Intro to Statistical Inference

M. Drew LaMar
August 31, 2016

> "To try to make a model of an atom by studying its spectrum is like trying to make a model of a grand piano by listening to the noise it makes when thrown downstairs."
>
> - Anonymous

Introduction to Quantitative Biology, Fall 2016

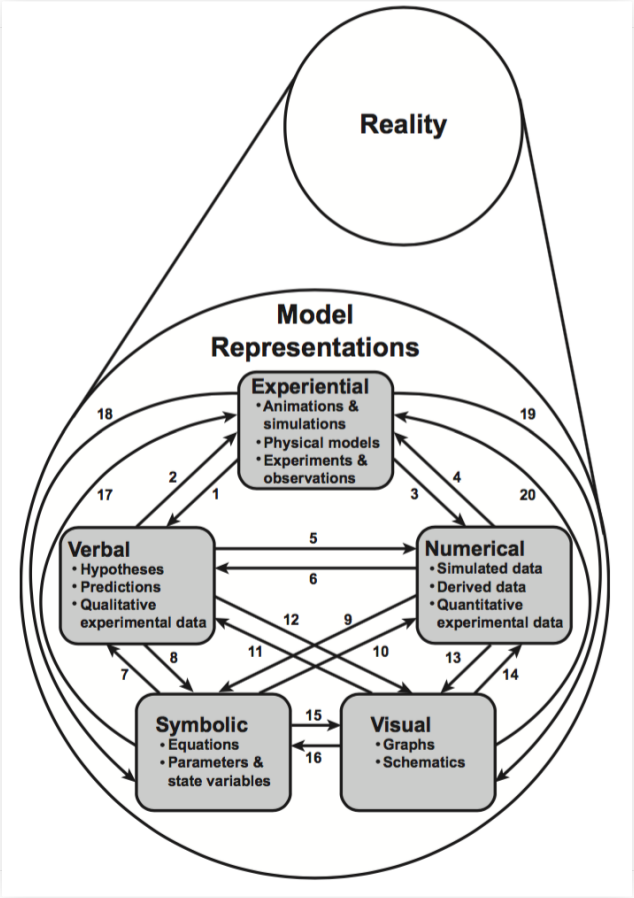
Class announcements

- Reading assignment for Friday: **OpenIntro Stats, Chapter 3: Distributions of random variables (optional - NO QUIZ)**
  - 3.1: Normal distribution
  - 3.2: Evaluating the normal distribution
- Homework Assignment #1 (due Monday, September 5):
  - OpenStats, Chapter 4: 4.6.1 Variability in estimates (p. 203) - #4.3, 4.5
  - …more coming on Friday

Model vs Reality

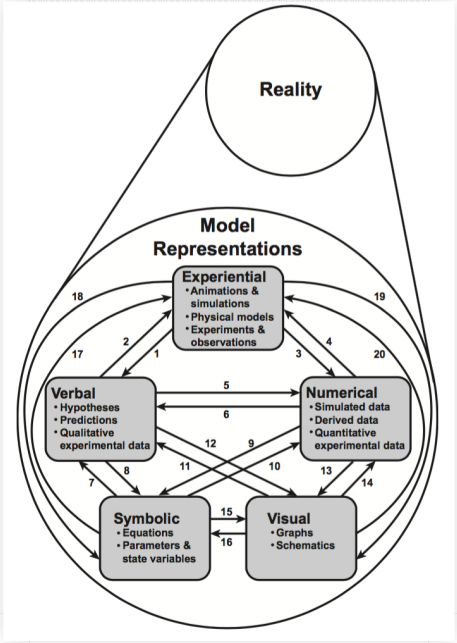
Model Representations

Multiple representations are important!!!



What is modeling?

Definition: ***Modeling*** is 1) the process of moving from observations of the real world to a model, 2) moving from one model representation to another model representation, or 3) comparing different models.



Why models and modeling are useful

Discuss: Why are models and modeling useful?

Answer: Models can be used for ***prediction***, ***explanation*** and ***understanding***.

Note: *You are already constructing models and doing modeling!!!*

In this course, we will work to become proficient in understanding, analyzing, and creating models - in particular *mathematical models*.

