

Introduction: The Population is the Point

PSCI 1801 · Class Handout

Dr. Marc Trussler

This handout is designed as an in-class supplement to the textbook. The material contained on this handout does not represent the full set of information you need to know for the exams. Exclusively studying off of this handout will not be sufficient for the exams.

What Statistics Is Actually About

The point of statistics is to make inferences about a population using a sample of data.

Describing the sample you have in front of you is not the goal. The goal is what your sample tells you about the population.

What defines statistics: making conclusions about populations from a sample.

What “Population” Means

A vague term used for a wide range of things:

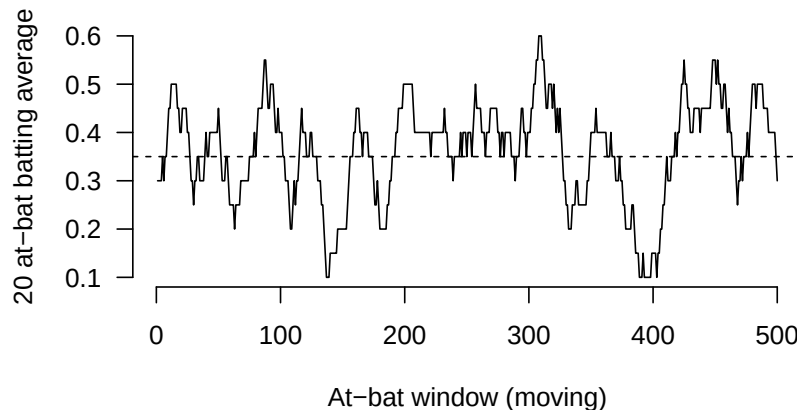
- **Obvious cases:** adult U.S. population (sampled by a poll of 8000)
- **Business earning potential:** a few months of finances = a sample from a “population” that is the business’s true earning distribution
- **Baseball hitting ability:** a stretch of at-bats = a sample from the player’s true hitting distribution

- **Data-generating process (DGP):** the underlying process that produced your data, whether or not there's a well-defined "population" behind it

Every observation is a draw from some DGP. Statistics is about inferring properties of the DGP from a sample.

The Central Insight (Baseball Example)

Suppose a player's true batting average is 0.350. Take a "sample" of 20 at-bats (about 4 games). What do we see?



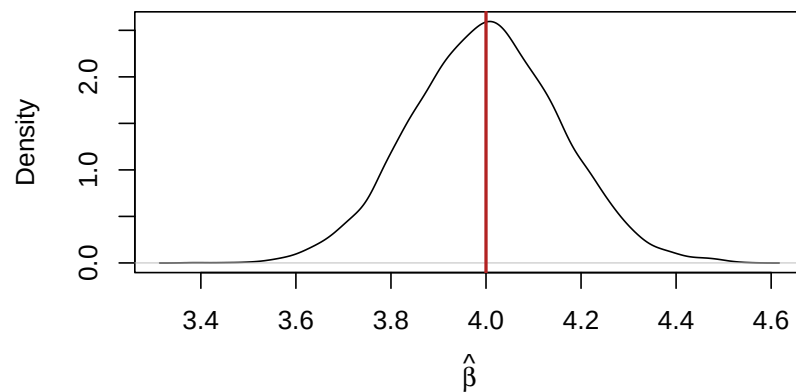
Even though the underlying truth is fixed at 0.350, 20-at-bat samples swing between 0.100 and 0.600 by pure chance. If we take the sample at face value we'll be wrong constantly.

We need tools to measure and express how much confidence to place in what the sample tells us.

Where This Course Is Headed

A quick preview:

5000 regression coefficients from the same population



If we could re-run our sampling process thousands of times, the distribution of any statistic we care about (mean, difference, regression coefficient) has a predictable shape — usually normal, with a knowable width.

The remarkable thing: the shape of this “sampling distribution” can be inferred from a single sample. That’s what standard errors are.

The Main Question of the Course

“If there is nothing going on — if there is no relationship in the population — how likely is it to observe the pattern I’m seeing in my sample?”

Every hypothesis test, every confidence interval, every p-value is a different formalization of that one question.

If the answer is “very unlikely,” we conclude something real is happening. If the answer is “not unlikely,” we can’t rule out that our result is just noise.