LESSON 16

Sampling

Lesson 17 Learning Objectives:

- **Explain the differences between a target population, a sampling frame and a sample**
- **List criteria for a sample to be considered a probability sample**
- **Provide examples of common probability sampling schemes**
- **Explain how to gather an IID sample**

Roadmap

Lesson 17:

Sampling

- Definitions
- Sampling Bias: A Case Study
- Random (aka Probability) Samples
- IID Samples

Learning Objectives:

- **Explain the differences between a target population, a sampling frame and a sample**

Sampling: Introduction

- Sampling: Definitions

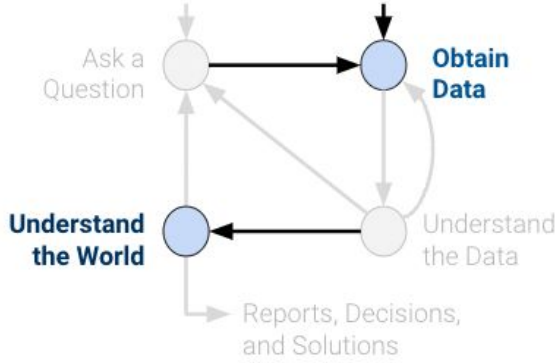
From Populations to Samples

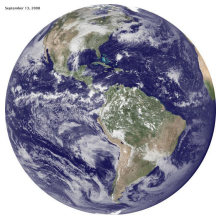
We've talked extensively about **populations**:

- If we know the **distribution of a random variable**, we can reliably compute expectation, variance, functions of the random variable, etc.

However, in Data Science, we often collect **samples**.

- We don't know the distribution of our population.
- We'd like to use the distribution of your sample to estimate/infer properties of the population.





Population

Inference

Sample size n

Course Scope for the Next Few Weeks: Statistical Inference

Questions of Interest:

In situations where we can't observe the entire population, what can we safely infer by polling a sample drawn from that population?

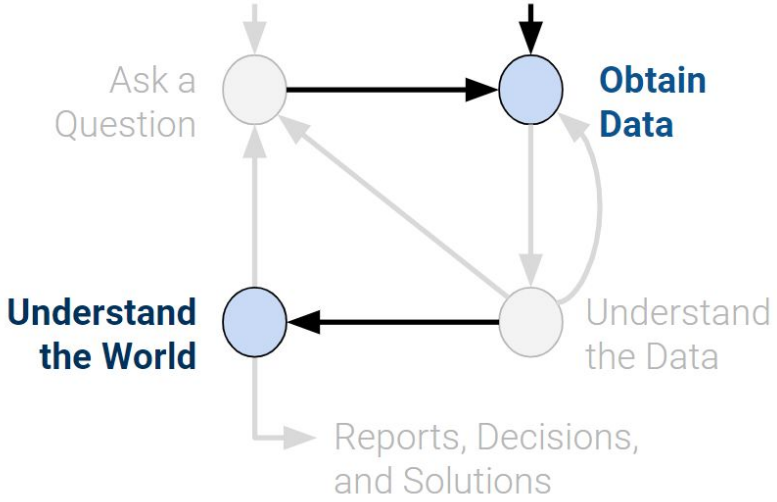
How large does your sample need to be before your conclusions become trustworthy, and how do we express confidence in what we conclude?

Sampling

Understanding the sampling process is what lets us go from **describing the data** to **understanding the world**

Without knowing / assuming something about how the data were collected:

- There is no connection between the **sample** and the **population**
- The **data set** doesn't tell us about the **world behind the data**



Population

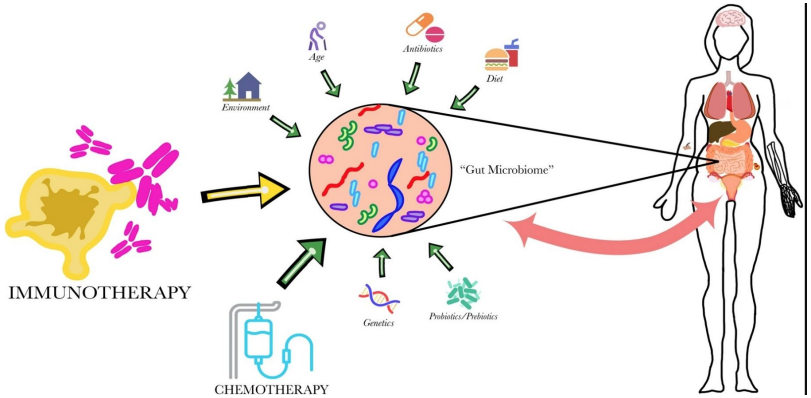
This is a **population**.

