Foundations of inference

Intro to Data Science

Shawn Santo

Announcements

- Soft deadline for Lab 05 is Thursday
- Soft deadline for Homework 03 is Friday

Today's agenda

- Discuss the foundations of inference
- Identify the conclusions we can and cannot make based on the statistical study

Recall

The statistical process

Statistics is a process that converts data into useful information, whereby practitioners

1. form a question of interest,
2. collect and summarize data,
3. and interpret the results.

The population of interest

The **population** is the group we'd like to learn something about. For example:

- What is the prevalence of diabetes among **U.S. adults**, and has it changed over time?
- Does the average amount of caffeine vary by vendor in **12 oz. cups of coffee at Duke coffee shops**?
- Is there a relationship between tumor type and five-year mortality among **breast cancer patients**?

The **research question of interest** is what we want to answer - often relating one or more numerical quantities or summary statistics.

If we had data from every unit in the population, we could just calculate what we wanted and be done!

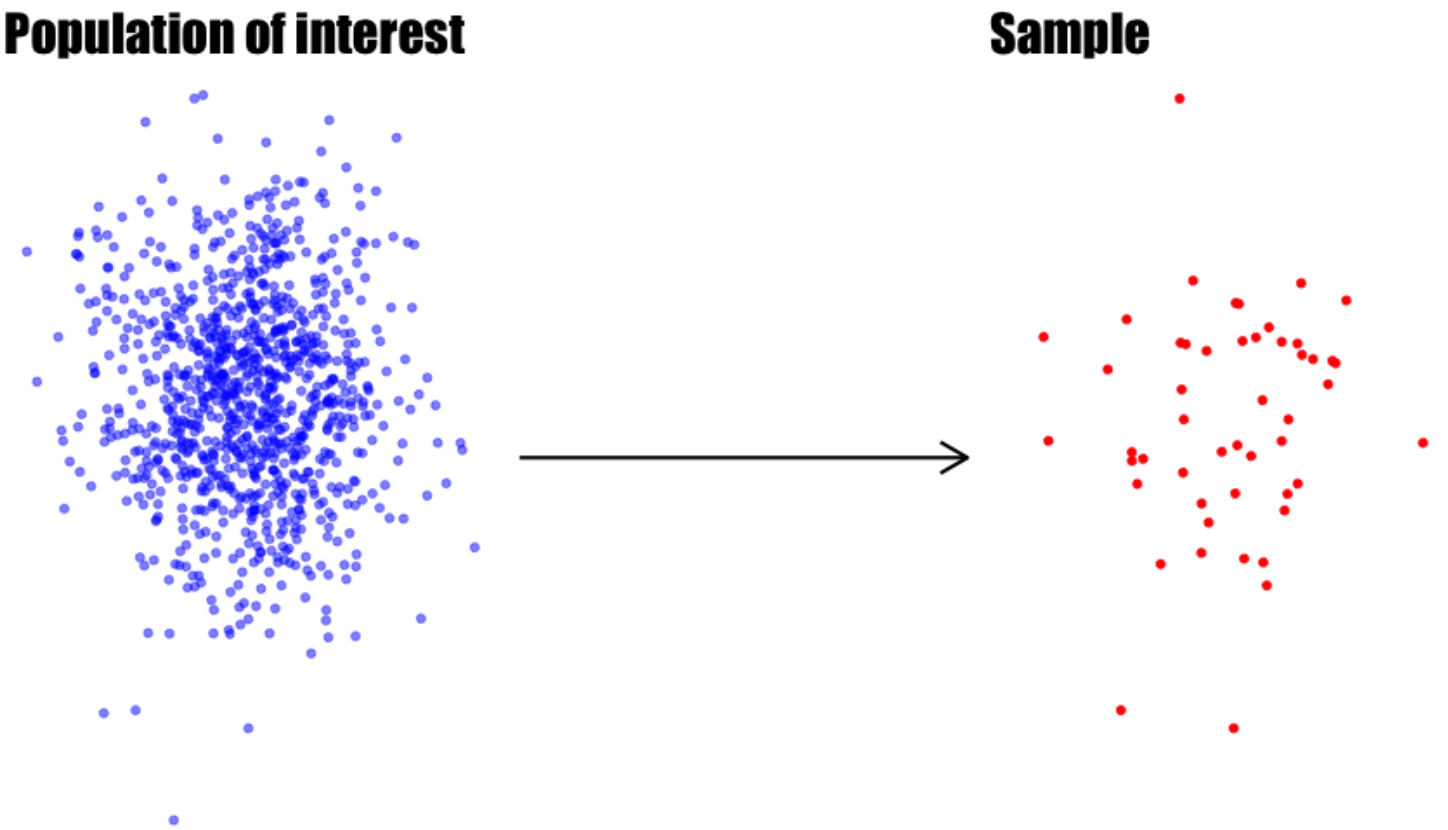
Sampling from the population

Unfortunately, we (usually) have to settle with a **sample** from the population.

Ideally, the sample is **representative**, allowing us to make conclusions that are **generalizable** to the broader population of interest.

In order to make a formal statistical statement about the broader population of interest when all we have is a sample, we need to use the tools of probability and statistical inference.

Big picture



We'll discuss a few population characteristics we'll be interested in

Terminology

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**. *Whether or not we can actually make this causal connection will depend on the type of statistical study (more on this shortly).*

$$\text{Explanatory Variable} \longrightarrow \text{Response Variable}$$

Do larger homes in good locations lead to higher home selling prices? What are the explanatory and response variables?

Population, parameter; sample, statistic

Population: a group of individuals or objects we are interested in studying

Parameter: a numerical quantity derived from the population (almost always unknown)

- Associate the parameter with the population
- Parameters could be the mean, median, correlation, maximum, etc.

If we had data from every unit in the population, we could just calculate population parameters and be done! **Unfortunately, we usually cannot do this.**

Sample: a subset of our population of interest

Statistic: a numerical quantity derived from a sample

- Associate the statistic with the sample
- Statistics could be the mean, median, correlation, maximum, etc.

Statistical inference

Statistical inference is the process of using sample data to make conclusions about the underlying population the sample came from.

- **Estimation**: estimating an unknown parameter based on values from the sample at hand
- **Testing**: evaluating whether our observed sample provides evidence for or against some claim about the population

In the coming lectures we'll discuss each of these inference approaches.

Before we get into this, let's discuss ways samples can be obtained and what type of conclusions we'll be be able to make and **not** make as a result of our statistical process.

