Applied Statistics

The course covers fundamental topics of statistical inference and introduces students to statistical programming.

Course website

https://polyhedron.math.uoc.gr/2223/moodle/course/edit.php?id=17

Class Hours

Mon 11.15-13.00 (recitation)

Tue 11.15-13.00

Thu 11.15-13.00

TA

Stelios Grammatikakis

MATERIAL

- (OS) OpenIntro Statistics,https://www.openintro.org/book/os/ (you can download it for free on the website)
- (ΣΤ) Στατιστική: Ανάλυση δεδομένων με χρήση της R, Witte Robert, Witte John, Ανδρουλάκης Γεώργιος, Κουνετάς Κωνσταντίνος.

LEARNING OUTCOMES

1. Understand the meaning of population and sample, recognize different sampling methods and identify their scope and limitations, understand the difference of experimental and observational studies.

2. Use python or R to summarize data numerically and visually, and to perform data analysis.

3. Apply estimation and testing methods (outlier detection, parameter and confidence interval estimation, hypothesis testing, linear and logistic regression) in order to understand natural phenomena and make data-based decisions.

4. Complete a research project using statistical software to analyze data and present your findings

Grading

June:

Midterm exam, 40%

Final Exam, 40%

Project, 20%

The project only counts towards your final grade if your average grade on the midterm and final exam is greater than or equal to 5.

September

Your final grade is computed as follows:

September Exam, 80%

Project, 20%

The project only counts towards your final grade if your grade on the final exam is greater than or equal to 5.

Project grades do not carry over to later years.

What Is Statistics?

- Probability: Language for dealing with uncertainty

e.g., Model a fair coin:

$$X: 0\ if\ tails, 1\ if\ heads.$$
$$X \sim Bernoulli(0.5)$$

- Statistics: You flip a (possibly biased) coin 100 times and record the values of X.
- Data: $0,0,1,1,1, \ldots, 1, 0,1$
- Given the data, find a probability model for X
- If 70% of the values are 1's, is the coin fair?

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 physicians in 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

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

Treatment Group: $19/27 \approx 0.70 \rightarrow 70\%$

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

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

- Proportion with good outcomes in control group

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

Understanding the results

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?

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

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**: 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

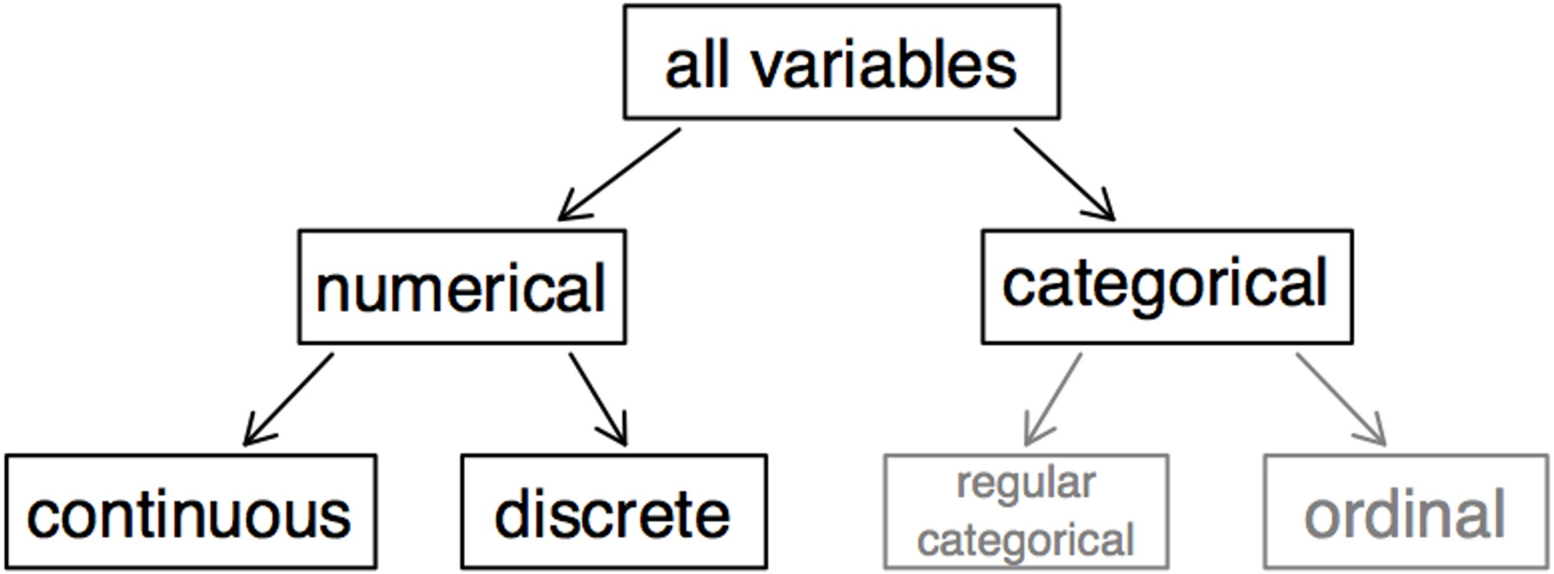
Data collected on students in a statistics class on a variety of variables:

