ECO239

Week 3

Fall 2018/2019

Announcement

- Next week, bring your laptop if you have one.

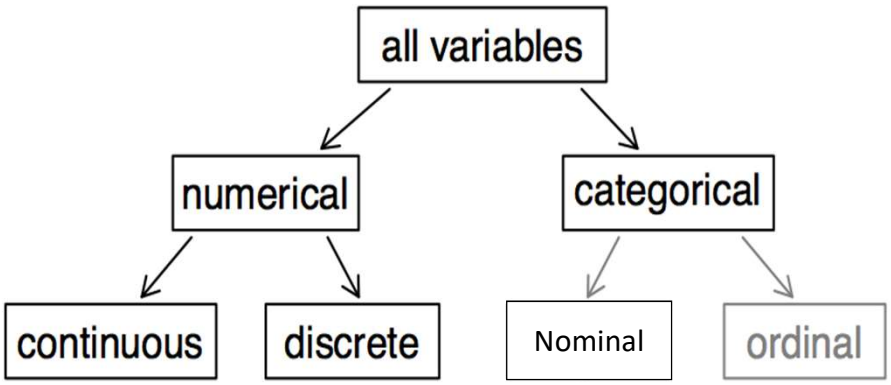
- Before coming to the class,

Download 1) R program (https://www.r-project.org/)

2) R Studio (https://www.rstudio.com/)

3) highgpa.csv file from course webpage

* Have to be downloaded in this order.



Example 1.3 Data were collected about students in a statistics course. Three variables were recorded for each student: number of siblings, student height, and whether the student had previously taken a statistics course. Classify each of the variables as continuous numerical, discrete numerical, or categorical.

- Number of siblings
- Student height
- Whether the students had taken a statistics course before.

Example 1.3 Data were collected about students in a statistics course. Three variables were recorded for each student: number of siblings, student height, and whether the student had previously taken a statistics course. Classify each of the variables as continuous numerical, discrete numerical, or categorical.

- Number of siblings <= Continuous, Discrete
- Student height <= Continuous, Numerical
- Whether the students had taken a statistics course before. <= Categorical, Nominal

Classify the answers to the following questions highgpa

0. GPA
1. Weekly Studying Hours
2. # classes taken in this semester
3. attendance to the classes (___ / 14 week)
4. which row you are sitting (1^{st} (1), 2^{nd} (2), middle (3), back (4))
5. telephone use during the classes (How many mins looking at your phone during the class)
6. studying style (Group vs Individual) (0: Individual, 1: Group)
7. Existence of Partner (girlfriend/boyfriend) (0: no , 1: yes)
8. if the students stay at home or at dorm (0: dorm (student house), 1: home)
9. the commuting distance (time spent – two ways in minutes)
10. working or not (if working , how many hours per week)
11. the length of sleep
12. target GPA (your expectation at graduation.)

Census

- Wouldn't it be better to just include everyone and "sample" the entire population?
- This is called Census
- Why don't we do this usually???

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

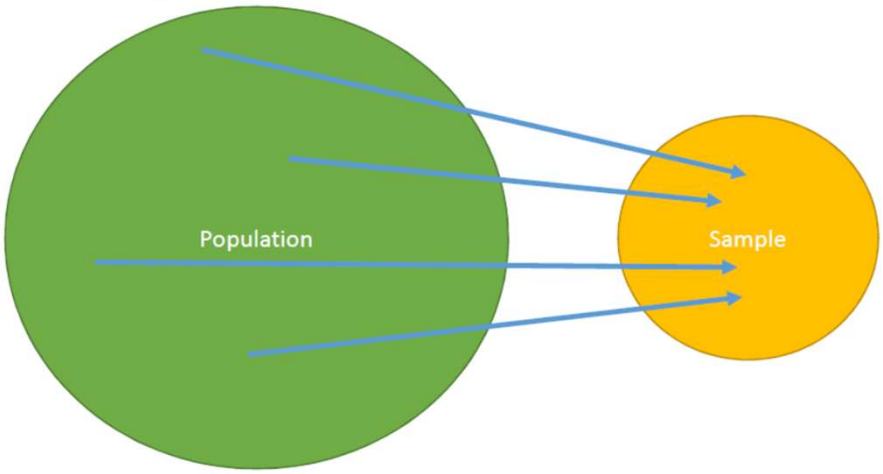
There is an effort underway to make sure Hispanics are accurately counted in the 2010 Census. Phoenix has some of the country's "hardest-to-count" districts. Some Latinos, especially illegal residents, fear that participating in the count will expose them to immigration raids or government harassment.