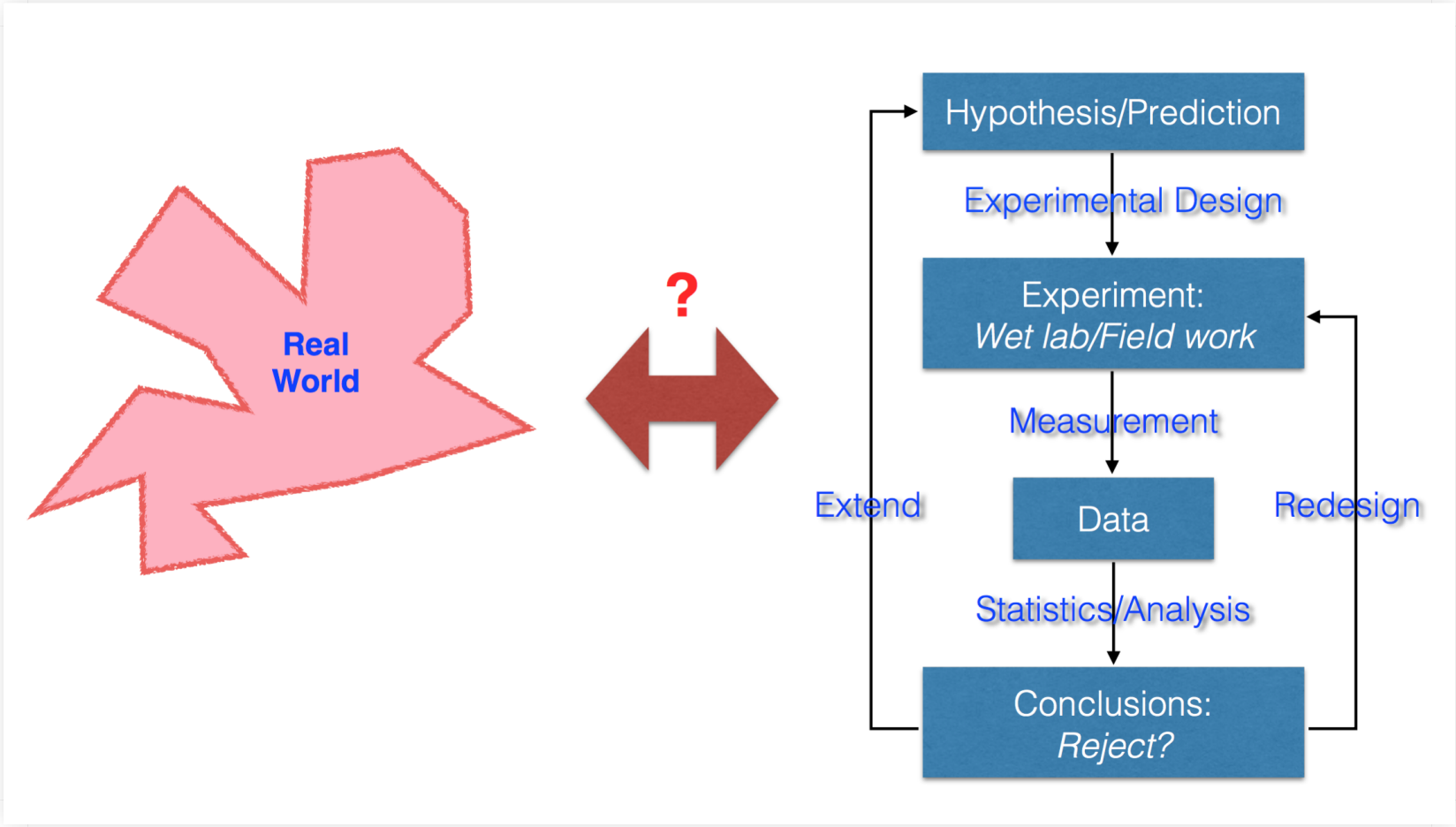
Process Of Science (POS)

Process Of Modeling (POM)

Why models and modeling are useful (little more sophisticated)

- models can make accurate predictions,
- models can generate causal explanations.
- simple, unrealistic models help scientists explore complex systems,
- models can be used to explore unknown possibilities,
- models can lead to the development of conceptual frameworks,
- models help expose assumptions,
- models help define expectations,
- models help us to devise new tests.