Other kinds of populations

The elements in a population are not always people!

Could be

- **Bacteria** in your gut (sampled using DNA sequencing)
- **Trees** of a certain species
- **Small businesses** receiving a microloan
- **Published results** in a journal / field



In any of these cases we might examine a sample and try to draw an inference about the population it came from.

- Simplest example: what % have some binary property (like voting intention)?

Censuses and Surveys

A **census** is “an official count or survey of a **population**, typically recording various details of individuals.”

A **survey** is a set of questions.

What is asked, and how it is asked, can affect:

- How the respondent answers.
- **Whether** the respondent answers.

There are entire courses on surveying!

Sampling from a finite population

A census is great, but expensive and difficult to execute.

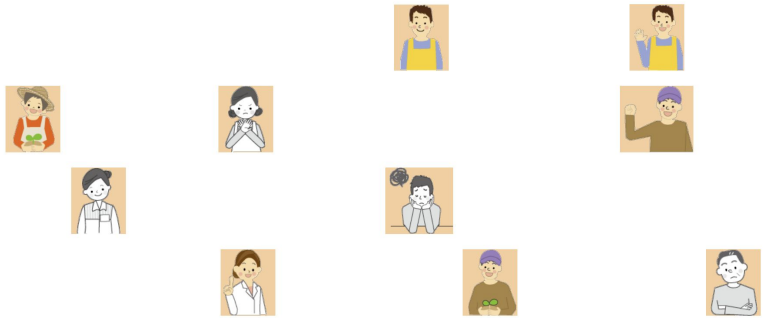
- Would **all** voters be willing to participate in a voting census prior to an actual election?

A **sample** is (usually) a subset of the population.

Population

This is a population.

Sample



A sample is selected from a population.

Target Population, sample, and sampling frame

Target Population: The group that you want to learn something about.

Sampling Frame: The list from which the sample is drawn.

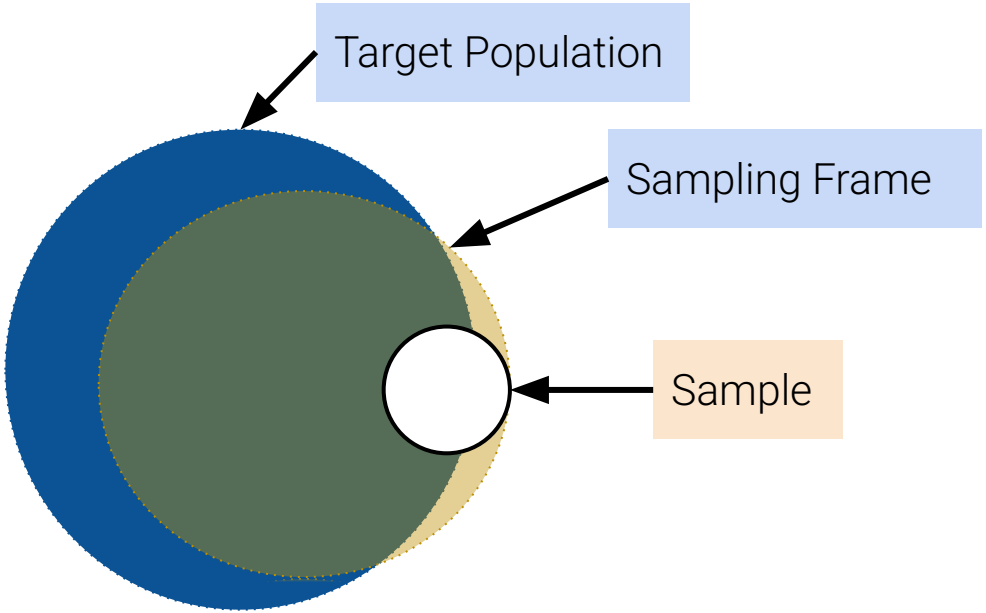
- If you're sampling people, the sampling frame is the set of all people that could possibly end up in your sample.

Sample: Who you actually end up sampling.

- A subset of your sampling frame.

There may be individuals in your **sampling frame** (and hence, your sample) that are **not** in your population!

Similarly, there might be individuals in your target population that are not in your sampling frame.

