Collecting Data: Sampling

SECTION 1.2

- Sample versus Population
- Statistical Inference
- Sampling Bias
- Simple Random Sample
- Other Sources of Bias

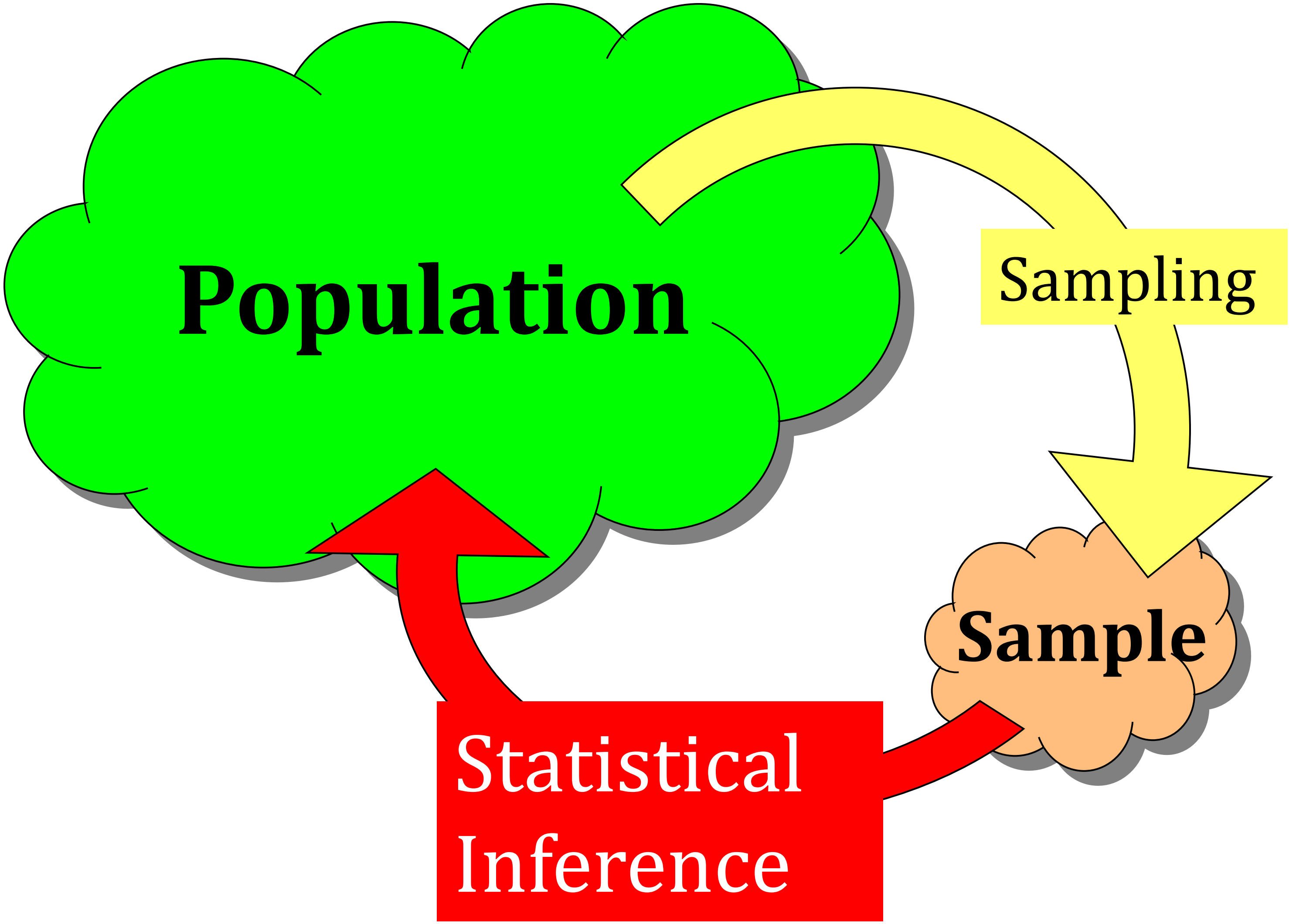
Sample versus Population

A ***population*** includes all individuals or objects of interest.