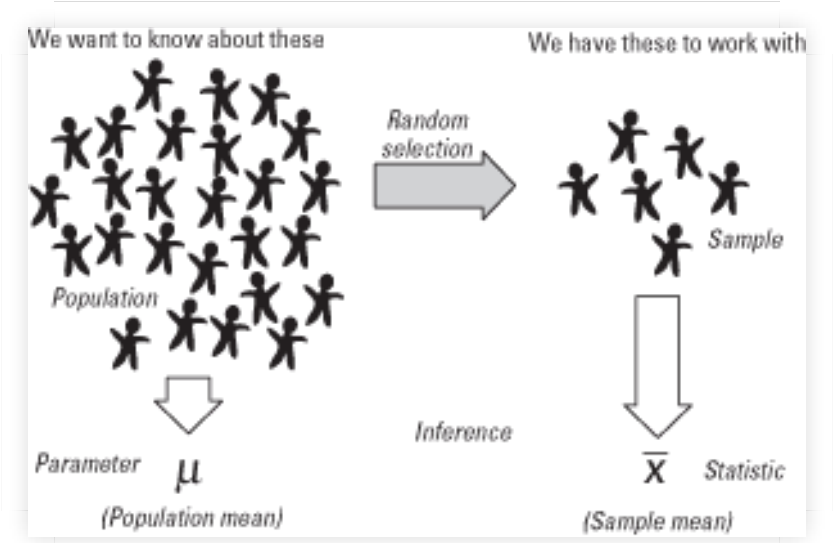
Introduction to Statistical Inference



Populations vs Samples

Definition: A ***parameter*** is a quantity describing a population, whereas an ***estimate*** or ***statistic*** is a related quantity calculated from a sample.

Parameter examples: Averages, proportions, measures of variation, and measures of relationship



What is statistics?

Statistics is a technology that describes and measures aspects of nature from samples.

Statistics lets us *quantify the uncertainty* of these measures.

Statistics makes it possible to determine the likely magnitude of measurements departure from the "truth".

Statistics is about *estimation*, the process of inferring an unknown quantity of a target population using sample data.

What is statistics?

The two sides of the statistical coin:

- Parameter estimation
- Hypothesis testing

Definition: A ***statistical hypothesis*** is a specific claim regarding a population parameter.

Definition: ***Hypothesis testing*** uses data to evaluate evidence for or against statistical hypotheses.

What is statistics? Parameter estimation

The two sides of the statistical coin:

- **Parameter estimation**
- Hypothesis testing

Example: A trapping study measures the rate of fruit fall in forest clear-cuts.