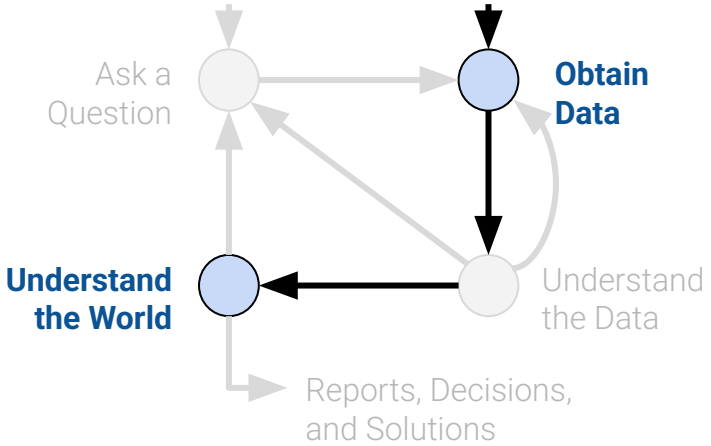


Sampling from a finite population

Samples are often used to make **inferences about the population**.

- How you draw the sample will affect the accuracy of the inference
- Two sources of error in a sample:
  - **chance error**: random samples can vary from what is expected in population, in any direction.
    - Can use sampling techniques so that we can quantify this error
  - **bias**: a systematic error in one direction.
    - Could come from our sampling scheme, and/or survey methods.
    - Often we can't assess the potential magnitude of the bias. Protocols are key to reducing these sources of bias.

Inference: drawing conclusions (and quantifying their reliability) about a population based on a sample.



Chance Error

The following types of variation results from a chance mechanism and have the advantage of being quantifiable:

Sampling variation

Results from using chance to select a sample. In this case, we can, in principle, compute the chance that a particular collection of elements is selected for the sample.

Assignment variation

Occurs in a controlled experiment when we assign units at random to treatment groups. In this situation, if we split the units up differently, then we can get different results from the experiment. This assignment process allows us to compute the chance of a particular group assignment.

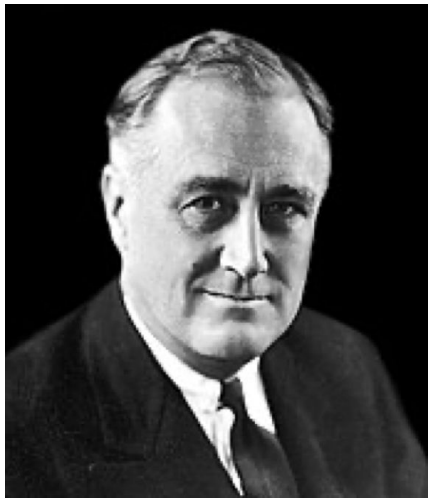
Measurement error

Results from the measurement process. If the instrument used for measurement has no drift or bias and a reliable distribution of errors, then when we take multiple measurements on the same object, we get random variations in measurements that are centered on the truth.

Sampling Bias:

- **Sampling Bias: A Case Study**

Case study: 1936 Presidential Election



Roosevelt (D)

Landon (R)

In 1936, President Franklin D. Roosevelt (left) went up for re-election against Alf Landon (right). As is usual, **polls** were conducted in the months leading up to the election to try and predict the outcome.

(Election result spoiler: Landon was not a U.S. President)

The Literary Digest: Election Prediction

The *Literary Digest* was a magazine. They had successfully predicted the outcome of 5 general elections coming into 1936.

They sent out their survey to **10,000,000** individuals, who they found from:

- Phone books.
- Lists of magazine subscribers.
- Lists of country club members.

	% Roosevelt	**# surveyed**
Actual election	**61%**	**All voters** (~45,000,000)
The Literary Digest poll	43%	10,000,000

How could this have happened?
They surveyed 10 million people!

The Literary Digest: What happened?

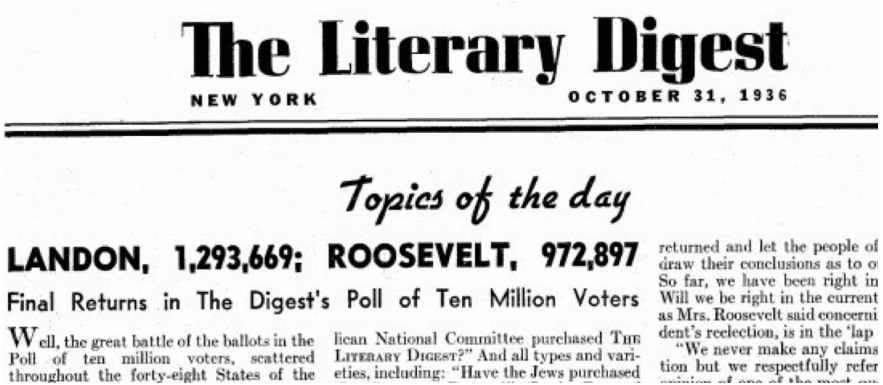
(1) The Literary Digest sample was **not representative** of the population.