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



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*

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:

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*

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:

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*

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:

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*

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:

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

Practice

What type of variable is a telephone area code?

(a) numerical, continuous

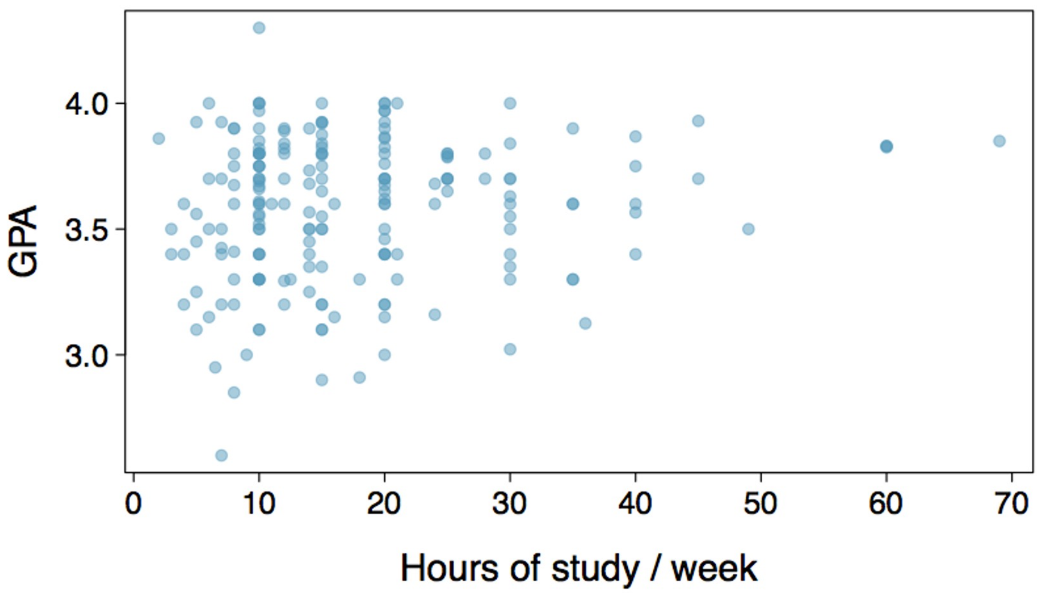
(b) numerical, discrete

(c) *categorical*

(d) categorical, ordinal

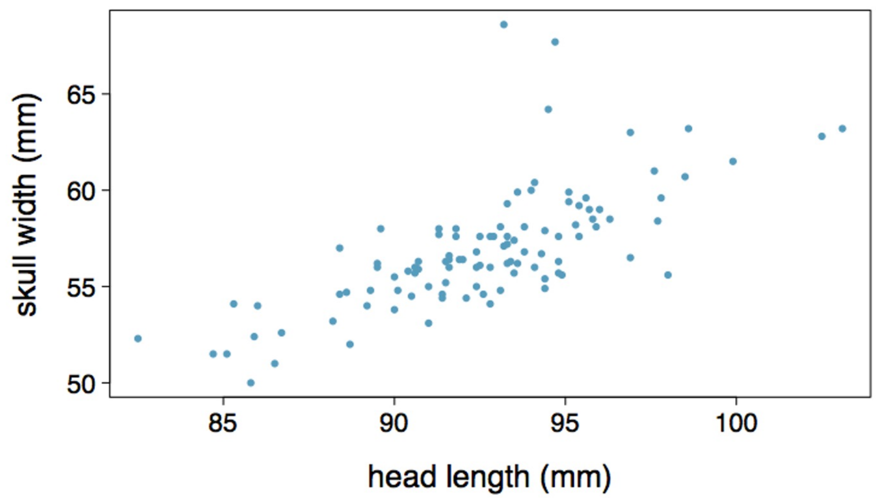
Relationships among variables

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



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.

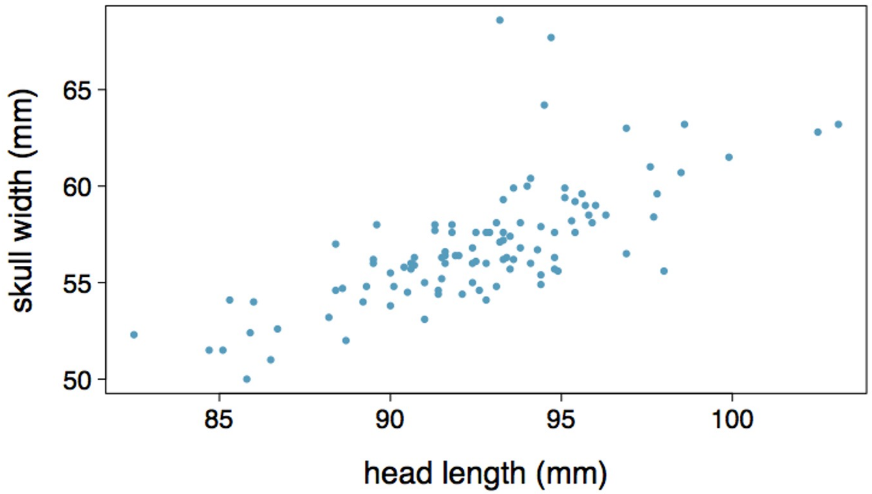
(c) Skull width and head length are negatively associated.

(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.

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.