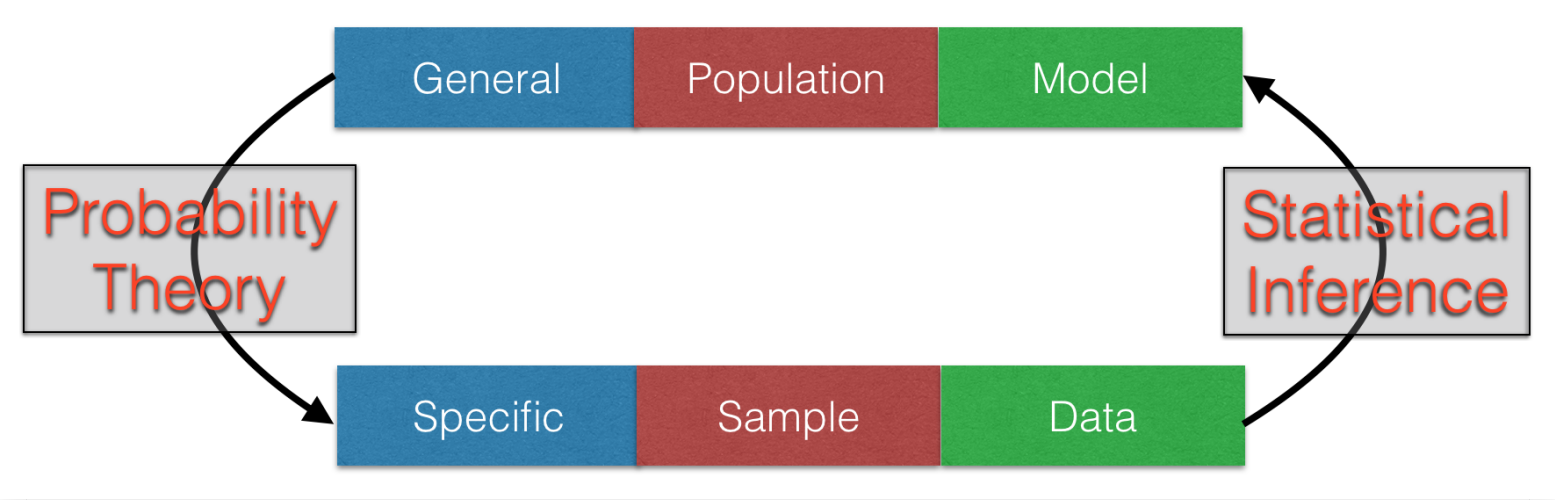
What is statistics? Hypothesis testing

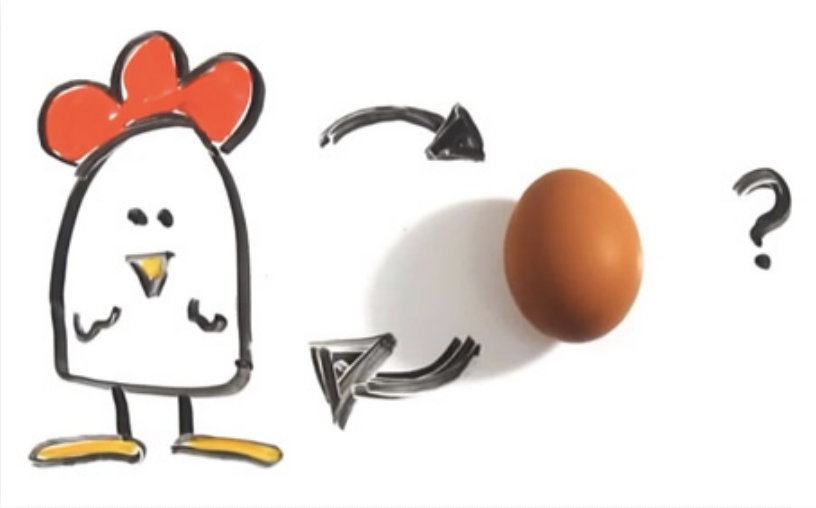
The two sides of the statistical coin:

- Parameter estimation
- **Hypothesis testing**

Example: A clinical trial is carried out to determine whether taking large doses of vitamin C benefits health of advanced cancer patients.

What is probability?





Probability comes first!

…well, most of the time.

- Many statistical techniques require assumptions about where your data is coming from (i.e. properties of the *population*)
- In other words, an assumed *probability model* describes the population
- Statistical techniques that are based on probability models are called **parametric techniques**, while those that are not are called **non-parametric techniques**.

Data as Information

For your question, there is **desired** and **undesired** information in your data.

Goals:

- Get *accurate* information by reducing ***bias***
- Get *precise* information by reducing ***sampling error*** due to ***random variation*** (increase signal-to-noise ratio)

Definition: ***Bias*** is a systematic discrepancy between the estimates we would obtain, if we could sample a population again and again, and the true population characteristic.

Data as Information

For your question, there is **desired** and **undesired** information in your data.

Goals:

- Get *accurate* information by reducing ***bias***
- Get *precise* information by reducing ***sampling error*** due to ***random variation*** (increase signal-to-noise ratio)

Definition: ***Sampling error*** is the difference between an estimate and the population parameter being estimated caused by chance.

