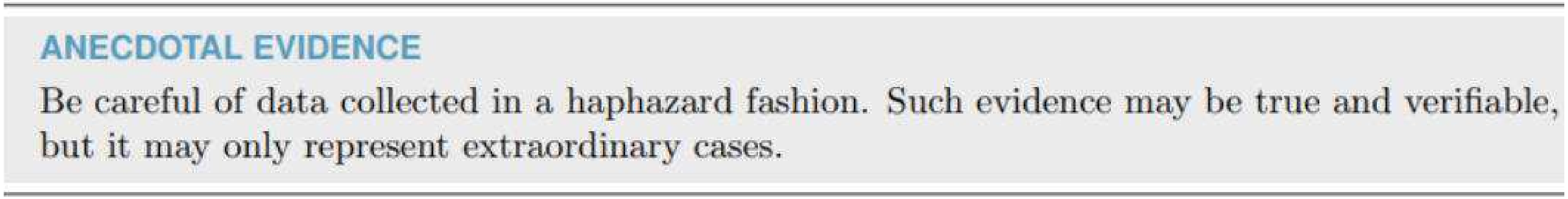
Anecdotal evidence

Brandt, **The Cigarette Century** (2009), Basic Books.

- 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.

- In time researchers were able to examine larger samples of cases (smokers), and trends showing that smoking has negative health impacts became much clearer.



ANECDOTAL EVIDENCE

Be careful of data collected in a haphazard fashion. Such evidence may be true and verifiable, but it may only represent extraordinary cases.

Census

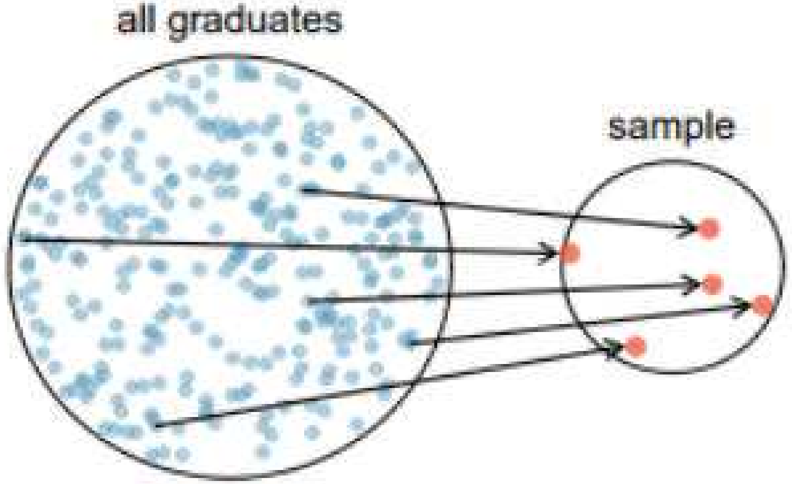
- Wouldn't it be better to just include everyone and "sample" the entire population?
  - This is called a *census*.

- 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.

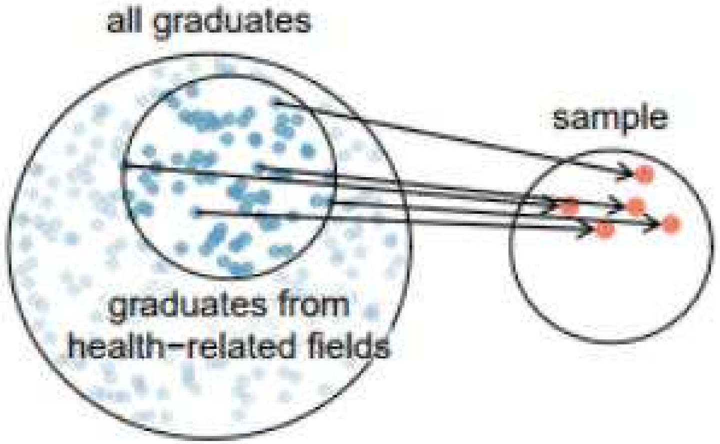
Exploratory analysis to inference

- 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.