- The Digest's **sampling frame**: people in the phonebook, subscribed to magazines, and went to country clubs.
- These people were more affluent and tended to vote Republican (Landon).

	% Roosevelt	# surveyed
Actual election	**61%**	**All voters** (~45,000,000)
The Literary Digest poll	43%	10,000,000

(2) Only 2.4 million people **actually filled out the survey!**

- 24% response rate (low).
- Who knows how the 76% **non-respondents** would have polled?



The Literary Digest

NEW YORK — OCTOBER 31, 1936

Topics of the day

LANDON, 1,293,669; ROOSEVELT, 972,897

Final Returns in The Digest's Poll of Ten Million Voters

Well, the great battle of the ballots in the Poll of ten million voters, scattered throughout the forty-eight States of the

lican National Committee purchased THE LITERARY DIGEST?" And all types and varieties, including: "Have the Jews purchased

returned and let the people of draw their conclusions as to o So far, we have been right in Will we be right in the current as Mrs. Roosevelt said concerni dent's reelection, is in the 'lap

"We never make any claims tion but we respectfully refer

Gallup's Poll: Election Prediction

George Gallup, a rising statistician, also made predictions about the 1936 elections.

His estimate was **much** closer despite having a smaller **sample size** of "only" 50,000

(Also more than necessary!)

	% Roosevelt	# surveyed
Actual election	**61%**	**All voters** (~45,000,000)
The Literary Digest poll	43%	10,000,000
George Gallup's poll	56%	50,000
George Gallup's prediction of Digest's prediction	44%	3,000

George Gallup also predicted what The Literary Digest was going to predict, within 1%, with a **sample size of only 3000 people**.

- He predicted the Literary Digest's **sampling frame** (phonebook, magazine subscribers, country clubs).
- So he sampled those same individuals!

Samples, while convenient, are subject to chance error and **bias**.

Adapted from *Statistics* by Freedman, Pisani, and Purves, W.W. Norton NY

Common Biases

Selection Bias

- Systematically excluding (or favoring) particular groups.
- **Example**: The Literary Digest poll excludes people not in phone books.
- **How to avoid**: Use chance mechanisms to select a sample from the sampling frame or to assign units to experimental conditions can eliminate selection bias. If sampling frame and population aren't equal, make sure sampling frame is representative of population.

Response (or Measurement) Bias

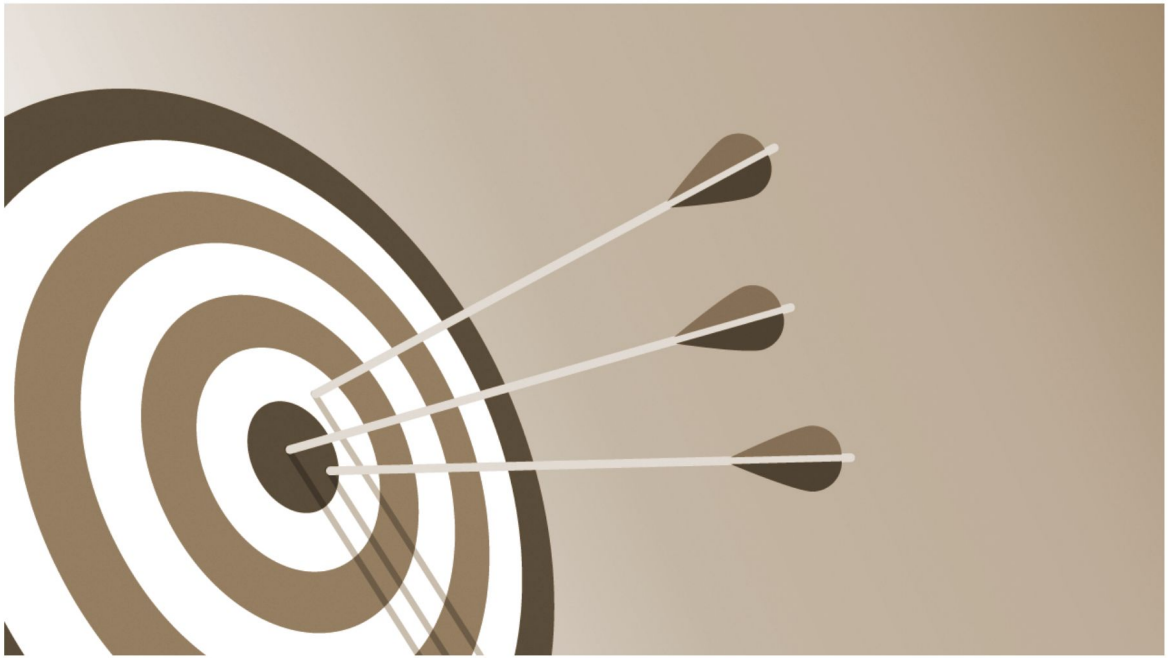
- People don't always respond truthfully, or questions lead to certain responses.
- **Example**: Asking citizenship questions on the census survey→illegal immigrants might not answer truthfully
- **How to avoid**: Response bias exists in ANY survey. However, we can try to minimize it by examining the nature of questions and the method of surveying.

