Study Design and Bias

Grinnell College

October 3, 2025

What have we done

Until this point, we have concerned ourselves with *descriptive statistics*

- Plots
- Tables
- Numerical Summaries

These have all been tools to help us understand and describe characteristics of our *sample*

Populations and Samples

A **population** in the context of statistics is an unambiguous and bounded set of items or events about which we may wish to make a statement

- Students at Grinnell College
- Iowa song birds
- Children with cochlear implants
- Vehicles made by a particular manufacturer

A **sample** is a smaller subgroup of a population

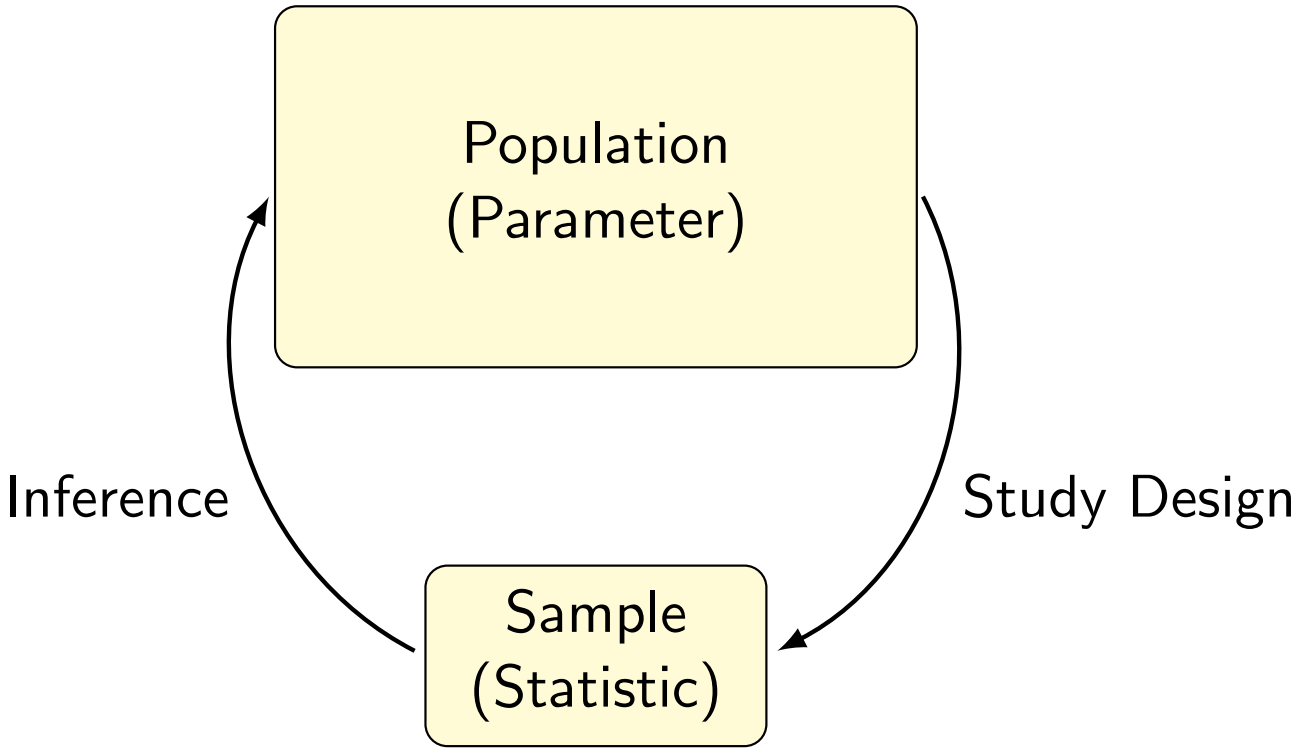
Inference

Statistical inference addresses the question: "How reliably can I expect trends in my sample to reflect what is true about the population"

A good starting point is to find a *point estimate*, or a statistic, to estimate the parameter in question

If a sample is **representative**, our point estimate should be *close* to the parameter we wish to know