Random sampling

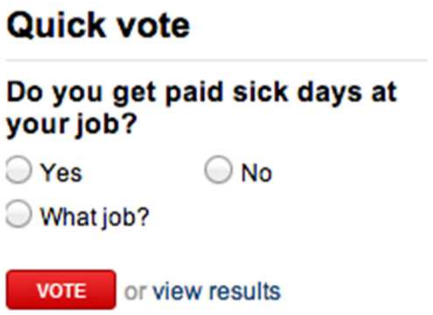


Sampling bias



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.

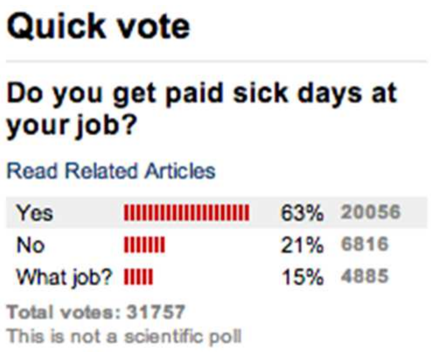


Quick vote

Do you get paid sick days at your job?

○ Yes ○ No

○ What job?



Quick vote

Do you get paid sick days at your job?

Yes	IIIIIIIIIIIIIIIIIIII	63%	20056
No	IIIIII	21%	6816
What job?	IIIII	15%	4885

Total votes: 31757

This is not a scientific poll

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

Obtaining Good Samples

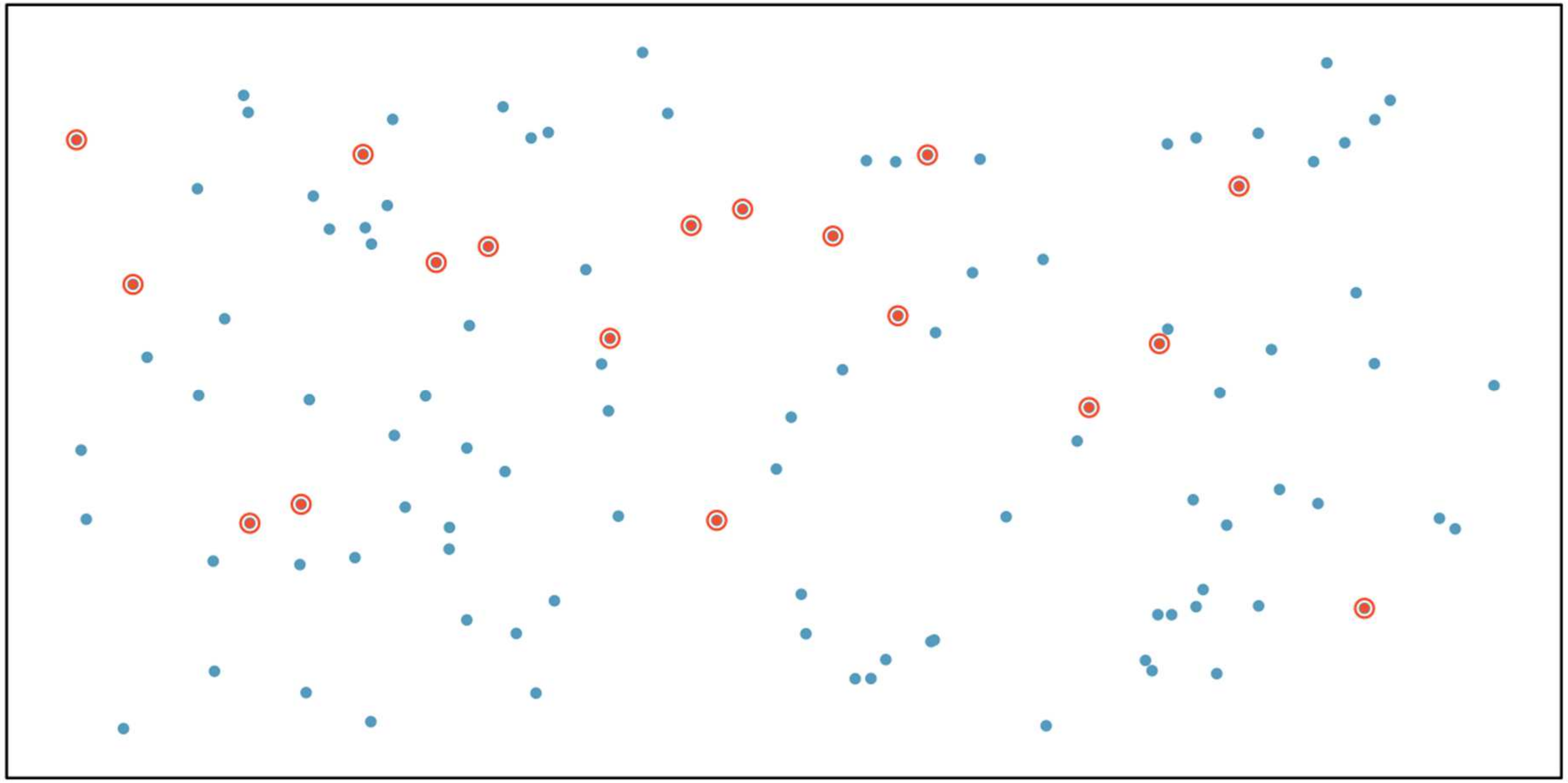
- Almost all statistical methods are based on the notion of implied randomness.

