Online CDS 101 *Introduction to Computational and Data Sciences*

# What are the computational and data sciences?

Data sampling

# 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).

# 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.

# 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.

# Populations and samples

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


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

Finding Your Ideal Running Form

By GRETCHEN REYNOLDS

David De Lossy/Getty Images

Source: http://well.blogs.nytimes.com/2012/08/29/finding-your-ideal-running-form

# 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

**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

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

Question: What is the 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

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

Question: What is the population of interest?

*Answer: 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

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

Question: What is the population of interest?

*Answer: All people*

**Study 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

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

Question: What is the population of interest?

*Answer: All people*

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

Question: 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

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

Question: What is the population of interest?

*Answer: All people*

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

Question: Population to which results can be generalized?

*Answer: 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.

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

# 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.

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

# 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.
- "Smoking is a complex human behavior [...] confounded by human variability."

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

# 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.
- "Smoking is a complex human behavior [...] 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.

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

# Sampling from a population: Census

- Wouldn't it be better to just include everyone and "sample" the entire population?

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

# Sampling from a population: Census

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

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

# Sampling from a population: 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:

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

# Sampling from a population: 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.**

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

# Sampling from a population: 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.

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

# Sampling from a population: 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.

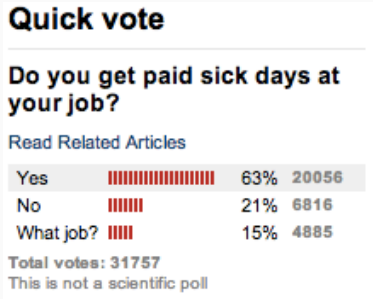
Source: http://www.npr.org/templates/story/story.php?storyId=125380052

# 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

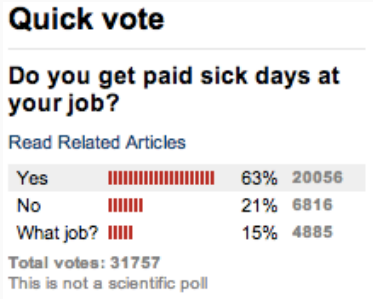
- **Non-response**: If only a small fraction of the randomly sampled people choose to respond to a survey, the sample may no longer be representative of the population.
- **Voluntary response**: Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the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


Source: 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.


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

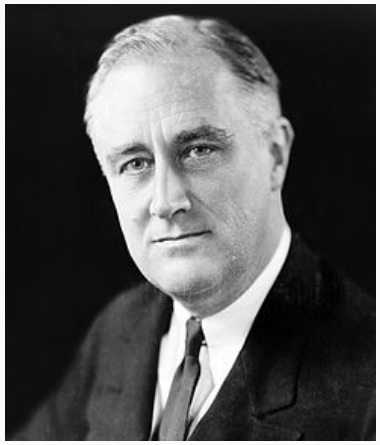
Source: cnn.com, Jan 14, 2012

# Sampling bias example: Landon vs. FDR

A historical example of a biased sample yielding misleading results:

# 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 Literary Digest Poll

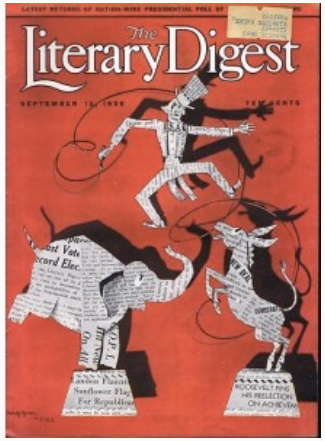
- The Literary Digest polled about 10 million Americans, and got responses from about 2.4 million.
- The poll showed that Landon would likely be the overwhelming winner and FDR would get only 43% of the votes.

# 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 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:

# The Literary Digest Poll - what went wrong?

The magazine had surveyed:

- Its own readers

# The Literary Digest Poll - what went wrong?

The magazine had surveyed:

- Its own readers
- Registered automobile owners

# The Literary Digest Poll - what went wrong?

The magazine had surveyed:

- Its own readers
- Registered automobile owners
- Registered telephone users

# The Literary Digest Poll - what went wrong?

The magazine had surveyed:

- Its own readers
- Registered automobile owners
- 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.

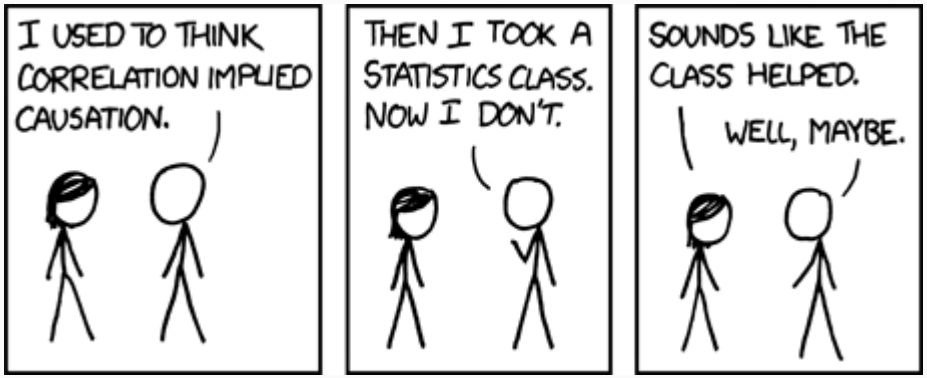
# Large samples are preferable, but...

- 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.

# 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.

# Correlation does not imply causation




Source: http://xkcd.com/552/

# Credits

License



Acknowledgments

Content adapted from the chapter 1 OpenIntro Statistics slides developed by Mine Çetinkaya-Rundel and made available under the CC BY-SA 3.0 license