Overview of data collection principles

Population and sample

Consider these possible research questions.

1. What is the average mercury content in swordfish in the Atlantic Ocean?
2. Over the last 5 years, what is the average time to complete a degree for BHCC students?
3. Does a new drug reduce the number of deaths in patients with severe heart disease?

Each question is about a **population**, but a researcher would probably need to use a **sample**.

Consider these possible responses to the three research questions:

1. A man on the news got mercury poisoning from eating swordfish, so the average mercury concentration in swordfish must be dangerously high.
2. I met two students who took more than 7 years to graduate from BHCC, so it must take a long time to graduate.
3. My friend's dad had a heart attack and died after they gave him a new heart disease drug, so the drug must not work.

Consider these possible responses to the three research questions:

1. A man on the news got mercury poisoning from eating swordfish, so the average mercury concentration in swordfish must be dangerously high.
2. I met two students who took more than 7 years to graduate from BHCC, so it must take a long time to graduate.
3. My friend's dad had a heart attack and died after they gave him a new heart disease drug, so the drug must not work.

Data collected in this haphazard fashion are called 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.

Consider these possible responses to the three research questions:

1. A man on the news got mercury poisoning from eating swordfish, so the average mercury concentration in swordfish must be dangerously high.
2. I met two students who took more than 7 years to graduate from BHCC, so it must take a long time to graduate.
3. My friend's dad had a heart attack and died after they gave him a new heart disease drug, so the drug must not work.

Data collected in this haphazard fashion are called 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.

Many news programs run on anecdotes. These tend to warp people's minds about the world (e.g. becoming more scared of terrorists/sharks than cars/drowning).

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:

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:

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 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.

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

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.

2010 Census Participation Rates:

https://www.census.gov/data/datasets/2010/dec/2010-participation-rates.html

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.

Sampling bias

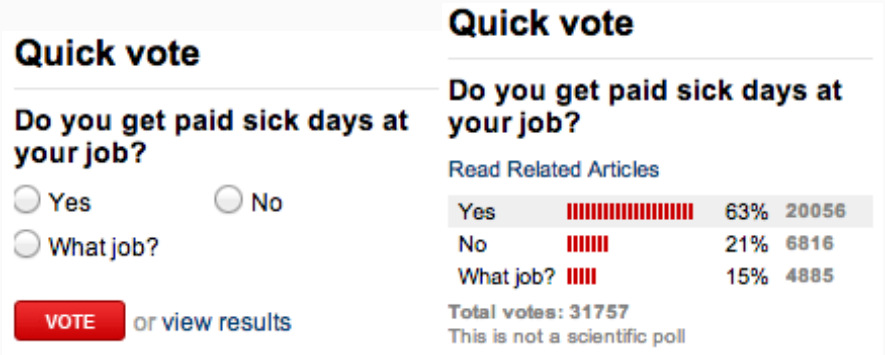
- *Non-response:* If only a small fraction of the randomly sampled people choose to respond to a survey, the sample may no longer be representative of the population.
- *Voluntary response:* Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the issue. 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.

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.



cnn.com, Jan 14, 2012

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.