Non-response Bias

- People don't always respond → People who don't respond aren't like the people who do!
- **Example**: Only 2.4m out of 10m people responded to The Literary Digest poll.
- **How to avoid**: Keep your surveys short, and be persistent.

Polling Strategies to Eliminate Bias:

5. Is accurate polling becoming harder to do?



Getty Images

Polls received a lot of criticism after the 2020 U.S. general election. Polls showed Joe Biden leading Donald Trump throughout the fall campaign season. He did win the election, but not by as large a margin or in as many states as the polls led people to expect.

https://www.pewresearch.org/course/public-opinion-polling-basics/#is-accurate-polling-becoming-harder-to-do

Learning Objectives:

- List criteria for a sample to be considered a probability sample
- Provide examples of common probability sampling schemes

Probability Samples

- Probability Samples

Quality, not quantity!

A **huge sample size** does not fix a **bad sampling method**!

We want the sample to be **representative** of the population.

Think about **tasting soup**: if it's **well-stirred**, a spoonful is all you need!

- Don't just try to get a BIG sample. If your method of sampling is BAD, and your sample is BIG, what you'll have is a BIG BAD sample
- This is a phenomenon you will explore in-depth in an assignment, where you will perform an analysis of the 2016 US Presidential Elections.

Easiest way to to get a representative sample is by using **randomness**.

Random (aka Probability) Samples

Definition of Random (aka Probability) sample:

- Before the sample is drawn, you have to know the selection probability of every group of people in the population
- Not all individuals / groups have to have equal chance of being selected

Why Use Random (aka Probability) sample?

- Since we know the source probabilities, we can measure the errors.
- Gives us a more representative sample of the population, which reduces bias.

  (Note: this is only the case when the probability distribution we're sampling from is accurate. Random samples using "bad" or inaccurate distributions can produce biased estimates of population quantities.)

- Probability samples allow us to estimate the bias and chance error, which helps us quantify uncertainty (more in a future lecture).

Probability Sample (aka Random Sample)

Why sample at random?

1. (As mentioned before) To get more representative samples → **reduce bias**
   - Random samples **can** produce biased estimates of population quantities.
2. More importantly, with random samples we can **estimate** the **bias** and **chance error** → **quantify uncertainty**

For a **probability sample**,

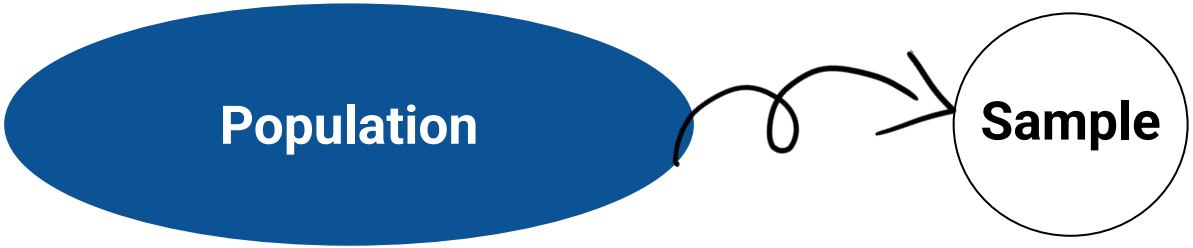
- We have to be able to provide the **chance** that any specified **set** of individuals will be in the sample.
- All individuals in the population **need not** have the same chance of being selected.
- Because we know all the probabilities, we will be able to **measure the errors**.

