# Introduction

Patrick Breheny

January 16

# What is statistics?

- Statistics is the science of learning from experience
- In principle, people do this every day of their lives, and should be really good at it . . .
- . . . but we're not

# Limitations of human reasoning

- Human beings are not natural statisticians
- We are not good at picking out patterns from a sea of noisy data
- On the flip side, we are *too good* at picking out non-existent patterns from small numbers of observations
- We also find it difficult to sort out the effects of multiple factors occurring simultaneously
- Finally, we are subject to all sorts of biases depending on our personalities, emotions, and past experiences

# Biostatistics

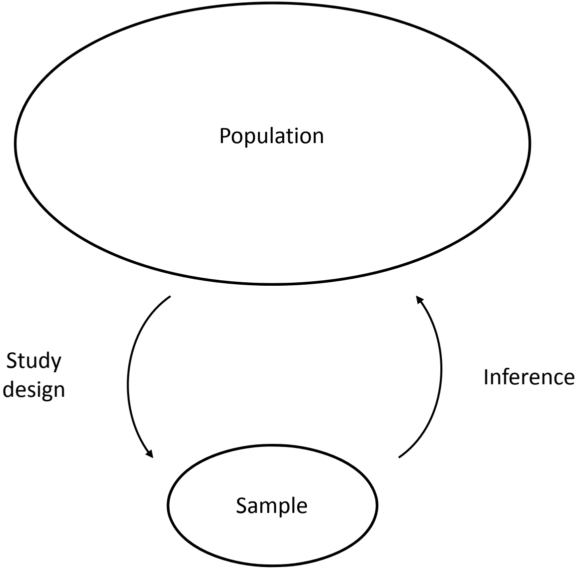
- In medicine, public health, and the biological sciences, we must often make decisions in the presence of uncertainty:
  - Which drug should a doctor prescribe to treat an illness?
  - An individual has a certain genetic mutation; what are the chances that she will develop breast cancer?
  - Does a certain pesticide cause neurological defects? Should it be banned?
- These questions are too important to be left to opinion, superstition, and conjecture, which is why there has been a tremendous push for objective, *evidence-based* decision making in medicine and public health in the past several decades

# Why do we need biostatistics?

- Statistics is the science which allows us to make these decisions
- Statistics is particularly important in research concerning human health, for several reasons:
  - Humans are incredibly diverse and variable
  - Humans are expensive to perform research upon
  - There is a moral imperative to make decisions on potentially life-saving therapies as fast as possible
- For these reasons, Biostatistics has emerged as an important field within statistics, and has become an important set of skills in fields such as medicine, biology, and public health

# Terms

- Scientists want to make generalizations about classes of people on the basis of their findings
- The class of people that they are trying to make generalizations about is called the *population*
- It is impractical to study the entire population, so people study only a small portion of it called the *sample*
- The researchers then make generalizations about the entire population based on studying the sample; this process is called *inference*


Population
Study design
Inference
Sample

# The three basic questions in statistics

- How should I collect my data?
- How should I describe and summarize the data that I've collected?
- What does my data tell me about the way that the world works (inference)?

# Parameters

- Specifically, there is some numerical fact about the population that the investigator is interested in, such as the percent of children who are obese or the average reduction in cholesterol upon taking a certain drug
- These numbers that describe the population are called *parameters*
- Parameters cannot be observed directly; they can only be *estimated* from a sample
- An *estimate* (or *statistic*) is a number that can be computed from a sample

# Example: Childhood obesity

- Many sources will just refer to childhood obesity rates as if they are known facts (i.e., parameters): example
- If you read the fine print, however, you will learn that these are in fact estimates coming from a sample – in this case, that sample is called the National Health and Nutrition Examination Survey, and is described in more detail in publications such as this one: Hales2018
- In particular, note that we are inferring what percent of the population ($\approx 75$ million U.S. children) are obese on the basis of a sample of just 3340 children . . . can we really justify this?

# How good are our estimates?

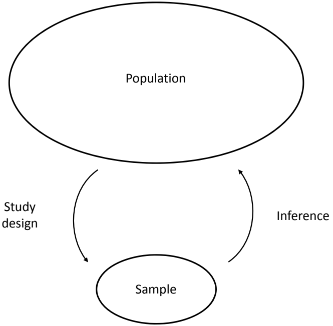
- As a reminder,
  - Estimates are what investigators know
  - Parameters are what investigators want to know
- We would like to know whether or not our estimate is measuring the parameter of interest well
- There are two major issues:
  - On average, does our estimate tend to be centered around the right answer, or is it *biased*?
  - How much *variability* is there likely to be in our estimate?
- The difference between these two issues may not be obvious at first, but the key difference is whether the error is systematic

# Comparing two groups

- These concepts are just as relevant when comparing two things: for example, whether patients suffering from a certain condition benefit from surgery or not
- Suppose that the success of the surgery depends greatly on the skill of the doctor that performs the operation:
    - In general (assuming a mix of skill in both surgery and non-surgery groups), this would increase the variability of our results
    - However, if all of the worst doctors are in the surgery group, this would be bias

# Summary

- We've discussed three important pairs of concepts:
    - Population / Sample
    - Parameter / Estimate
    - Bias / Variability
- The conceptual framework of statistics is represented in this figure (slide 7):


Population
Study design
Inference
Sample