Sampling

Sampling strategies

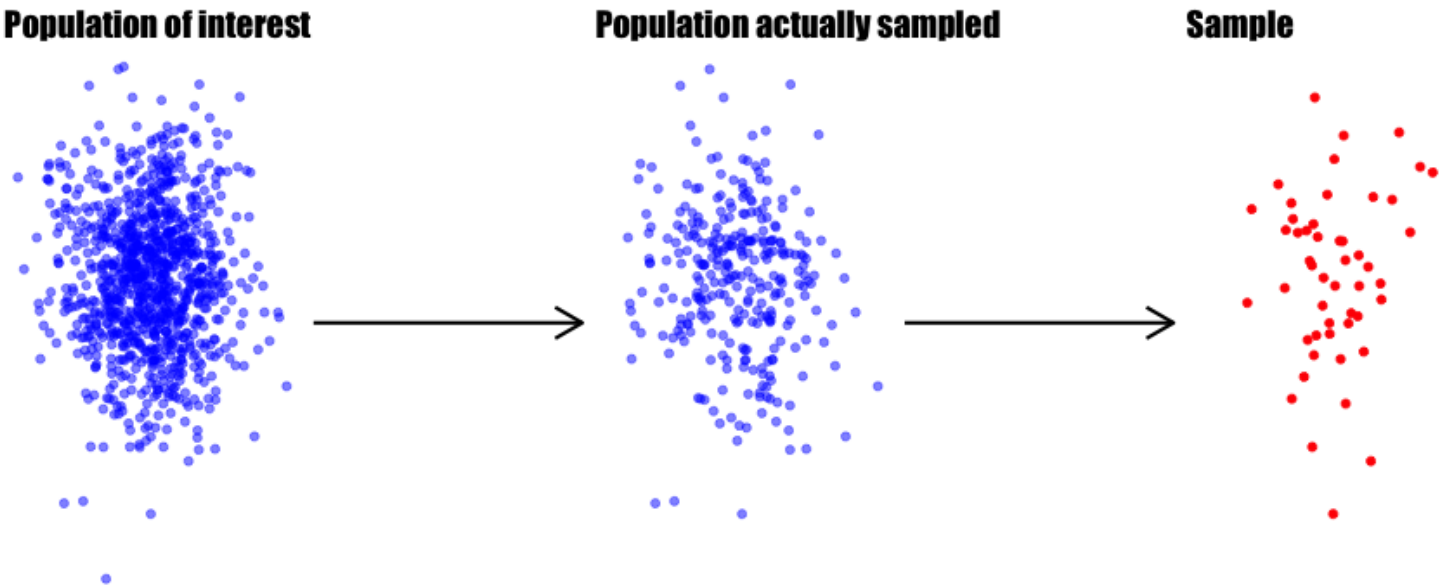
- In our discussions on probability, we considered randomly selecting individuals from studies, where each individual was equally likely to be selected. This form of random sampling is known as **simple random sampling**.
- **Stratified sampling** divides the population into **strata** such that each strata is homogenous. Then a simple random sample is applied within each stratum.
    - Can you think of a reason why we would employ this technique?
- **Cluster sampling** first partitions the population into **clusters**, where each cluster is representative of the population. A fixed number of clusters is selected and all observations within the cluster are included in the sample.
- **Multistage sampling** is similar to cluster sampling, but rather than keep all observations in each cluster, only a random sample of observations is kept.

Example

Suppose we are interested in estimating the malaria rate in a densely tropical portion of rural Indonesia. We learn that there are 30 villages in that part of the Indonesian jungle, each more or less similar to the next. Our goal is to test 150 individuals for malaria. What are the costs and benefits to using the four aforementioned sampling techniques?

Sample bias

- The four sampling strategies help reduce **bias** in our sample. A biased sample can lead to erroneous conclusions.
- Bias can still appear if the non-response rate is very high.
    - Is our sample representative of the population or is it representative of the population that "responded" to the survey?



Statistical studies and conclusions

Observational studies and experiments

- Observational
  - Collect data in a way that does not interfere with how the data arise ("observe")
  - Only establish an **association**
  - Data often cheaper and easier to collect

- Experimental
  - Randomly assign subjects to treatments
  - Establish **causal connections**
  - Often more expensive
  - Sometimes it is impossible or unethical to design an experiment