- 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.

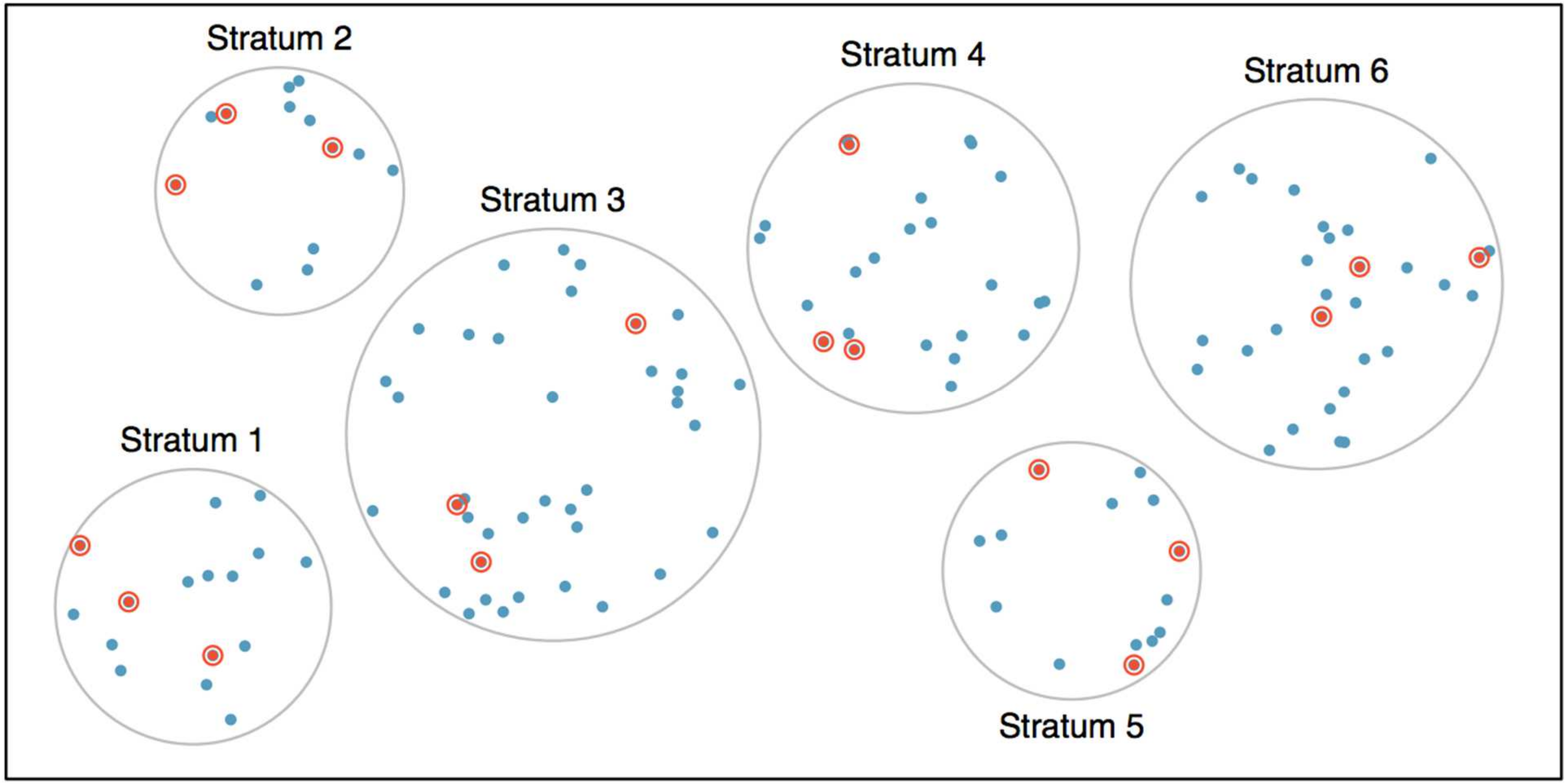
Simple Random Sample

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