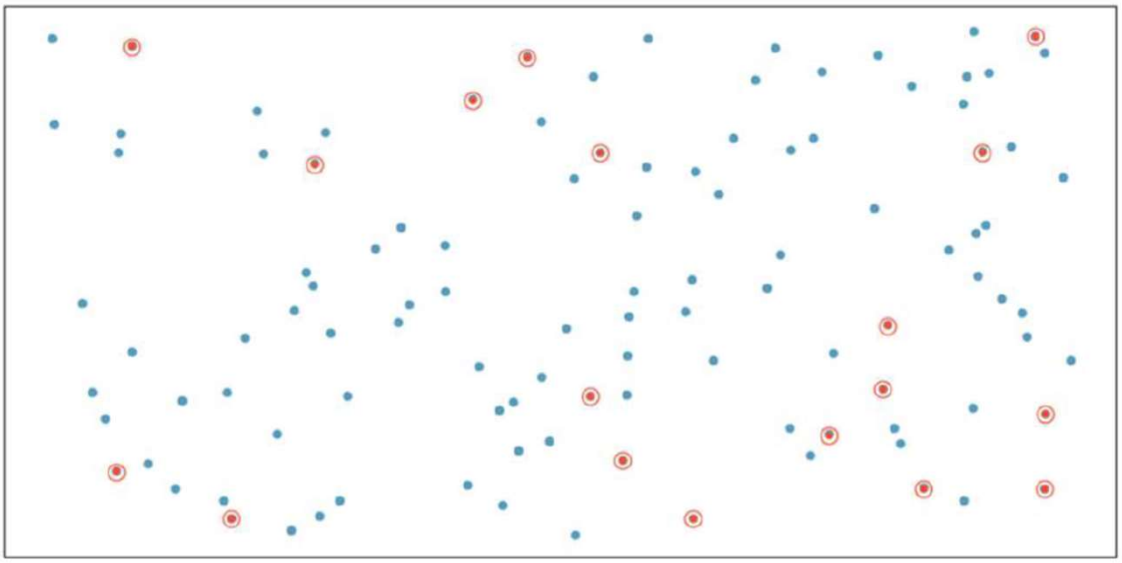
http://www.npr.org/templates/story/story.php?storyId=125380052

Sampling

Question: How should we sample 1000 observations which represent Ankara population?

Radom 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

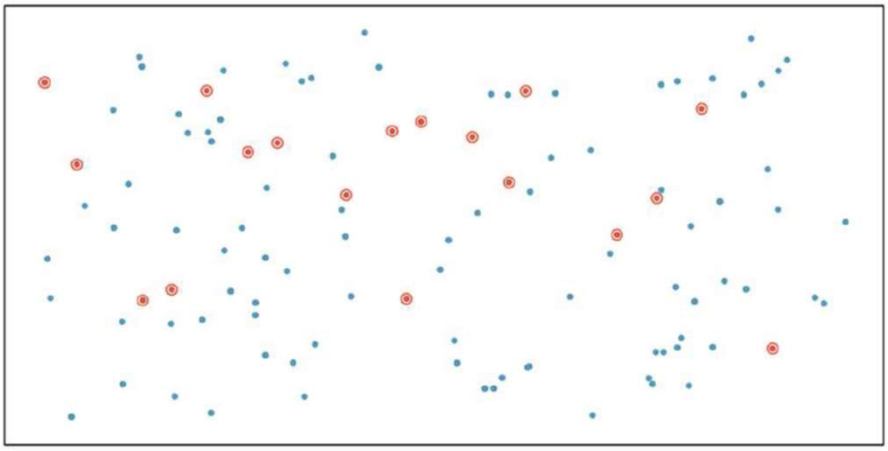
- Every object has an equal probability of being selected.