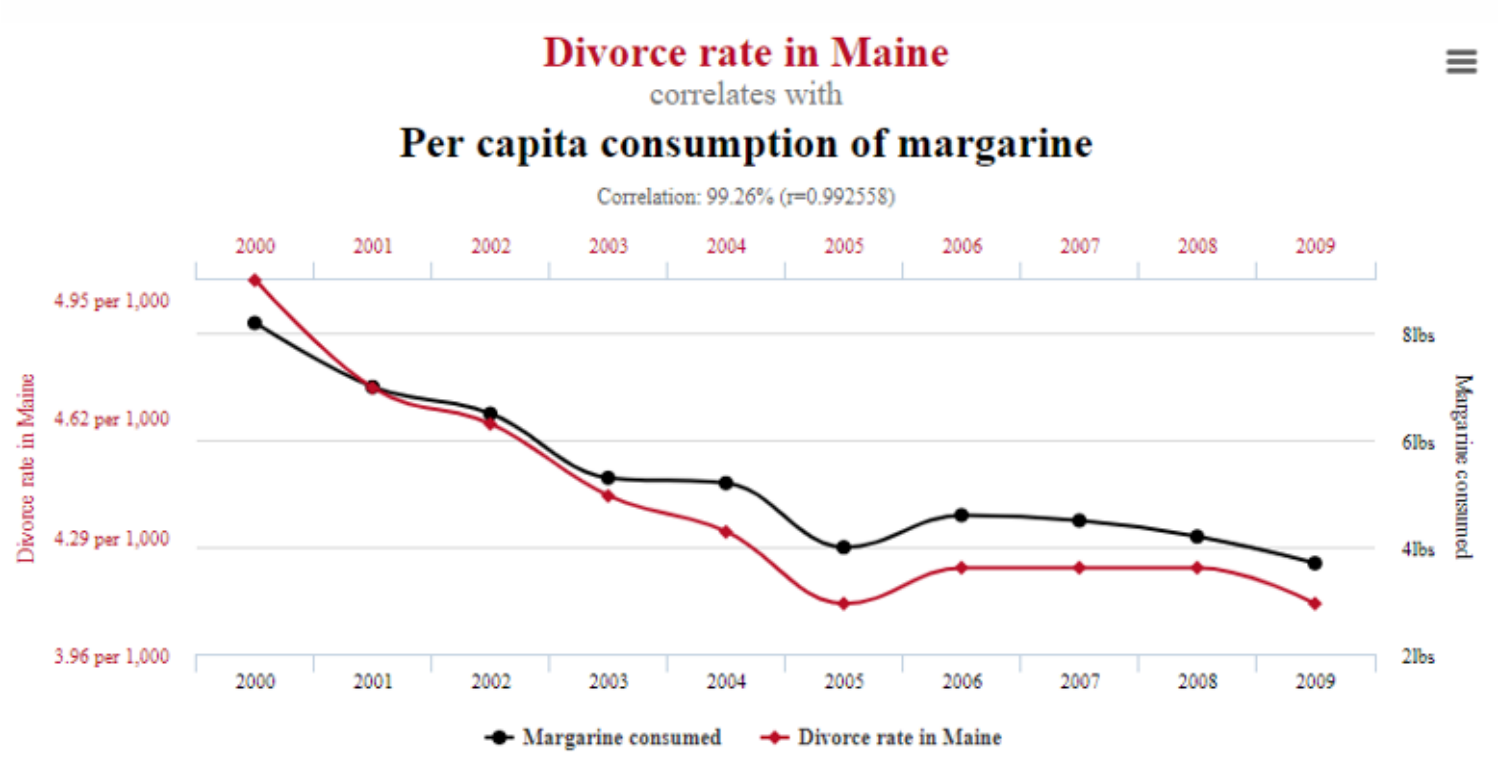
Random sampling vs. random assignment

	Random assignment	No random assignment	
Random sampling	Causal and generalizable	Not causal, but generalizable	Generalizable
No random sampling	Causal, but not generalizable	Neither causal nor generalizable	Not generalizable
	Causal	Not causal	

What do you think Pfizer did in their trials for the COVID-19 vaccine development?

