Sampling

Grinnell College

February 26, 2024

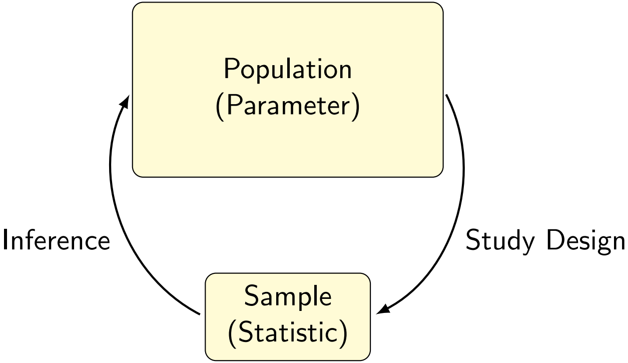
What have we done

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*

The Statistical Framework



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 best guess, of the statistic in question

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

Notation

Typically, statisticians use special notation to differentiate *population parameters* (things we wish to know) from *statistics* computed from our sample:

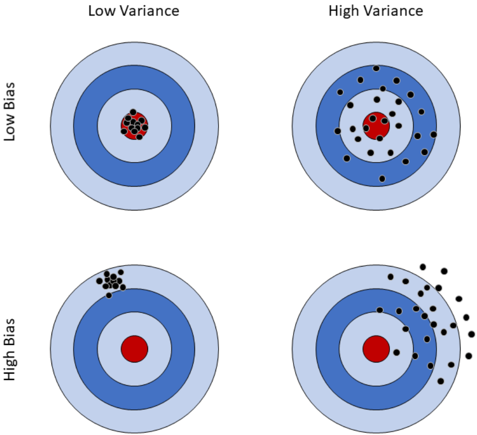
	Population Parameter	Sample Statistic
Mean	μ	$\overline{x}$
Standard Deviation	σ	s
Proportion	p	$\hat{p}$
Correlation	ρ	r
Regression	β	$\hat{\beta}$

Sources of Error

Thre are two main reasons why our sample statistic may differ from the population parameter:

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

Bias/Variability



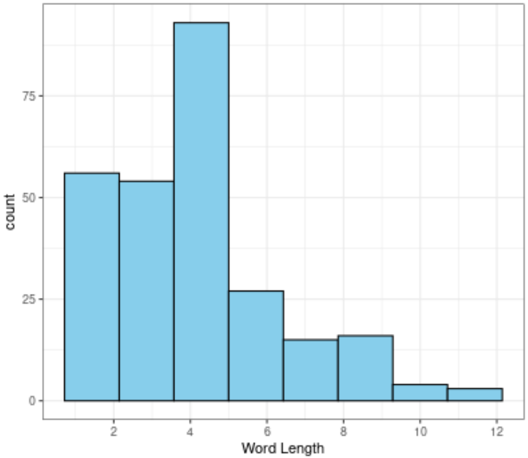
Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal.

Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battle-field of that war. We have come to dedicate a portion of that field, as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this.

But, in a larger sense, we can not dedicate – we can not consecrate – we can not hallow – this ground. The brave men, living and dead, who struggled here, have consecrated it, far above our poor power to add or detract. The world will little note, nor long remember what we say here, but it can never forget what they did here. It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us – that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion – that we here highly resolve that these dead shall not have died in vain – that this nation, under God, shall have a new birth of freedom – and that government of the people, by the people, for the people, shall not perish from the earth.

Abraham Lincoln, 1864

Word Length



Word Length	1	2	3	4	5	6	7	8	9	10	11
Frequency	7	49	54	59	34	27	15	6	10	4	3

Types of Bias

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

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

- Exit polling
- Literary Digest and FDR
- Online polls

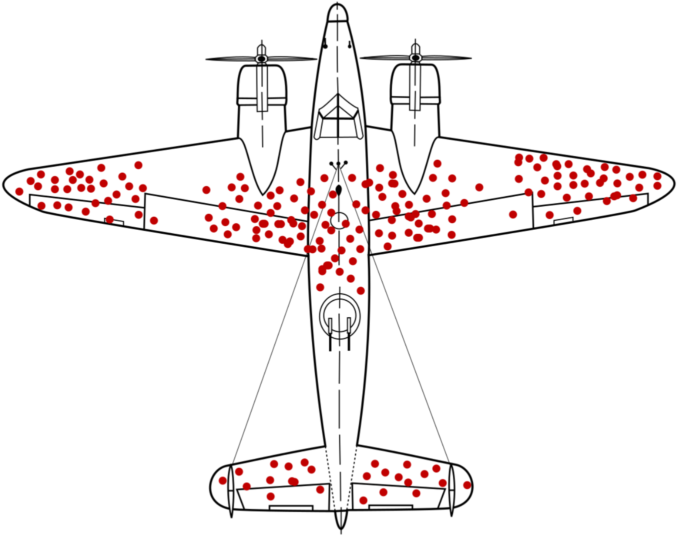
Non-reponse Bias – describes situation in which willingness to response may be associated with outcome