How can you do this?

Task: Consider the way to sample 10 students from the students sitting in this classroom.

Simple Random Sample

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



Simple Random Sample

- If you have a large population, you may use Random Number Generator.

R-command

- sort(sample(*population size*, *sample size*, replace=*False*))

For example, if you want to sample 10 sample out of 50 students,

```
sort(sample(50, 10, replace= False))
```

```
[1]  5 12 13 15 19 22 37 38 39 42
```

IF you don't have access to random sample generator, but need to sample urgently…try Systematic Sampling

- Choose sample size n.
- Set k = N/n.
- Select one random number (R) from 1 and k.
- Sample R, R+k, R+2k, R+3k…. R+(n-1)k.

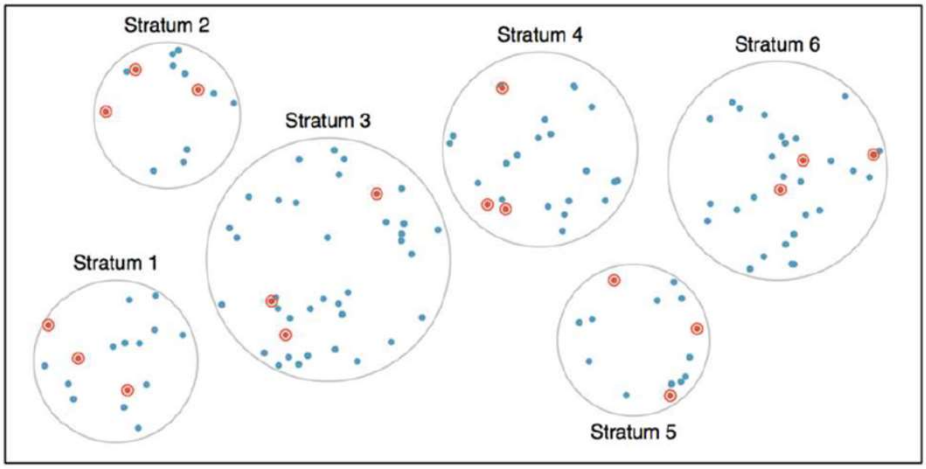
Example.
If N = 46000, n = 46.
k= 1000
R = 596

Then IDs sampled are
596, 1596, 2596,…. 596+(46-1)*1000=45596.

*Not same as simple random sampling. Simple random sampling is usually preferred.