- If two variables are not associated, i.e. there is no evident connection between the two, then they are said to be *independent*.

Observational studies and sampling 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?

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

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

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

Census

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

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.

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

↓ Download

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

Exploratory analysis to inference

- Sampling is natural.

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.

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*.

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*.

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.

Sampling bias

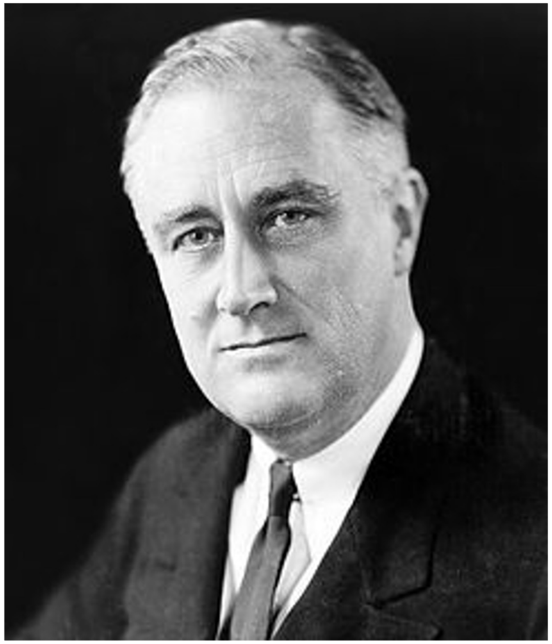
- Non-response: If only a small fraction of the randomly sampled people choose to respond to a survey, the sample may no longer be representative of the population.
- Voluntary response: Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the issue.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