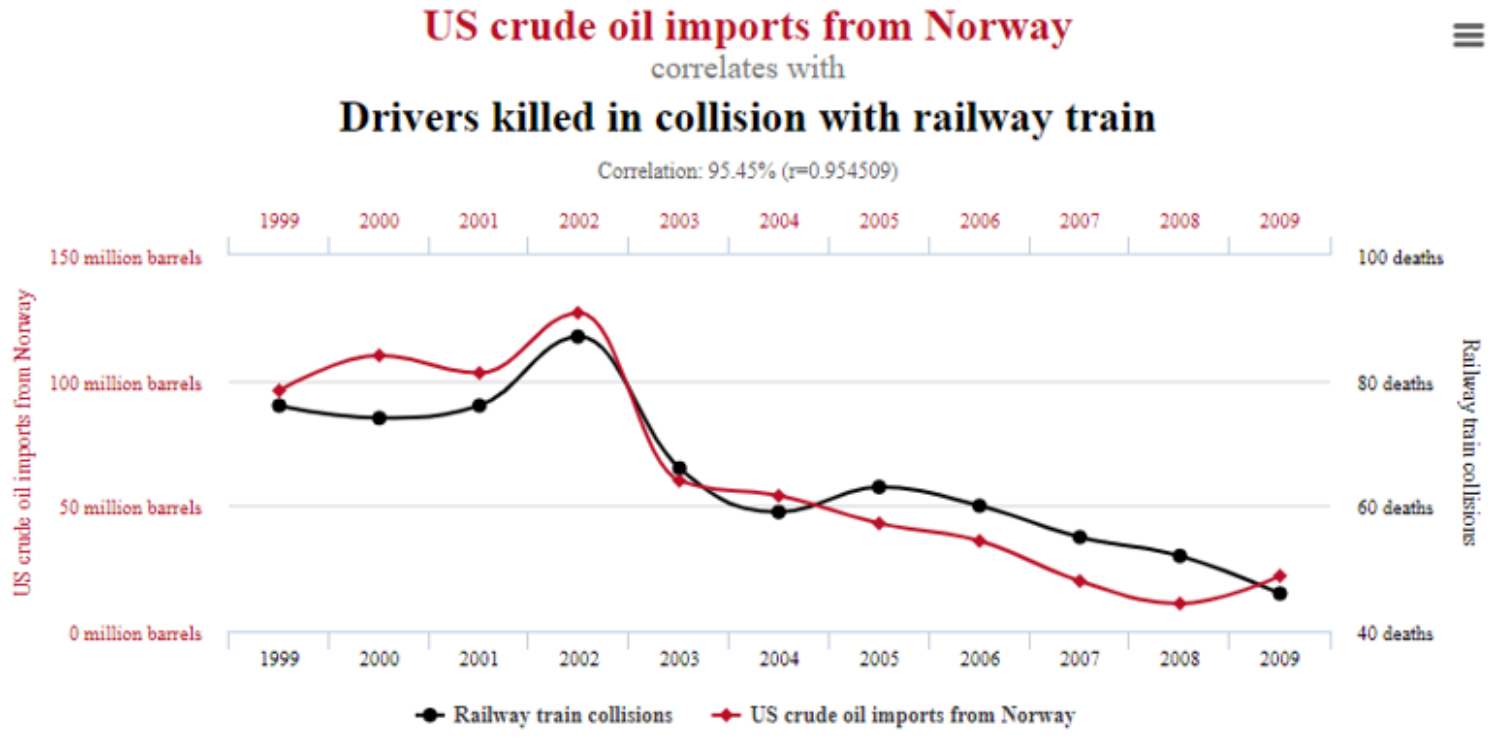
Pitfalls

"Lucky coincidences"