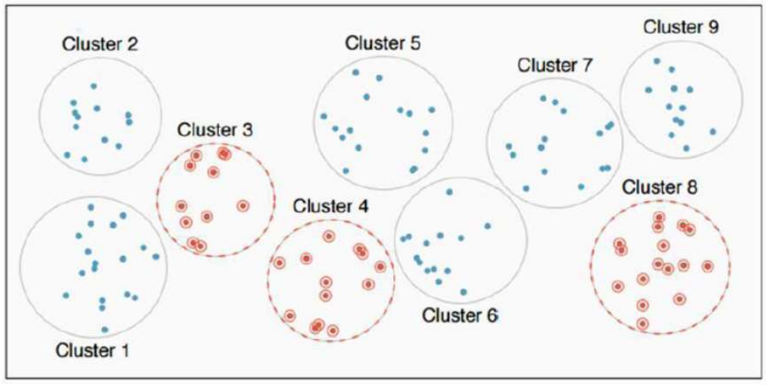
Stratified Sample

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



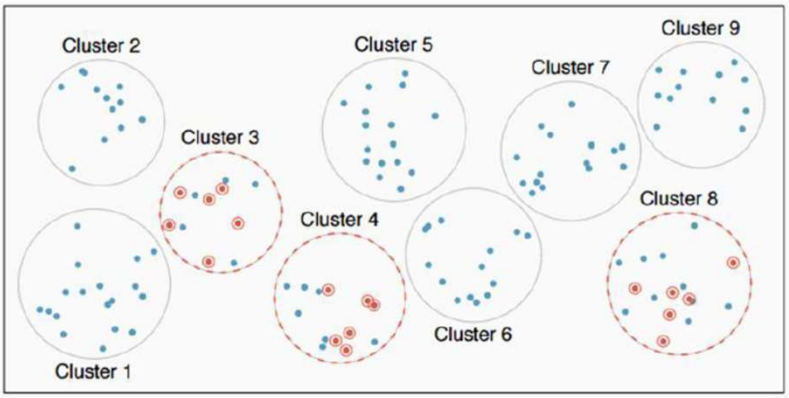
Cluster Sample

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

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

- How should we sample 1000 observations which represent Ankara population?

If you are interested in
- Household Income
- Education level
- Job

Biased Sample

Example:
Reviews on Products, Hotels, Instructors….

- If 50% of online reviews for a product are negative, do you think this means that 50% of buyers are dissatisfied with the product?

(Public) Opinions.
- If 80% of WhatsUp (or FB or any other social media) messages state negative comments about A high-school, does it mean that majority is unsatisfied with the school?

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.

Sampling bias

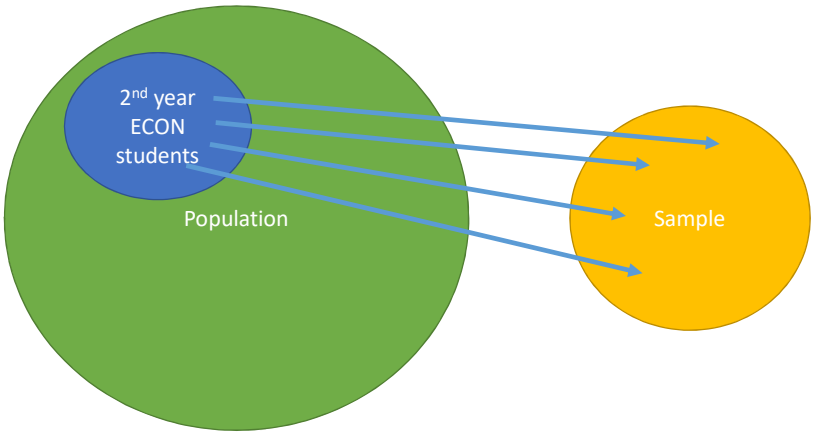
- *Non-response*: If only a small fraction of the randomly sampled people choose to respond to a survey, the sample may no longer be representative of the population.
- *Voluntary response*: Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the issue.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

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.

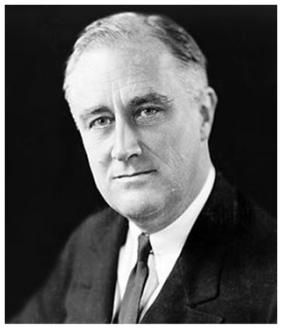
Radom Sample …. ???



Sampling bias example:
Landon vs. FDR

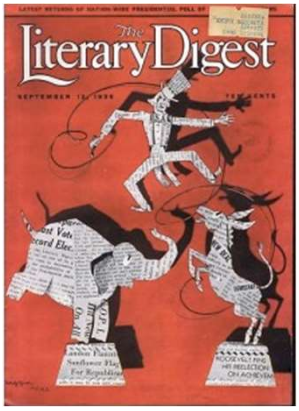
A historical example of a biased sample yielding misleading results

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.

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.

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

Explanatory and Response Variables

- Explanatory variable (a.k.a. X variable, independent variable)
- Response variable (a.k.a. Y variable, dependent variable)

Q: Any Explanatory variable => Response variable relationship?

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:

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

- 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.