A ***sample*** is all the cases that we have collected data on (a subset of the population).

Statistical inference is the process of using data from a sample to gain information about the population.

The Big Picture



Dewey Defeats Truman?

- The paper was published before the conclusion of the 1948 presidential election, and was based on the results of a large telephone poll
- Harry S. Truman won the election
- What went wrong?

Sampling Bias

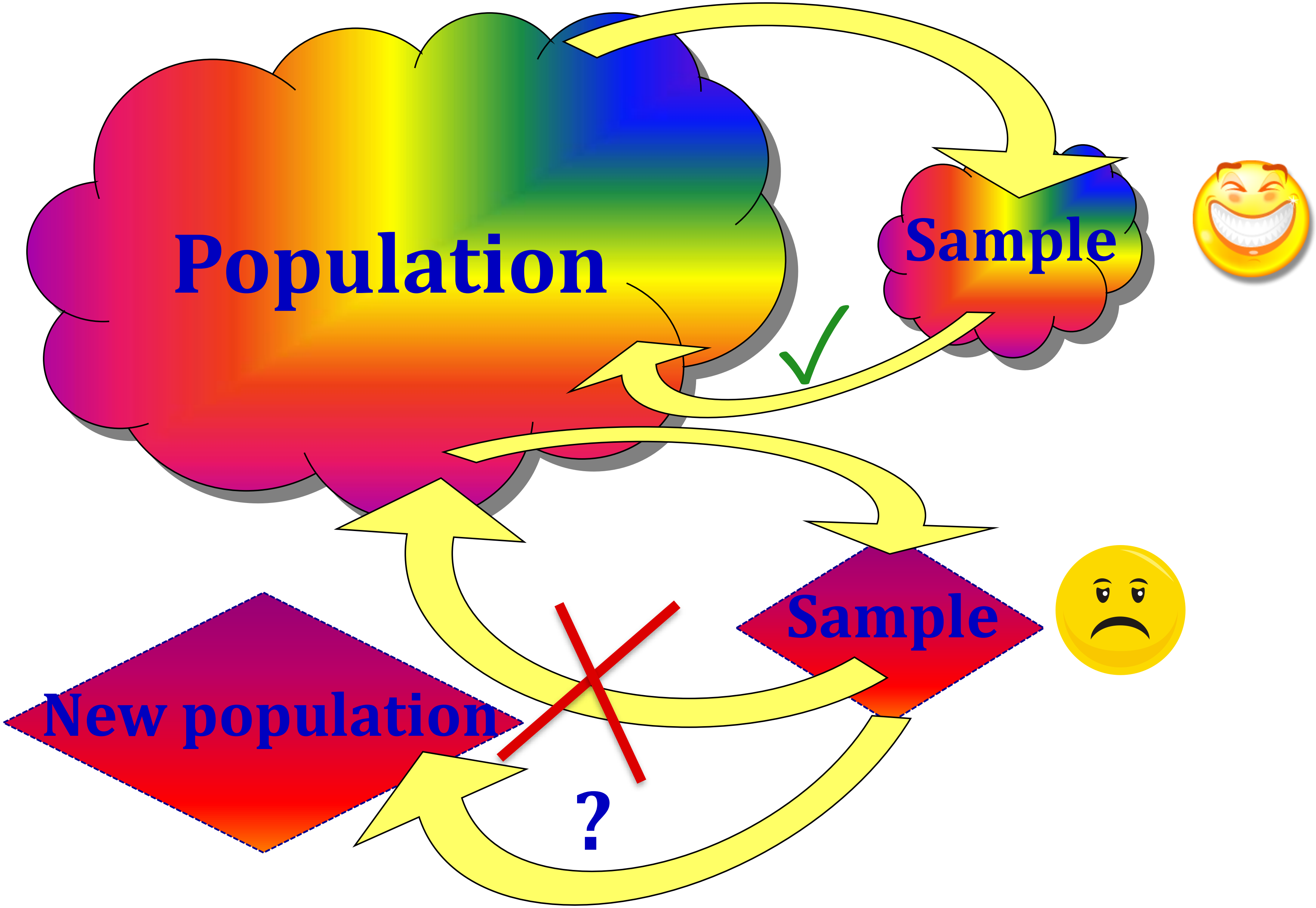
Sampling bias occurs when the method of selecting a sample causes the sample to differ from the population in some relevant way.