The real world is usually more complicated!

- Election polling: When Gallup calls, most people don't answer.
- Bacteria: We don't know the probability a given bacterium will get into a microbiome sample.

If the sampling / measurement process isn't fully under our control, we try to **model it**.

So far

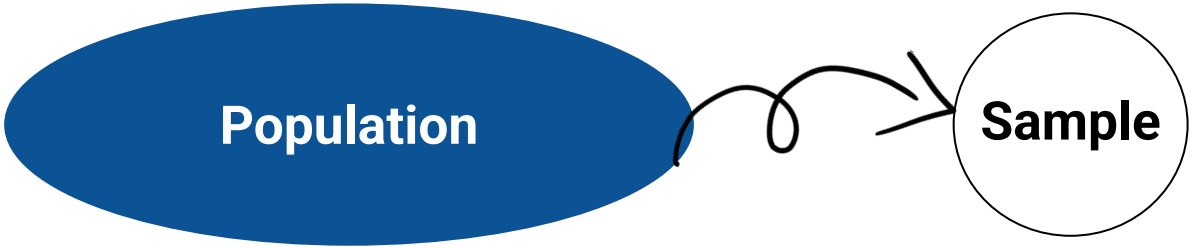


If we have a probability sample (aka a random sample):

- We can quantify error and bias.
- **Given the population distribution**, we can compute the probability of us getting a **particular sample**.

Note: We almost **never** know the population distribution! But this is a good start.

So far



If we have a probability sample:

- We can quantify error and bias (to be covered later).
- **Given the population distribution**, we can compute the probability of us getting a **particular sample**.

Note: We almost **never** know the population distribution! But this is a good start.

Special case: Random sampling with replacement of a **categorical population** produces **Multinomial Probabilities (See HW 7 & Lesson 16)**.

Multinomial Random Variable

Consider an experiment of n independent trials:

- Each trial results in one of m outcomes. $P(\text{outcome } i) = p_i$, $\sum_{i=1}^{m} p_i = 1$
- Let X_i = # trials with outcome i

Joint PMF

$$P(X_1 = c_1, X_2 = c_2, \ldots, X_m = c_m) = \binom{n}{c_1, c_2, \ldots, c_m} p_1^{c_1} p_2^{c_2} \cdots p_m^{c_m}$$

where $\sum_{i=1}^{m} c_i = n$ and $\sum_{i=1}^{m} p_i = 1$

Multinomial # of ways of ordering the outcomes

Probability of each ordering is equal + mutually exclusive

Common random sampling schemes

A **random sample with replacement** is a sample drawn **uniformly** at random **with** replacement.

- Random doesn't always mean "uniformly at random," but in this specific context, it does.
- Some individuals in the population might get picked more than once

A **simple random sample (SRS)** is a sample drawn **uniformly** at random **without** replacement.

- **Every individual has the same chance of being selected.**
- Every pair has the same chance as every other pair.
- Every triple has the same chance as every other triple.
- And so on.

A **stratified random sample:** if you have a heterogeneous population that can be broken up into homogeneous groups, randomly sample from each group proportionate to their prevalence in the population.

Sample of Convenience (NOT random)

A **convenience sample:**

- A **convenience sample** is whoever you can get ahold of.
  - Example: *sample consists of whoever visits your website*

    sample consists of whoever walks by your polling table

- Just because you think you're **sampling "randomly"**, doesn't mean you have a random sample.
- If you can't figure out **ahead of time**
  - what's the population
  - what's the **chance of selection**, for each group in the population

then you **don't have a random sample**

Warning:

- Haphazard ≠ **random**.
- Many potential sources of bias!