County Data

- Explanatory variable (a.k.a. X variable, independent variable)
- Response variable (a.k.a. Y variable, dependent variable)
- E.g. : Is federal spending, on average, higher or lower in counties with high rates of poverty?
- Which one is the explanatory variable, and which one is response variable?

Q: The higher rate of poverty => the higher federal spending?
The higher federal spending => the lower rate of poverty?

Relationships Between Variables

We, social scientists (incl. economists) are often interested in the relationship between two variables.

Q: Is federal spending, on average, higher or lower in counties with high rates of poverty?

⇒Do we expect any relationship between Government Spending and Rate of Poverty?

⇒What kind of relationship do we expect? (Positive, Negative)

⇒How can we answer these questions???

Relationships between Variables

- Collected Data (ECO239_GPA)

Objective: To learn the relationship between GPA and other variables.

1. Weekly Studying Hours
2. # classes taken in this semester
3. attendance to the classes (____/ 14 week)
4. which row you are sitting (1^{st}, 2^{nd} , 3^{rd}, middle, back)
5. telephone use during the classes (How many mins looking at your phone during the class)
6. studying style (Group vs Individual)
7. Existence of Partner (girlfriend/boyfriend)
8. if the students stay at home or at dorm
9. the commuting distance (time spent – two ways)
10. working or not (if working , how many hours per week)
11. the length of sleep
12. target GPA (your expectation at graduation.)

Explanatory variables => Response variable

How can we analyze these relationship???

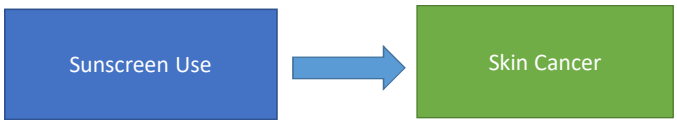
Observational Study vs. Experimental Study

- Observational Study: Researchers collect data in a way that does not directly interfere with how the data arise.
- Data collection: Surveys, Reviewing various records, follow a cohort of many similar individuals.
- Result: Association between explanatory and response variables.

- Experimental Study: Researchers conduct experiments to reveal causations between explanatory and response variables.
- Data collection: Conduct experiments. Set-up Control and Treatment groups.
- Result: Causation between explanatory and response variables.

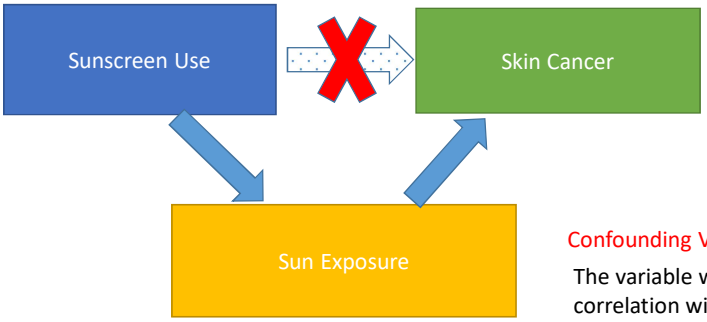
Observational Study

Guided Practice 1.10 Suppose an observational study tracked sunscreen use and skin cancer, and it was found that the more sunscreen someone used, the more likely the person was to have skin cancer. Does this mean sunscreen *causes* skin cancer?[12]



Observational Study

Guided Practice 1.10 Suppose an observational study tracked sunscreen use and skin cancer, and it was found that the more sunscreen someone used, the more likely the person was to have skin cancer. Does this mean sunscreen *causes* skin cancer?[12]



Confounding Variable

The variable which has correlation with explanatory AND response variables.

New study sponsored by General Mills says that eating breakfast makes girls thinner

Study: Breakfast Helps Girls Stay Slim
I love these studies....and finding out who sponsored them!

By ALEX DOMINGUEZ, Associated Press

Girls who regularly ate breakfast, particularly one that includes cereal, were slimmer than those who skipped the morning meal, according to a study that tracked nearly 2,400 girls for 10 years.