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

Sampling bias example:
Landon vs. FDR

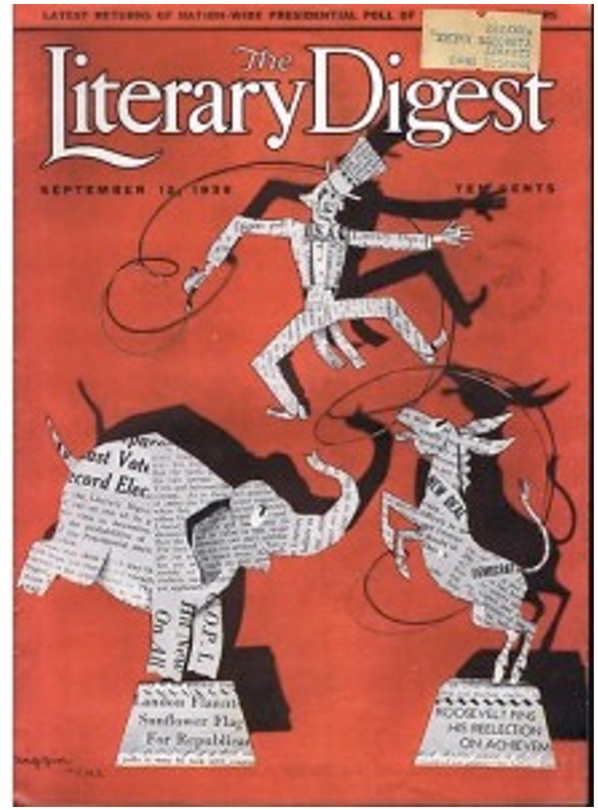
A historical example of a biased sample yielding misleading results

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



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.



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.

Large samples are preferable, but...

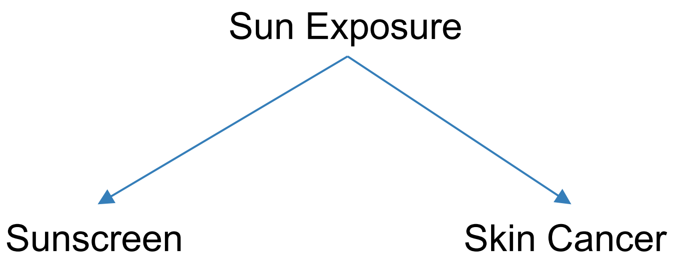
- The Literary Digest election poll was based on a sample size of 2.4 million, which is huge, but since the sample was *biased*, the sample did not yield an accurate prediction.

- 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.

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.

Example: A 1999 study found that sunscreen is associated with skin cancer (i.e, the more sunscreen you use the more likely to get skin cancer.