The Statistical Framework



Today

1. Types of studies
2. Types of sampling
3. Types of error

Types of Studies

Experiments are studies that involve manipulating the *treatment* or *exposure* that a participant receives:

- Individuals are *randomly assigned* to different treatments (i.e., drug/placebo, prisoner/guard)
- The participants' *responses* to the treatment are measured
- Can be used to establish causal relationships

Observational studies are studies that do not involve manipulating explanatory variables:

- We simply "observe" what is already there (i.e., children living near airports)
- No assignments to groups are made
- Causal relationship cannot be established (only association)

Experimental Studies

Experimental studies, in particular the double-blind randomized clinical trial, is considered the "gold-standard" of study design

- Participants assigned groups at random, with groups balanced to be as similar as possible
- Neither investigator nor subject knows which groups they are in
- *Intent to Treat (ITT)* analyzes results based on treatment assigned rather than treatment received

Intent to Treat (Example)

The Coronary Drug Project Research Group published an article in the *New England Journal of Medicine* (1980) describing a randomized controlled double-blind experiment involving the drug clofibrate, which reduces the level of cholesterol in the blood

	Clofibrate	
	Number	Deaths
Adherers	708	15%
Nonadherers	357	25%
Total	1,103	20%

Subjects who took more than 80% of their prescribed medicine were called "adherers"

Clofibrate and Placebo Results (Example)

	Clofibrate		Placebo	
	Number	Deaths	Number	Deaths
Adherers	708	15%	1,813	15%
Nonadherers	357	25%	882	28%
Total	1,103	20%	2,789	21%

- Taking into account the placebo results as well, clofibrate no longer looks effective
- One possibility is that adherers are more concerned with their health, and take better care of themselves in general
- Take-home message: comparing subjects *as they were randomized* is the only completely valid way of carrying out a controlled experiment; all other comparisons are subject to confounding and bias

Types of Observational Studies

Case-control: Two existing groups are collected based on outcome and compared on the basis of a supposed causal attribute. Basically a snapshot in time. For example, collecting 750 individuals with and without lung cancer and asking smoking status

Longitudinal: Also called prospective. Here, participants are collected based on some exposure and then followed for a period of time, prior to outcomes being known. For example, the ABCD study is collecting brain scans and diagnostics on 10,000 US children. We can partition groups based on pre-term birth status and evaluate outcomes over time

Retrospective: Similar to a longitudinal, but is done following the outcome in question. This is most common in cases with rare outcomes. To study exposures related to Parkinson's, for example, a very large prospective study would be needed to ensure enough positive outcomes would be collected

How is our sample collected?

Sampling Method

We need to randomly select observations from our population to study

- Important to have a *representative sample* so that our results will generlize
- Must balance with concerns of logistics/feasability

Sample Size

How many observations are we going to study?

- More observations means more data
- Controls major source of variability
- Marginal benefit descreases as more included (though costs continue to rise)

Sampling Methods

Census – We include the entire population in our study

- Pros: Have exact answers
- Cons: Difficult, expensive, no statistics :(

Sampling Methods

Convenience sampling – select all cases from our target population that are easily accessible

- Pros: easy to collect data
- Cons: high potential for sampling bias

Sampling Methods

Simple random sampling – randomly select cases from target population

- Pros: eliminates sampling bias
- Cons: difficult to execute

Sampling Methods

Stratified or clustered random sampling – randomly select cases separately from different segments of population

- Pros: low potential for sampling bias, more flexible than random sampling
- Cons: data analysis complicated, expensive

Samples and Populations

Ultimately, our entire conversation on sampling methods is in pursuit of having a *sample* that resembles our *population*