Girls who ate breakfast of any type had a lower average body mass index, a common obesity gauge, than those who said they didn't. The index was even lower for girls who said they ate cereal for breakfast, according to findings of the study conducted by the Maryland Medical Research Institute. The study received funding from the National Institutes of Health and cereal-maker General Mills.

"Not eating breakfast is the worst thing you can do, that's really the take-home message for teenage girls," said study author Bruce Barton, the Maryland institute's president and CEO.

The fiber in cereal and healthier foods that normally accompany cereal, such as milk and orange juice, may account for the lower body mass index among cereal eaters, Barton said.

The results were gleaned from a larger NIH survey of 2,379 girls in California, Ohio and Maryland who were tracked between ages 9 and 19. Results of the study appear in the September issue of the Journal of the American Dietetic Association.

Nearly one in three adolescent girls in the United States is overweight, according to the association. The problem is particularly troubling because research shows becoming overweight as a child can lead to a lifetime struggle with obesity.

As part of the survey, the girls were asked once a year what they had eaten during the previous three days. The data were adjusted to compensate for factors such as differences in physical activity among the girls and normal increases in body fat during adolescence.

http:// www.peertrainer.com/ LoungeCommunityThread.aspx?ForumID=1&ThreadID=3118

What type of study is this, observational study or an experiment?

"Girls who regularly ate breakfast, particularly one that includes cereal, were slimmer than those who skipped the morning meal, according to a study that tracked nearly 2,400 girls for 10 years. [...] As part of the survey, the girls were asked once a year what they had eaten during the previous three days."

What is the conclusion of the study?

Who sponsored the study?

What type of study is this, observational study or an experiment?

"Girls who regularly ate breakfast, particularly one that includes cereal, were slimmer than those who skipped the morning meal, according to a study that tracked nearly 2,400 girls for 10 years. [...] As part of the survey, the girls were asked once a year what they had eaten during the previous three days."

This is an observational study since the researchers merely observed the behavior of the girls (subjects) as opposed to imposing treatments on them.

What is the conclusion of the study?

There is an association between girls eating breakfast and being slimmer.

Who sponsored the study?

What type of study is this, observational study or an experiment?

"Girls who regularly ate breakfast, particularly one that includes cereal, were slimmer than those who skipped the morning meal, according to a study that tracked nearly 2,400 girls for 10 years. [...] As part of the survey, the girls were asked once a year what they had eaten during the previous three days."

This is an observational study since the researchers merely observed the behavior of the girls (subjects) as opposed to imposing treatments on them.

What is the conclusion of the study?

There is an association between girls eating breakfast and being slimmer.

Who sponsored the study?

General Mills.

3 Possible Explanations

1. Eating breakfast causes girls to be thinner.

3 Possible Explanations

1. Eating breakfast causes girls to be thinner.

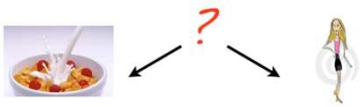
2. Being thin causes girls to eat breakfast.

3 Possible Explanations

1. Eating breakfast causes girls to be thinner.

2. Being thin causes girls to eat breakfast.

3. A third variable is responsible for both. What could it be? An extraneous variable that affects both the explanatory and the response variable and that make it seem like there is a relationship between the two are called confounding variables.



Images from: http:// www.appforhealth.com/ wp-content/ uploads/ 2011/ 08/ ipn-cerealfrijo-300x135.jpg, http:// www.dreamstime.com/ stock-photography-too-thin-woman-anorexia-model-image2814892.

Experimental Study

- Task: Statistically test if a newly introduced drug A is effective to reduce the risk of heart attack.