IID samples

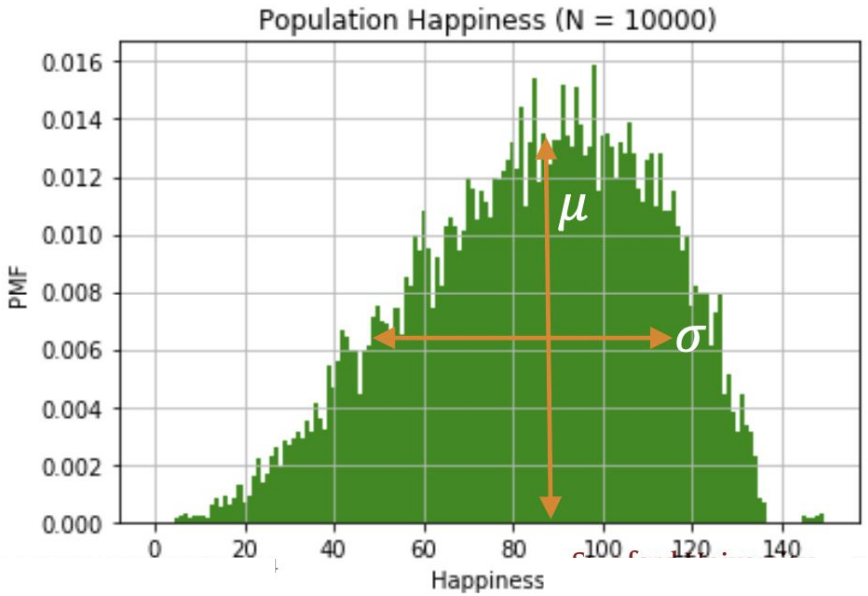
Samples and IID RV

An IID sample, mathematically

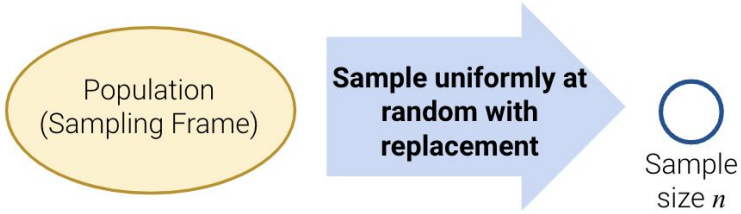
Consider n random variables $X_1, X_2, \ldots, X_n$.

The sequence $X_1, X_2, \ldots, X_n$ is **an IID sample** from distribution F if:

- X_i are all independent and identically distributed (iid)
- X_i all have same distribution function F (the underlying distribution), where $E[X_i] = \mu$, $\text{Var}(X_i) = \sigma^2$



An IID sample, mathematically

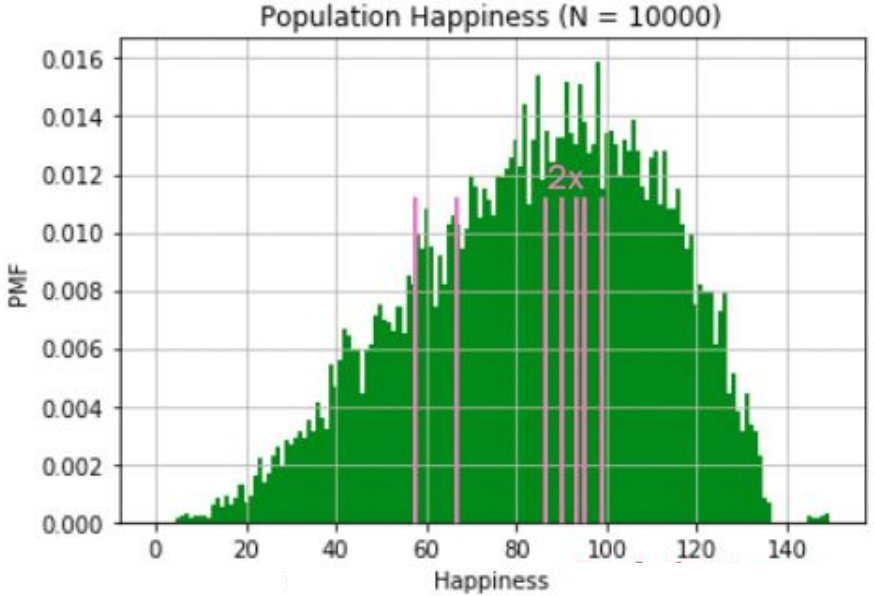


A sample of size 8:

$$(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$

The realization of a sample of size 8:

$$(59, 87, 94, 99, 87, 78, 69, 91)$$



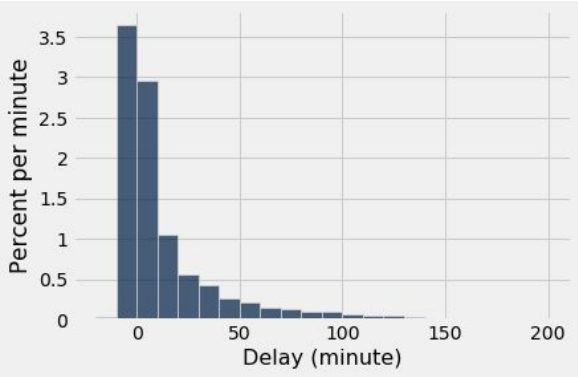
Empirical Distribution of an IID Sample

If the **sample size is large**, then

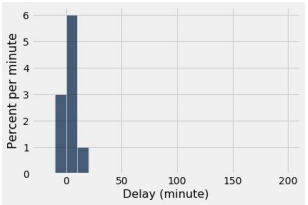
the empirical distribution of a sample (specifically a random sample with replacement)

resembles the probability distribution of the population
with high probability.