- If sampling bias exists, we cannot trust generalizations from the sample to the population

Scope of Inference

- Inference can only extend to a population that "looks like" the sample
- Examples:
  - Drug testing on animals
  - Agricultural experiment in a certain location
  - 23andme genetic data
  - Tracking disease by those who visit doctors

Sampling



Population

Sample

Sample

New population

?

How does this apply to you?

Come up with an example of how sampling bias might arise in your field or where the sample clearly differs from the population:

Random Sampling

- How do we get a sample that looks like the population???

Take a RANDOM sample!

- A random sample will resemble the population!
- Random sampling avoids sampling bias!

Bowl of Soup Analogy

Think of tasting a bowl of soup…

- Population = entire bowl of soup
- Sample = whatever is in your tasting bites
- If you take bites non-randomly from the soup (if you stab with a fork, or prefer noodles to vegetables), you may not get a very accurate representation of the soup
- If you take bites at random, only a few bites can give you a very good idea for the overall taste of the soup

Simple Random Sample

In a ***simple random sample***, each unit of the population has the same chance of being selected, regardless of the other units chosen for the sample

- Like drawing names out of a hat
- More complicated random sampling schemes exist, but will not be covered in this course

Random Sample Examples

- Random sample of Americans to infer the proportion of Americans who are obese
- Random sample of plants in a field to infer average yield per plant in that field
- Random sample of patients to measure satisfaction
- Random sample of square meters in a forest to measure tree blight
- Note: Won't get answer exactly (how close you get depends on sample size), but will avoid bias!

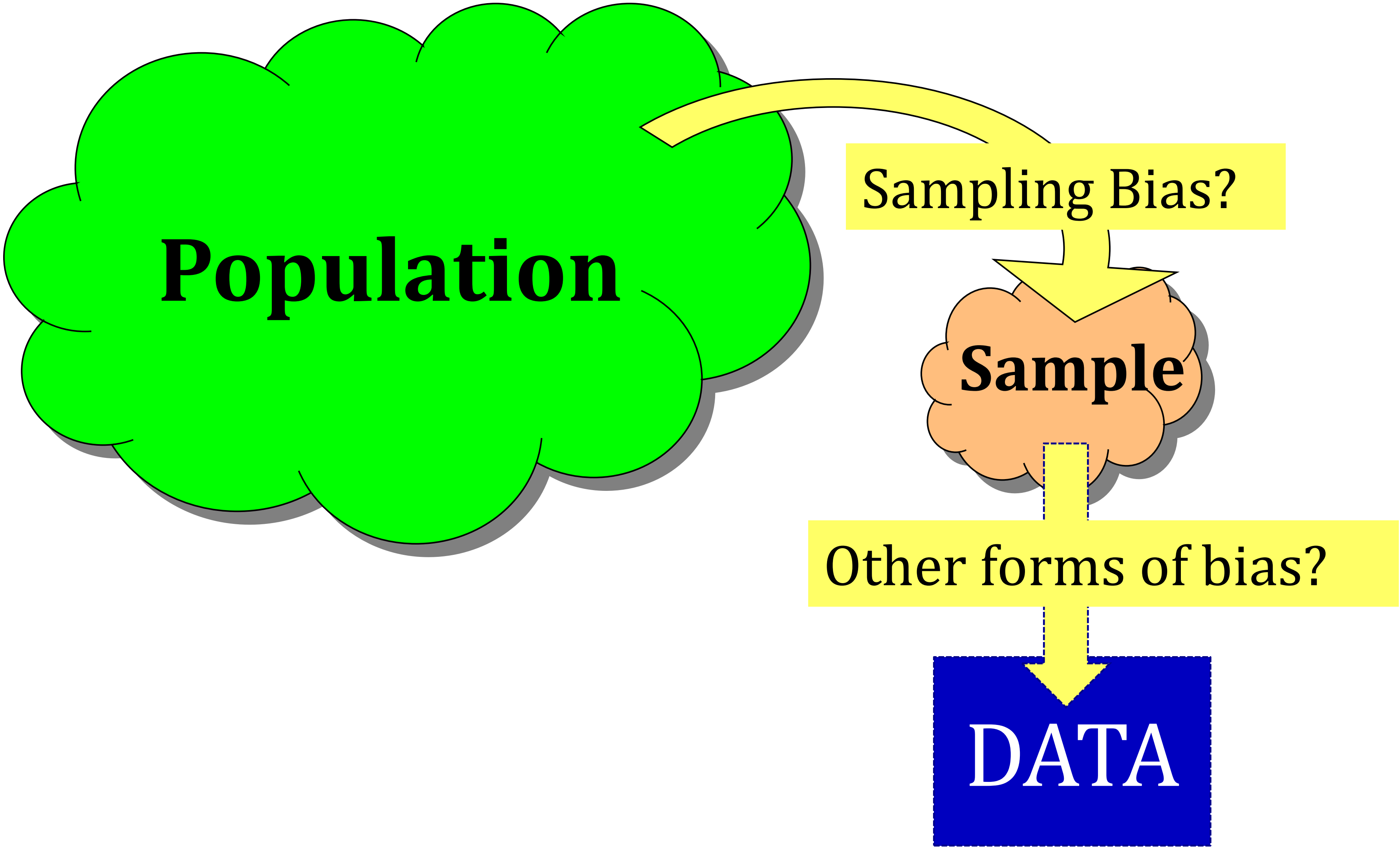
How does this apply to you?