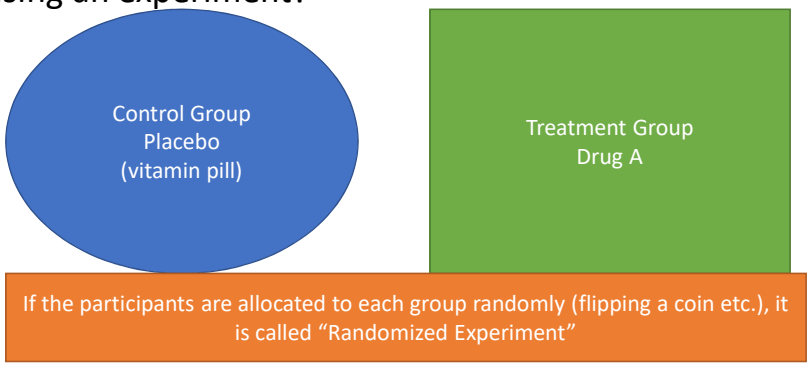
- Q: How can we test the effectiveness of the drug using an experiment?

=> Group Work

Experimental Study

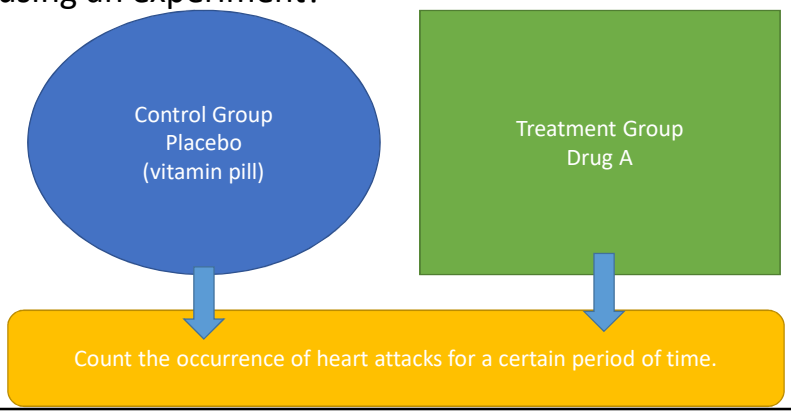
- Task: Statistically test if a newly introduced drug A is effective to reduce the risk of heart attack.

- Q: How can we test the effectiveness of the drug using an experiment?



Experimental Study

- Task: Statistically test if a newly introduced drug A is effective to reduce the risk of heart attack.
- Q: How can we test the effectiveness of the drug using an experiment?



Causation

TIP: association $\neq$ causation
In general, association does not imply causation, and causation can only be inferred from a randomized experiment.

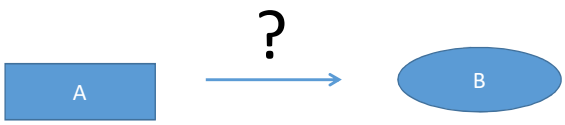
Problem with "Causation"

- Statistic/Econometric analysis do not prove any "causation".

e.g. Cause – Effect ?

A: The number of hours a kid play "violent" video games.

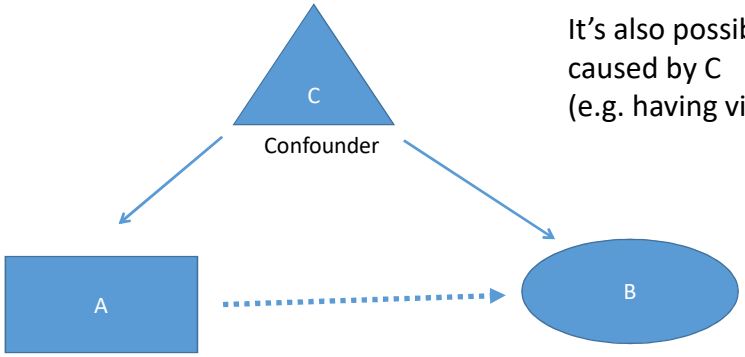
B: Crime committed by a kid.



e.g. Cause – Effect ?

A: The number of hours a kid play "violent" video games.

B: Crime committed by a kid.



It's also possible that A and B are caused by C
(e.g. having violent characteristics).

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.

More on Blocking

- We would like to design an experiment to investigate if energy gels makes you run faster:

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

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:

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

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

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.

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

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