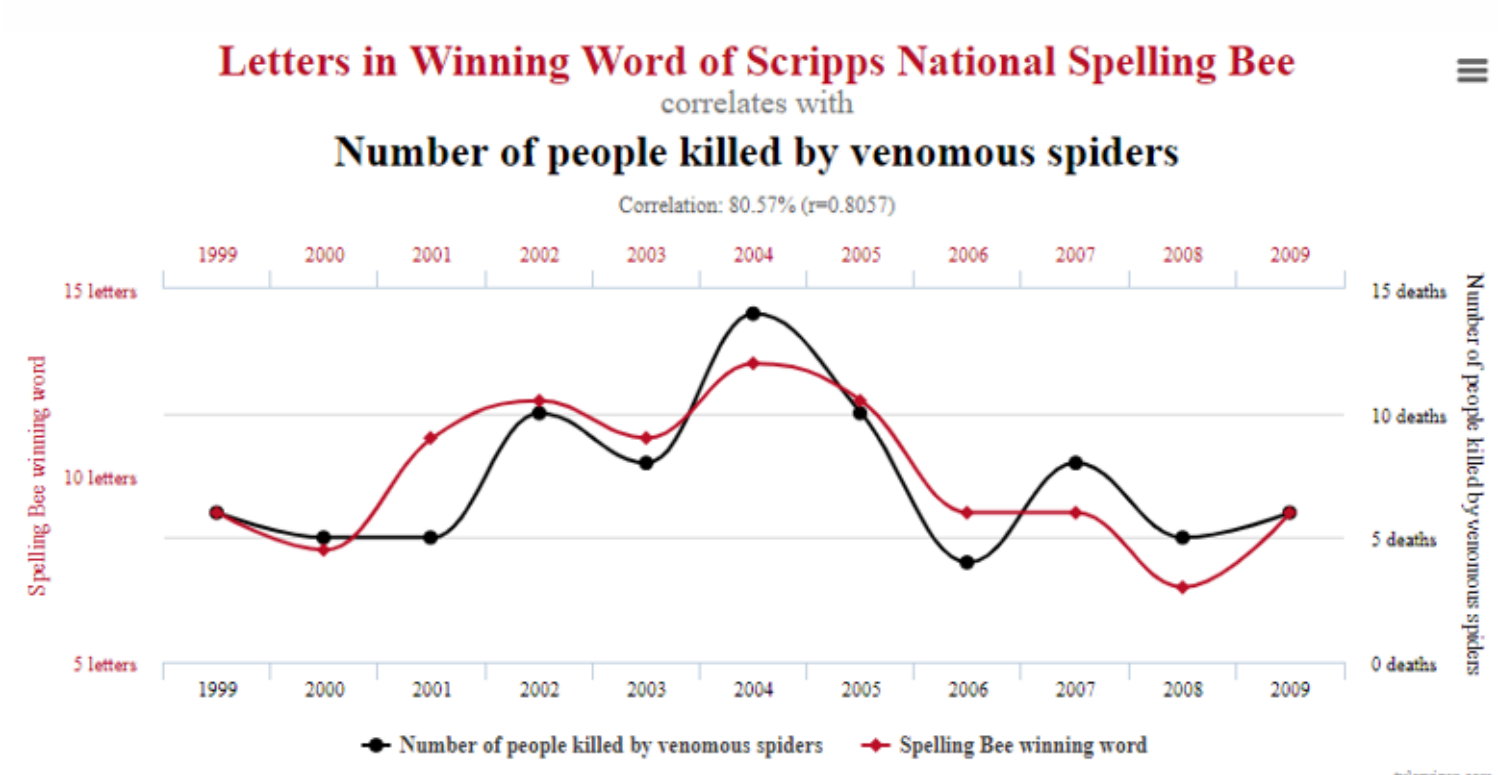
Source: Tyler Vigen's site of spurious correlations:

"Lucky coincidences"



Source: Tyler Vigen's site of spurious correlations:

"Lucky coincidences"



Source: Tyler Vigen's site of spurious correlations:

Confounding variables

A **confounding** variable is an an extraneous variable that affects both the explanatory and the response variable, and makes it seem like there is a relationship between them.

Identify the confounding variable in each of the following statements:

1. As the amount of ice cream sales increases, the number of shark attacks also increases.

2. The higher the number of firefighters at a fire is, the greater the amount of damage caused by that fire.

3. Taller children are better at both reading and math compared to shorter children.

One method to justify making causal conclusions from observational studies is to exhaust the search for confounding variables, there is no guarantee that all confounding variables can be examined or measured. Therefore, it is best to only discuss associations between variables from observational studies.

Getting some practice

- Create your personal private repository by clicking https://classroom.github.com/a/SW9XSOkk