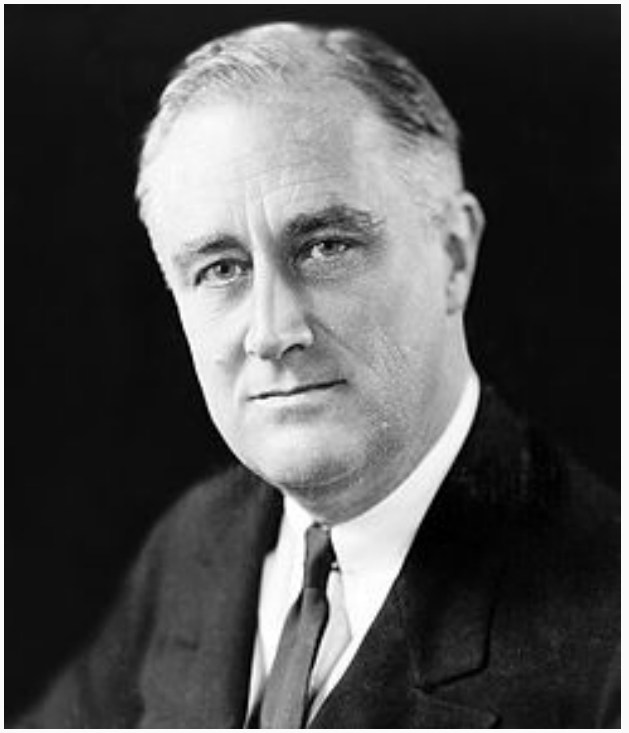
cnn.com, Jan 14, 2012

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

Sampling bias example: Landon vs. FDR

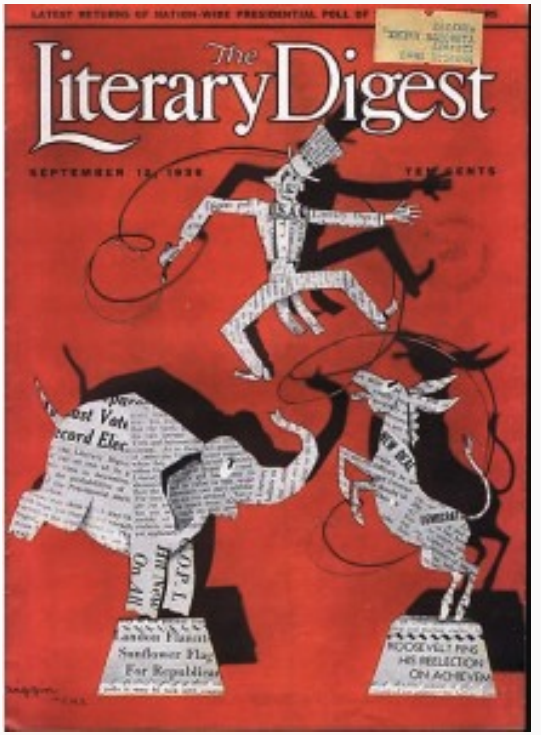
A historical example of a biased sample yielding misleading results:

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.

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

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

- 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.

Observational studies and experiments

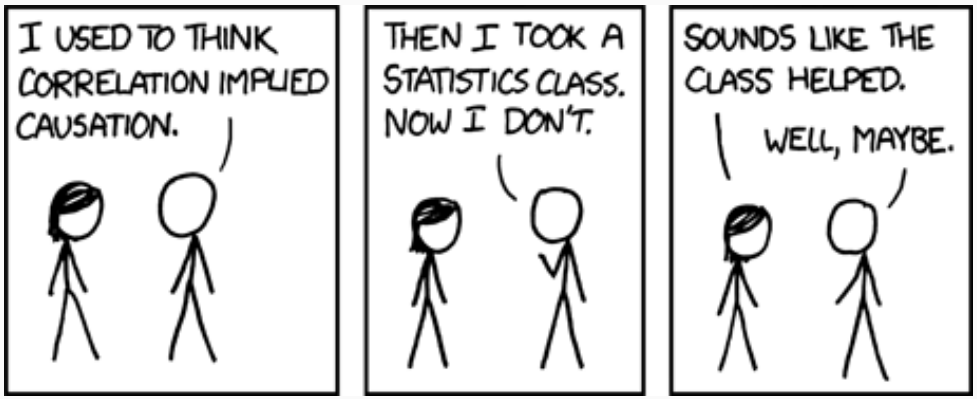
- *Observational study:* Researchers collect data in a way that does not directly interfere with how the data arise, i.e. they merely "observe", and can only establish an association between the explanatory and response variables.

Observational studies and experiments

- *Observational study:* Researchers collect data in a way that does not directly interfere with how the data arise, i.e. they merely "observe", and can only establish an association between the explanatory and response variables.
- *Experiment:* Researchers randomly assign subjects to various treatments in order to establish causal connections between the explanatory and response variables.

Observational studies and experiments

- *Observational study:* Researchers collect data in a way that does not directly interfere with how the data arise, i.e. they merely "observe", and can only establish an association between the explanatory and response variables.
- *Experiment:* Researchers randomly assign subjects to various treatments in order to establish causal connections between the explanatory and response variables.
- If you're going to walk away with one thing from this class, let it be "correlation does not imply causation".



http://xkcd.com/552/