Obtaining good samples

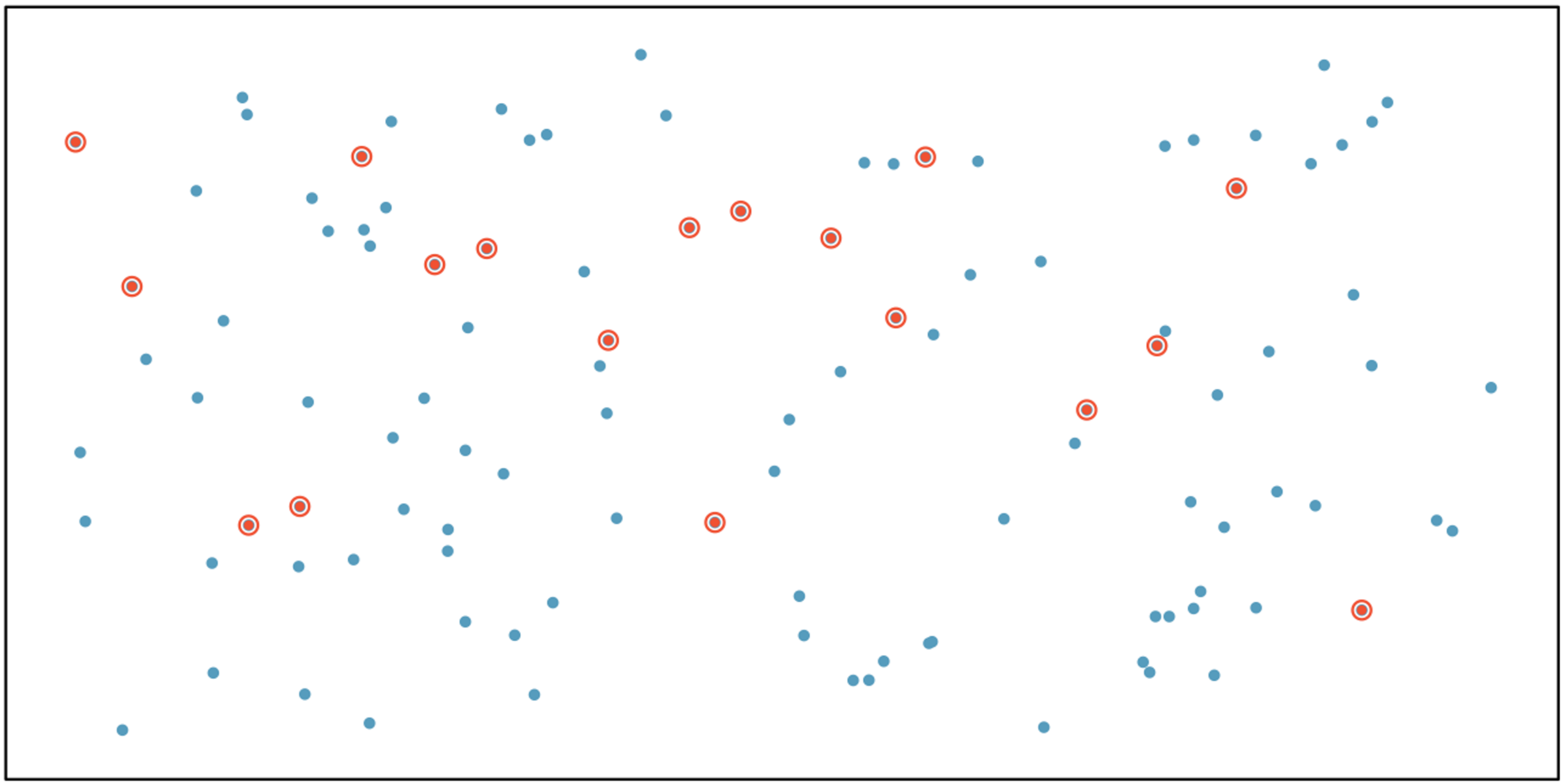
- Almost all statistical methods are based on the notion of implied randomness.

- 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.