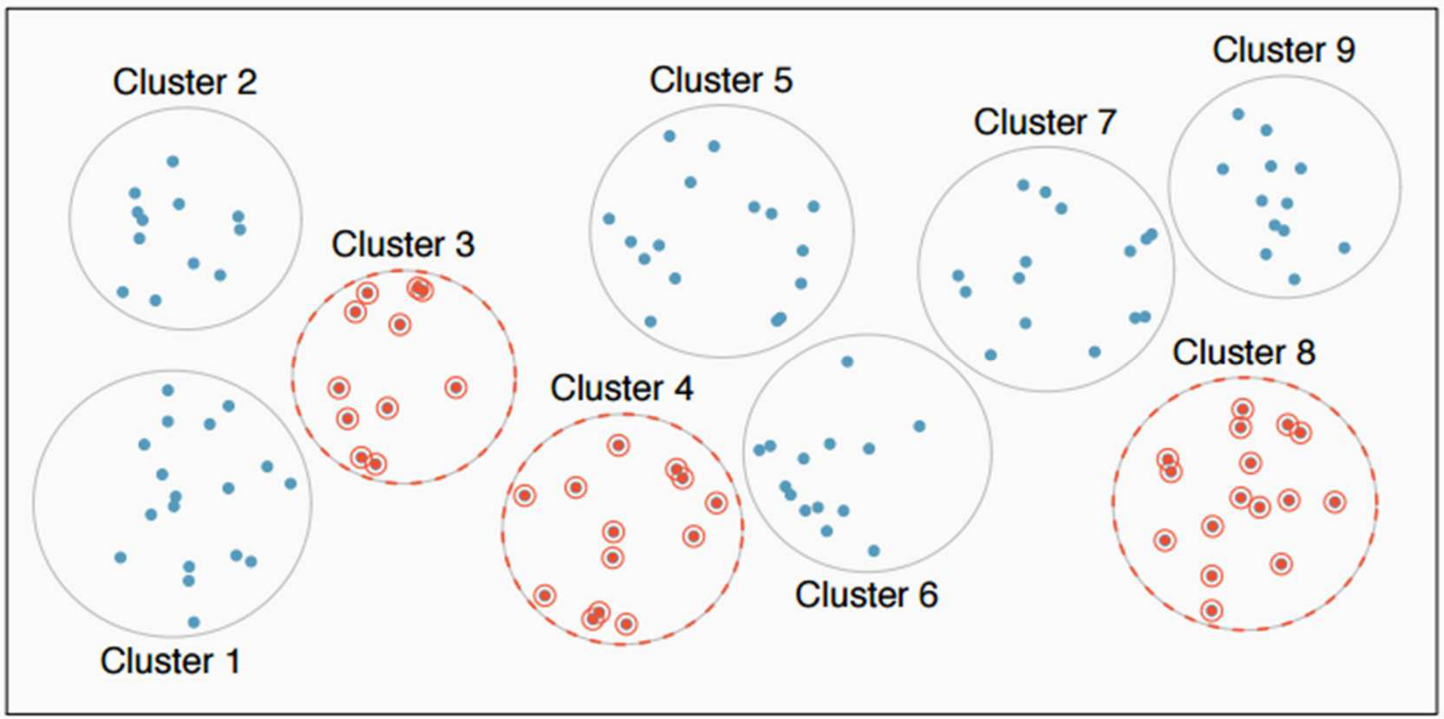
Stratified Sample

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



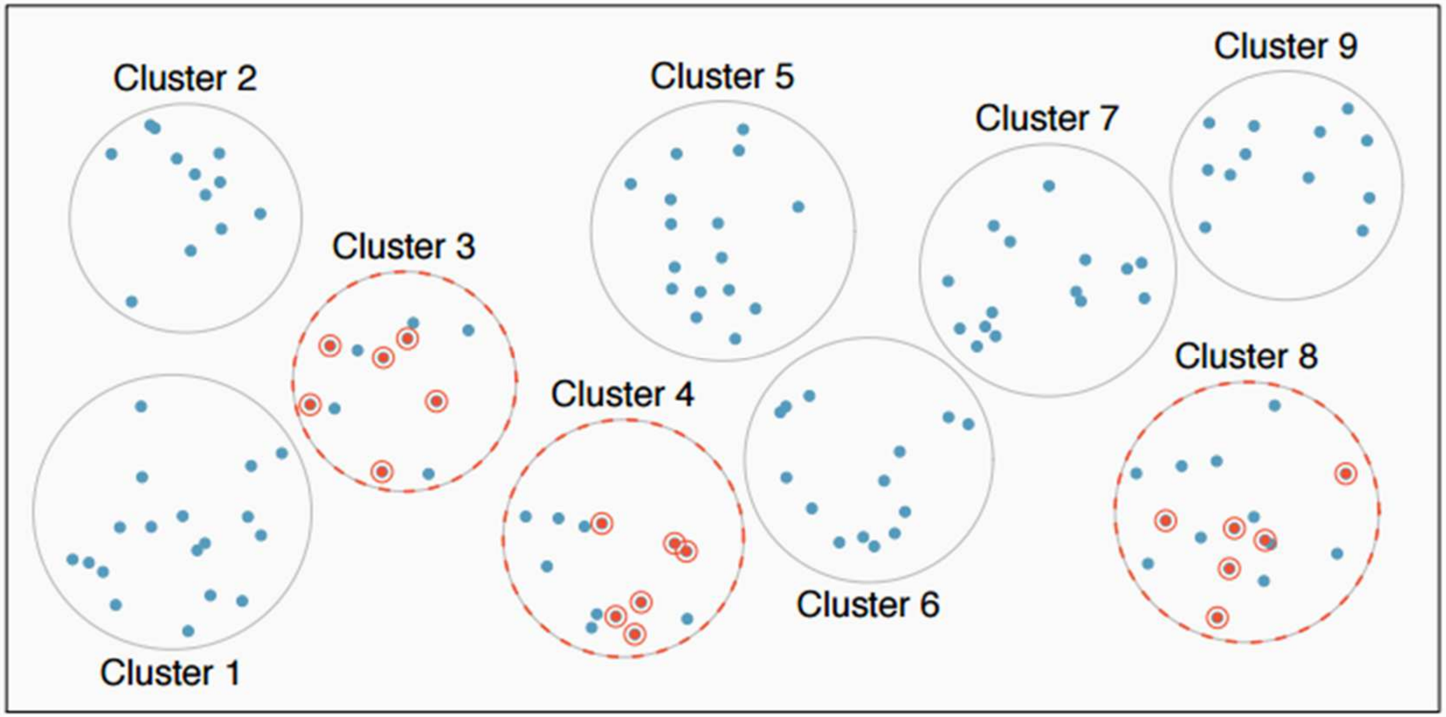
Cluster Sample

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.



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



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) Multistage sampling