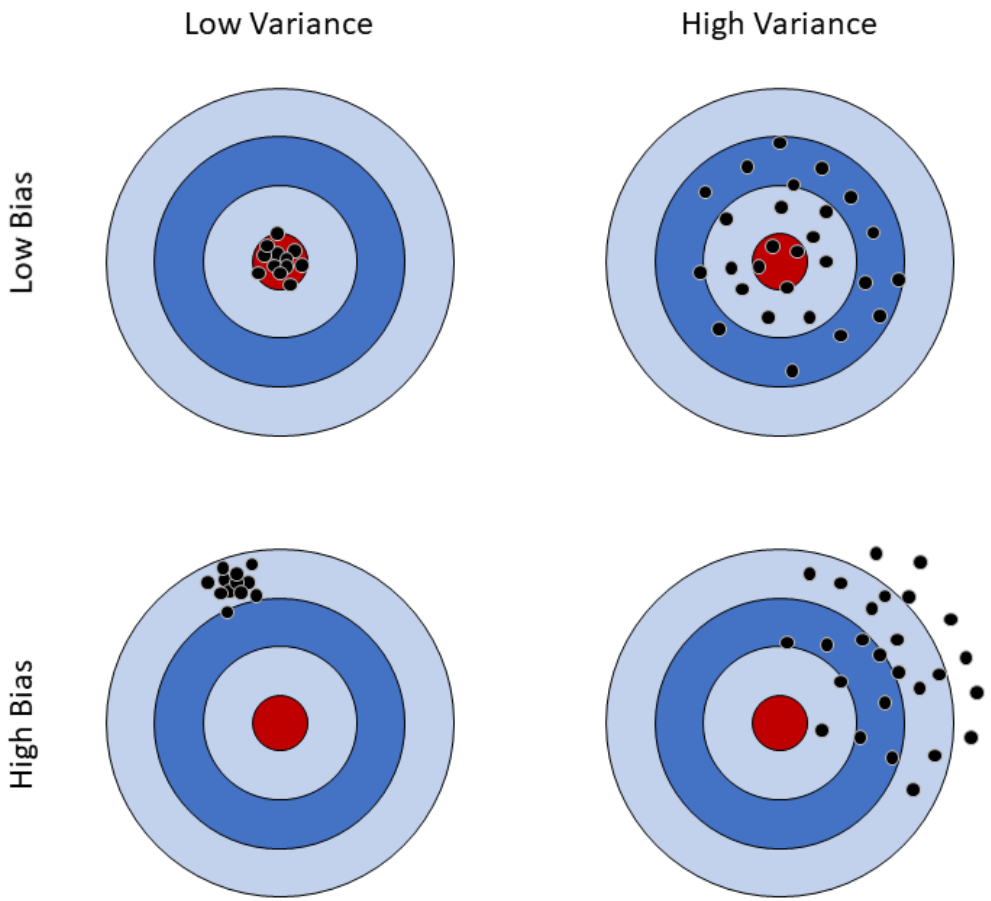
Sources of Error

Error can broadly be defined as the degree to which our sample statistic differs from our population parameter

Thre are two main reasons why this may occur:

1. **Sampling Bias** – A systemic flaw in how the sample was collected
2. **Sampling Variability** – Differences between samples due to *random chance*

Bias/Variability



Types of Bias

Bias describes ways in which our sample may be *non-representative* of our population

Selection Bias – describes situation in which the method whereby observations are sampled may be associated with the outcome in question:

- Exit polling
- Literary Digest and FDR
- Online polls

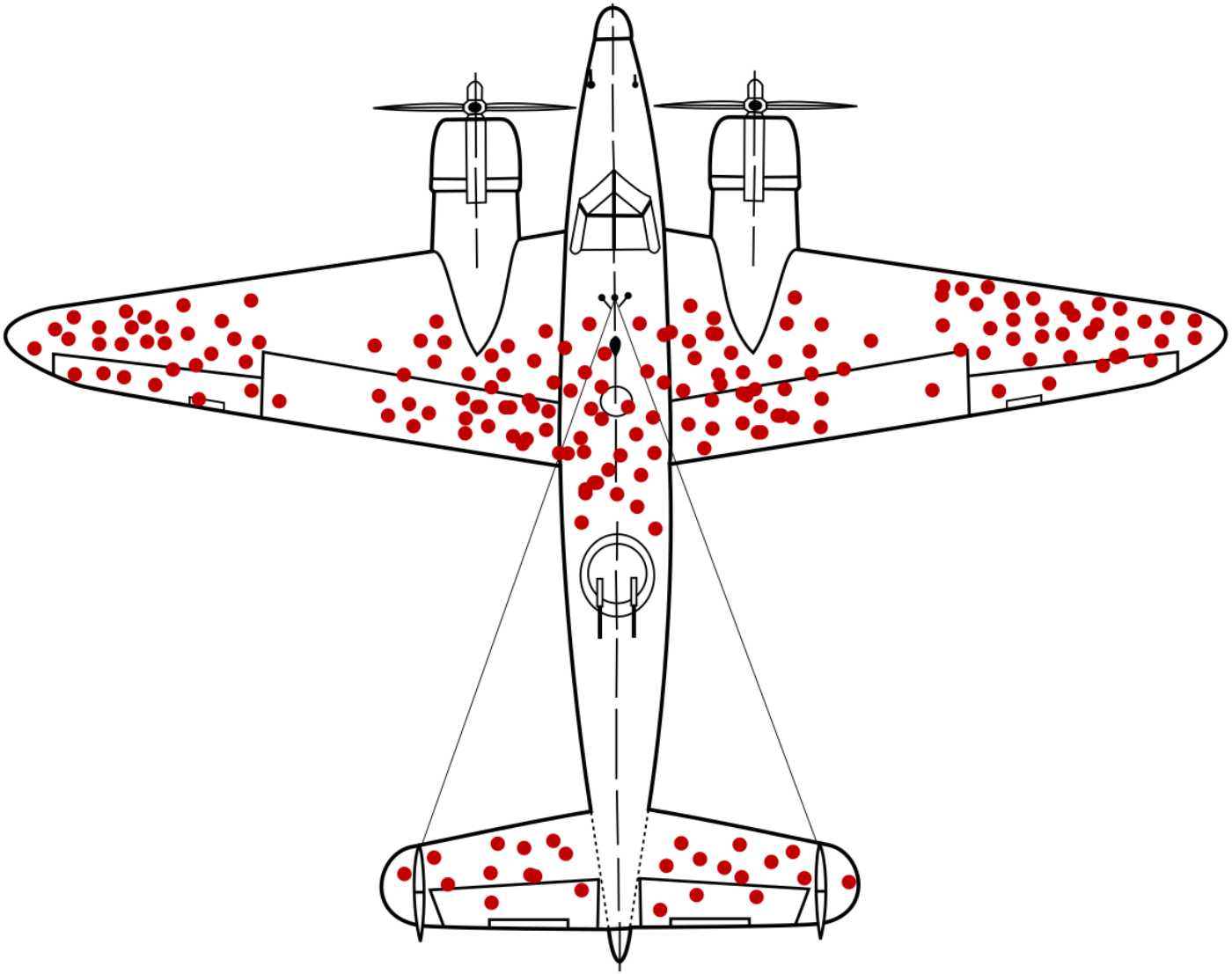
Non-response Bias – describes situation in which willingness to response may be associated with outcome

- Online store reviews
- Customer service
- Health outcomes

Example

In 2017, the Speak Out Iowa survey for sexual misconduct and dating violence was sent out to all degree seeking undergraduate, graduate, and professional students (N = 30,458). A total of 6,952 responses were collected with 67% of respondents identifying as female and 38% identifying as male. Is this sample representative of the population in question? Why or why not?

Example



Bias

"Are there any factors associated with the collection of our data that may have *any* relationship to the outcome we are intending to study?"

If yes, the outcome in our sample may systematically deviate from that in our population

Review

- **Inference** is the process of using an estimate *from a sample* to describe a characteristic of a *population*
- Estimates from sample can deviate from truth in two ways:
  - **Sampling bias**
  - **Sampling variability**
- Sampling bias is a result of *how we collect our sample*
- Sampling variability is multifaceted, primarily involving *sample size* and *variability within the population*