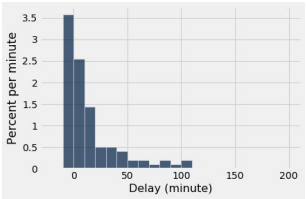
Population (Theoretical Probability) Distribution



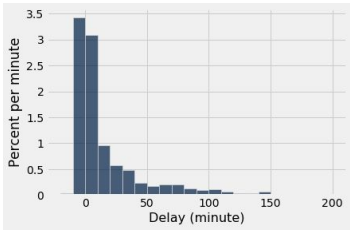
Empirical distribution of random sample of size n with replacement



n=10



n=100



n=1000

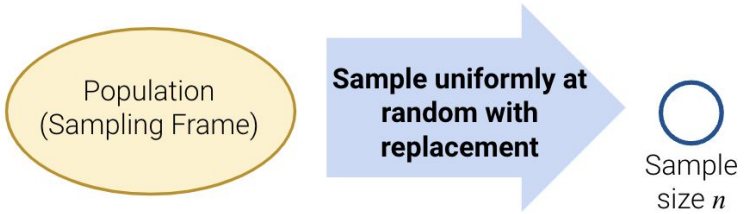
IID Random Variables

Recall:

$X_1, X_2, \ldots, X_n$ are independent and identically distributed if

- $X_1, X_2, \ldots, X_n$ are independent, and
- All have the same PMF (if discrete) or PDF (if continuous).

A Random Sample With Replacement is a Set of IID Random Variables



A common situation in data science:

- We have an enormous population.
- We can only afford to sample a relatively small number of individuals (without replacement)

What about a Simple Random Sample? Is it also IID?

As the **population gets very large** compared to the sample, then random sampling **with** replacement becomes a **good approximation** to random sampling **without**.

A very common approximation for gathering an IID sample

A common situation in data science:

- We have an enormous population.
- We can only afford to sample a relatively small number of individuals.

As the **population gets very large** compared to the sample, then random sampling **with** replacement becomes a **good approximation** to random sampling **without**.

Example: Suppose there are 10,000 people in a population.
Exactly 7,500 of them like Reese's; the other 2,500 like Snickers.

What is the probability that in a random sample of 20, **all people like Reese's**?

Random Sample Without Replacement (aka Simple Random Sample)

$$\overbrace{\left(\frac{7500}{10000}\right)}^{0.75}\overbrace{\left(\frac{7499}{9999}\right)}^{0.74997}\cdots\overbrace{\left(\frac{7482}{9982}\right)}^{0.7495}\overbrace{\left(\frac{7481}{9981}\right)}^{0.7495} \approx .003151$$

Random Sample With Replacement

$$\left(\frac{7500}{10000}\right)^{20} \approx .003171$$

Probabilities of sampling with replacement are much easier to compute!

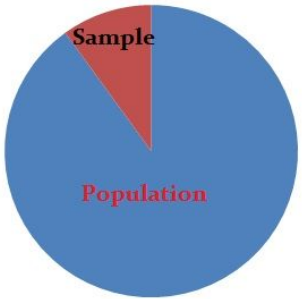
10% Rule for gathering an IID sample

As the **population gets very large** compared to the sample, then **random sampling with replacement** becomes a **good approximation** to random sampling **without replacement**.

10% rule: When using a simple random sample:

If sample size < 10% of population size:

Then we can treat the sample as if it is a set of IID RV



Simple Random Sample(Random Sample Without Replacement)

$$\overbrace{\left(\frac{7500}{10000}\right)}^{0.75}\overbrace{\left(\frac{7499}{9999}\right)}^{0.74997}\cdots\overbrace{\left(\frac{7482}{9982}\right)}^{0.7495}\overbrace{\left(\frac{7481}{9981}\right)}^{0.7495}\approx .003151$$

Random Sample With Replacement

$$\left(\frac{7500}{10000}\right)^{20}\approx .003171$$

20< 0.10*(10000)