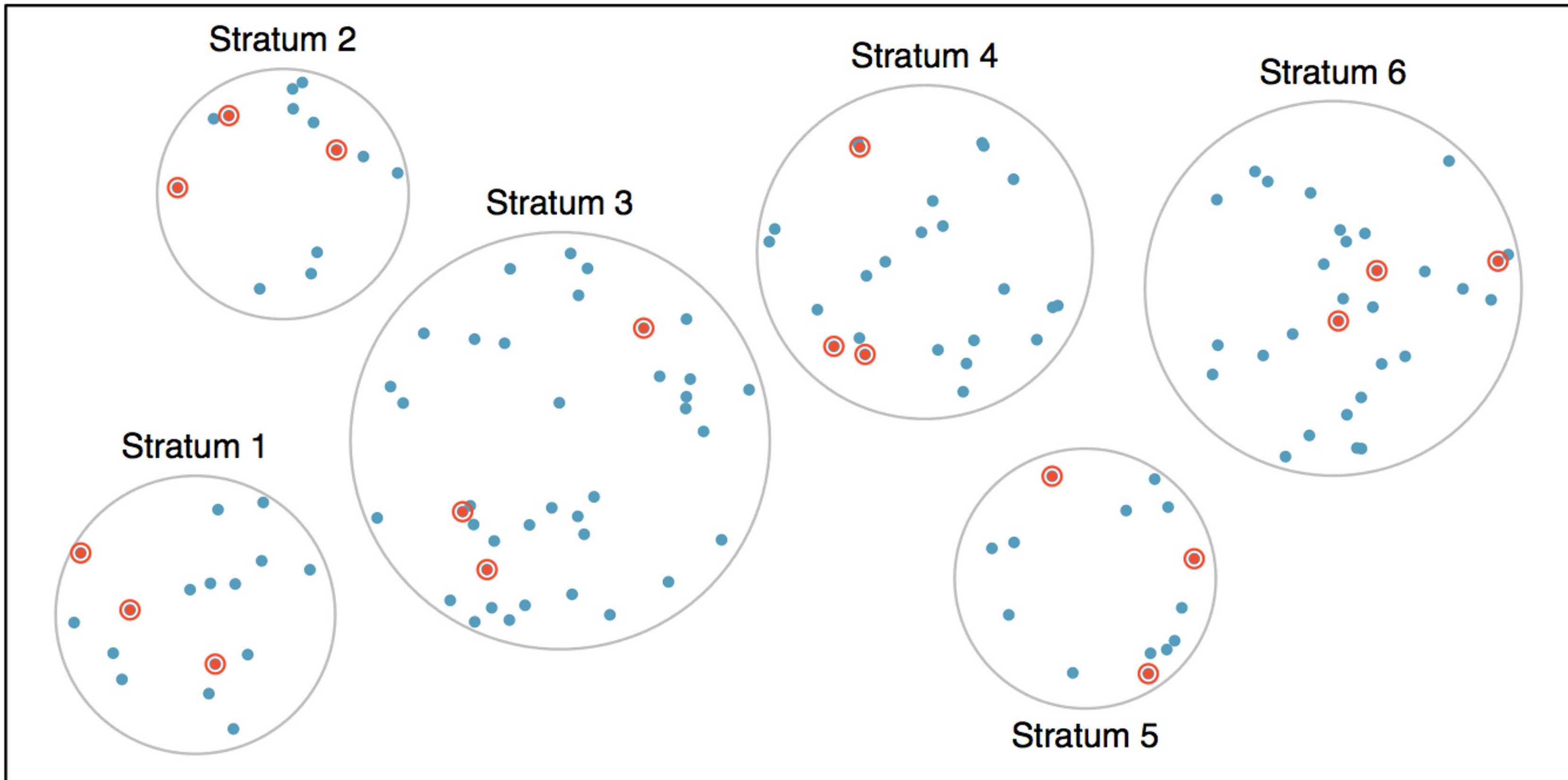
Simple Random Sample

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



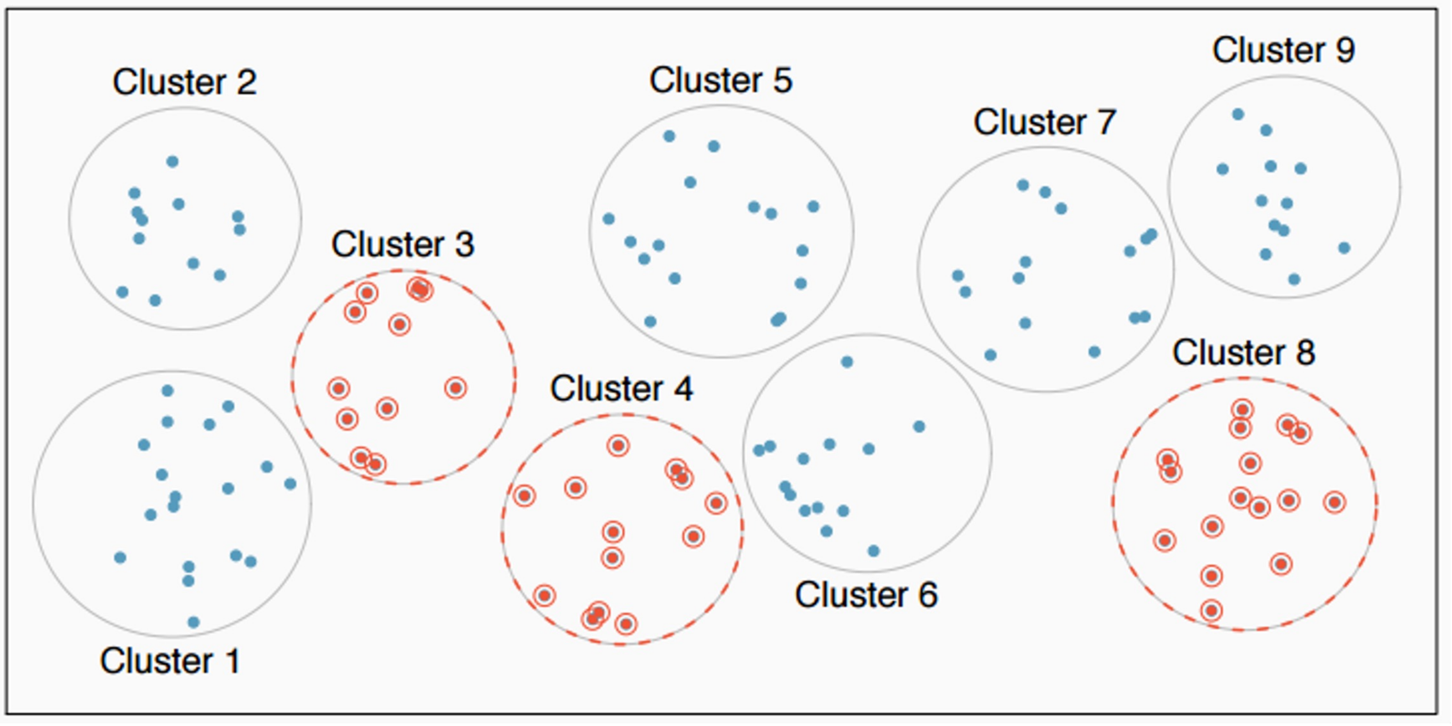
Stratified Sample

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.



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

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

Experiments

Principles of experimental design

1. **Control**: Compare treatment of interest to a control group.
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.

Randomize

- Split patients in two groups: Control and Treatment
- Give treatment to Treatment group, do not give treatment to control group

- We would like to design an experiment to investigate if energy gels makes you run faster:
    - Treatment: energy gel
    - Control: no energy gel
- If on average treatment group improves running times more than control group (statistically significant), conclude the gel works

What could go wrong?

Energy gels might affect pro and amateur athletes differently

Blocking

- 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

What could go wrong?

Knowing you are in the treatment group boosts your confidence.

Placebo

- 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

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)

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)

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.

Practice

What is the main difference between observational studies and experiments?

A. Experiments take place in a lab while observational studies do not need to.

B. In an observational study we only look at what happened in the past.

C. Most experiments use random assignment while observational studies do not.

D. Observational studies are completely useless since no causal inference can be made based on their findings.

Practice

What is the main difference between observational studies and experiments?

A. Experiments take place in a lab while observational studies do not need to.

B. In an observational study we only look at what happened in the past.

C. *Most experiments use random assignment while observational studies do not.*

D. Observational studies are completely useless since no causal inference can be made based on their findings.

Random Assignment vs. Random Sampling