Data as Information

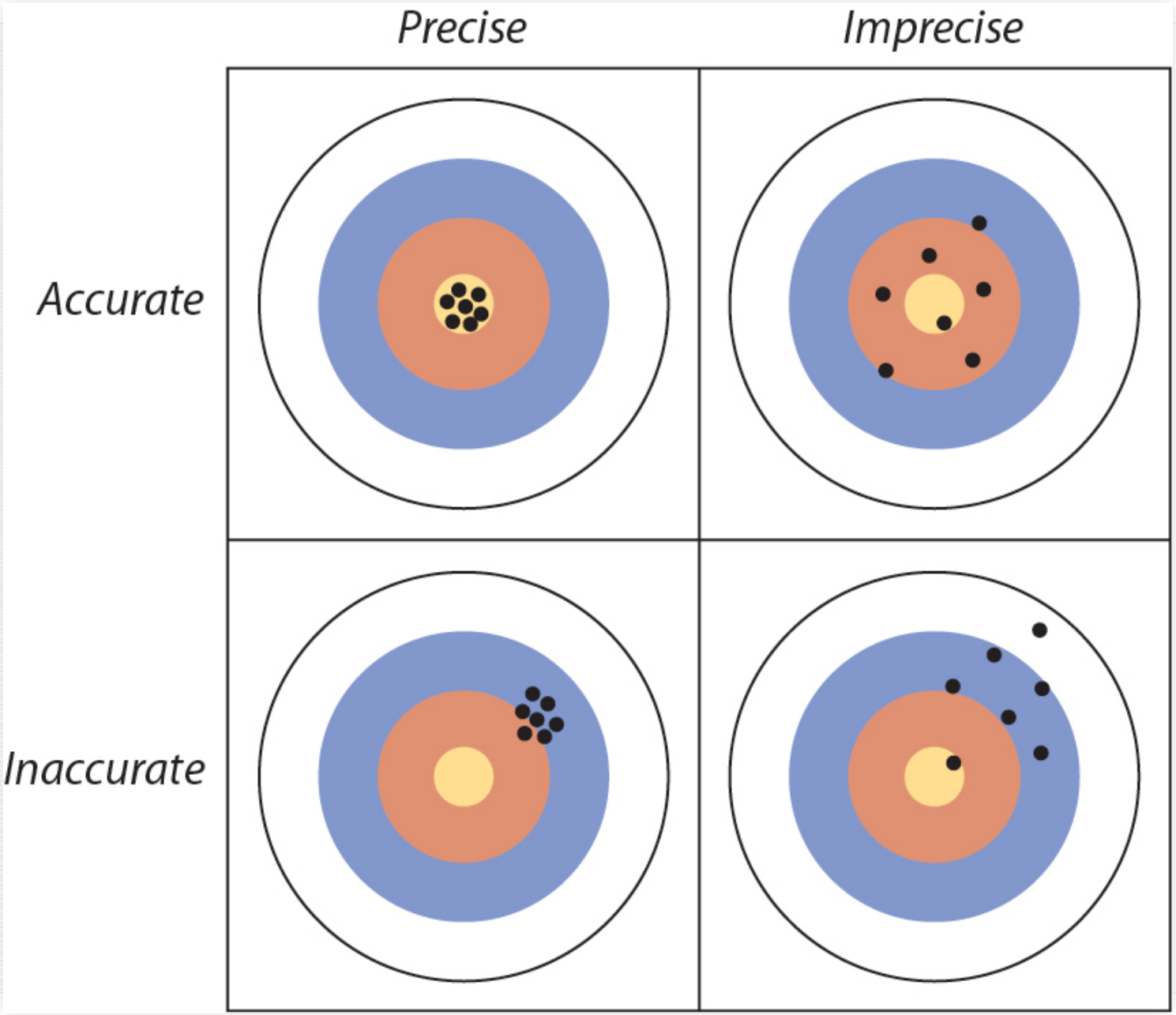
For your question, there is **desired** and **undesired** information in your data.

Goals:

- Isolate *desired* information by *reducing* or *controlling* for ***confounding factors*** (i.e. *undesired* information)

> "The aim … is to provide a clear and rigorous basis for determining when a causal ordering can be said to hold between two variables or groups of variables in a model…"
>
> - H. Simon

Precision vs Accuracy



Random sampling

The main assumptions of all statistical techniques is that your data come from a *random sample*.

Definition: In a ***random sample***, each member of a population has an equal and independent chance of being selected.

Random sampling

1. minimizes bias (**equal**) and
2. makes it possible to measure the amount of (*quantify precision*) sampling error (**independent**)