- 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



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 executre

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*

Distribution

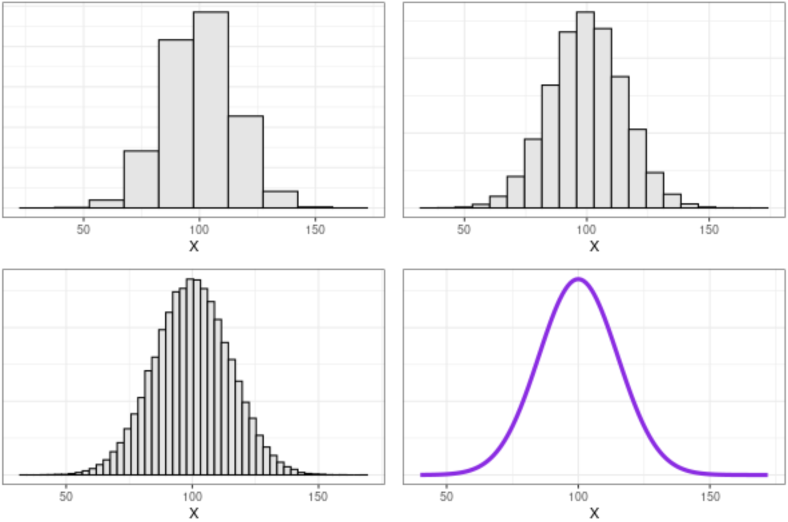
Recall that a **distribution** describes:

- What values our variable can take
- How frequently they occur

Often, the parameters of a population we are interested in learning are associated with their distribution

The most common distribution we will be working with in this couse will be the *normal distribution*

Normal Distribution



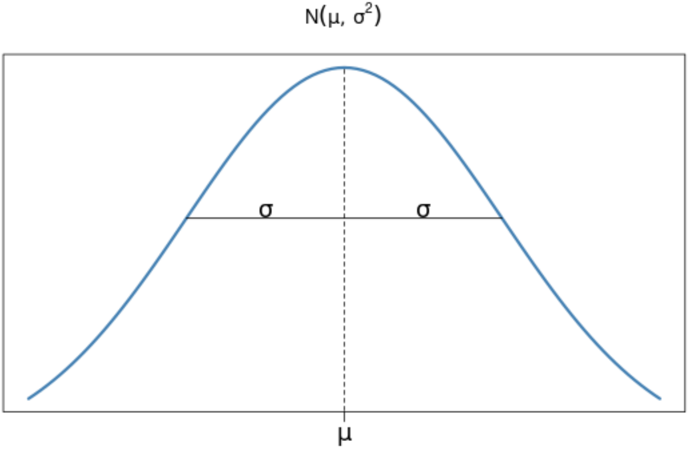
Normal Distribution

A variable X is said to be *normally distributed* with *parameters* mean μ and variance σ^2, which we express as

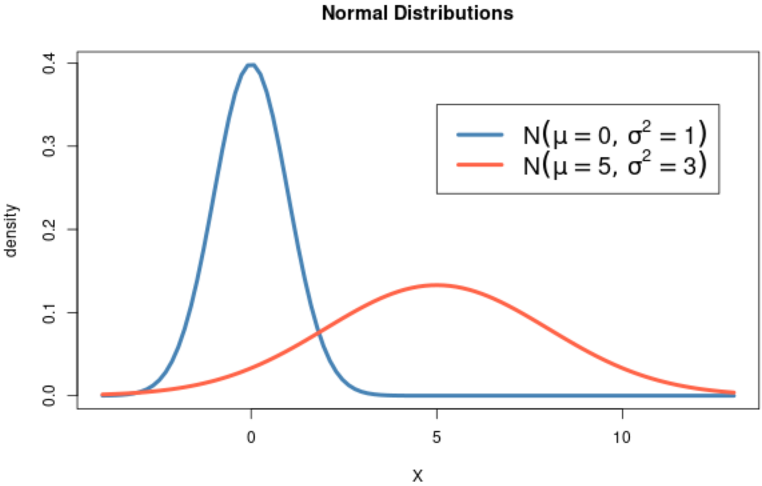
$$X \sim N(\mu, \sigma^2)$$

Here, μ and σ^2 are the parameters of the normal distribution

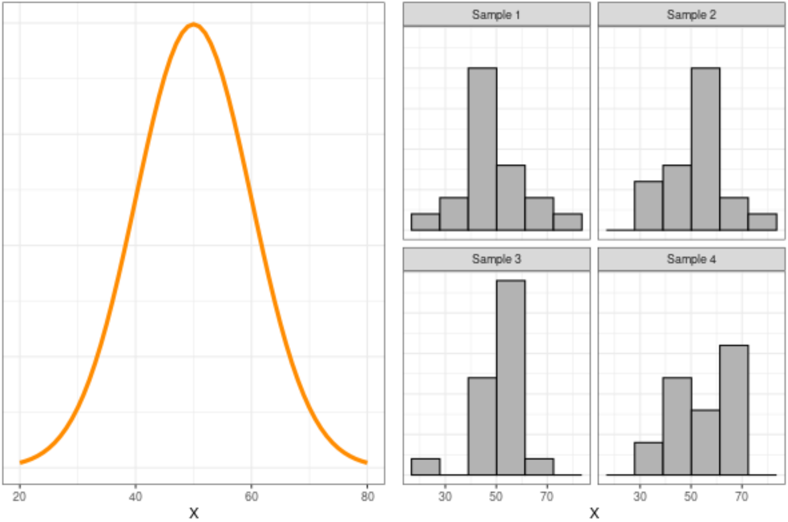
Normal Distribution



Normal Distribution



Representative Samples



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*
- Normal Distribution
    - "Theoretical"
    - Parameters location and scale