Come up with a situation where random sampling could be feasible and help to answer a question in your field:

Reality

- Unfortunately, obtaining random samples are often not possible
  - Clinical trials – can’t force someone to participate
  - Random sample of everyone with a particular disease?
  - Random sample of all humans?
  - Random sample of any species of animal or plant?
  - Some people simply won’t respond
- May have to alter the population to be feasible
- In reality, just watch out for ways in which the sample may differ from the population

Data Collection and Bias



Population

Sampling Bias?

Sample

Other forms of bias?

DATA

Other Forms of Bias

- Other forms of bias to watch out for in data collection:
  - Inaccurate responses (self-reporting, measurement error)
  - Influential question wording
  - Context
  - Many other possibilities – examine the specifics of each study!

Self-Reported Values

- NHANES (random sample!) got self-reported height and weight and measured height and weight
  - ○ Men…
  - ○ Women…

Merrill, R.M. and Richardson, J.S. (2009). Validity of Self-Reported Height, Weight, and Body Mass Index: Findings From the National Health and Nutrition Examination Survey, 2001 – 2006. *Preventing Chronic Disease: Public Health Research, Practice, and Policy*. **6**(4).

Question Wording

- A random sample was asked: "Should there be a tax cut, or should money be used to fund new government programs?"

- A different random sample was asked: "Should there be a tax cut, or should money be spent on programs for education, the environment, health care, crime-fighting, and military defense?"

Context

- Ann Landers column asked readers

“If you had it to do over again, would you have children?

- The first request for data contained a letter from a young couple which listed worries about parenting and various reasons not to have kids

- The second request for data was in response to this number, in which Ann wrote how she was “stunned, disturbed, and just plain flummoxed”

Having Children

If we were to run the question all by itself in the newspaper with a request for responses, could we trust the results?

(a) Yes

(b) No

Having Children

Newsday conducted a random sample of all US adults, and asked them the same question, without any additional leading material

Do you think the true proportion of US parents who are happy they had children is close to 91%?

a) Yes

b) No

2016 Election Polls

4 Possible Reasons The Polls Got It So Wrong This Year (www.npr.org, 11/16/16):

- The national polls weren't that off (Clinton won the popular vote, as predicted)
- Some people just don't answer the phone
- Did people lie to pollsters?
- It's hard to capture enthusiasm (to predict who will vote)

Summary

Always think critically about how the data were collected, and recognize that not all forms of data collection lead to valid inferences

Random sampling is a powerful way to avoid sampling bias