LECTURE 16

Sampling

How to sample effectively, and how to quantify the samples we collect.

CSCI 3022

Maribeth Oscamou

Content credit: Acknowledgments

Course Logistics: 7th Week At A Glance

Mon 2/26	Tues 2/27	Wed 2/28	Thurs 2/29	Fri 3/1
Attend & Participate in Class	(Optional): Attend Notebook Discussion with our TA (5-6pm Zoom)	Attend & Participate in Class	HW 6 Due 11:59pm	Attend & Participate in Class NO QUIZ!

Today's Roadmap

CSCI 3022

- Finish Lesson 15:
  - IID
  - Multinomial Probabilities
- Theoretical vs Empirical Distributions
- Sampling: Definitions
- Sampling Bias: A Case Study
- Probability Samples
- IID Samples

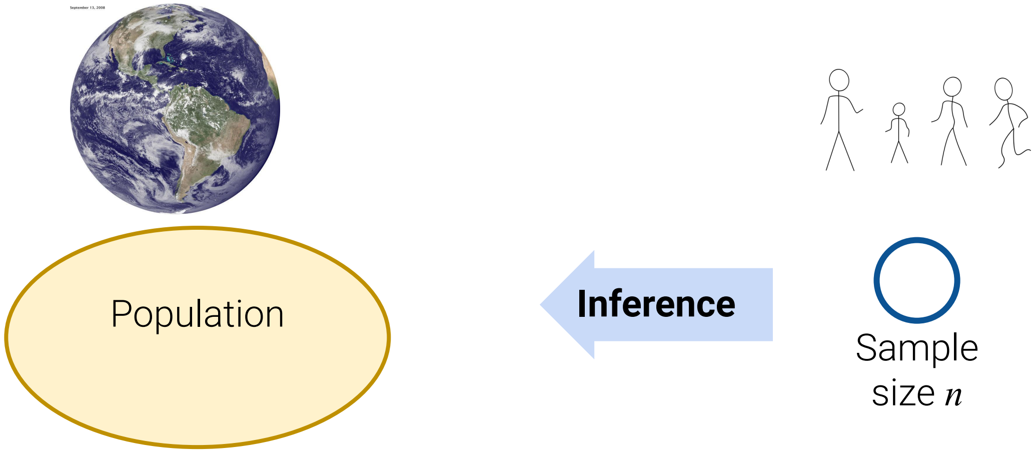
From Populations to Samples

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.



Inference

- Statistical Inference:

  Making conclusions based on data in **random samples**

- **Example**:

  fixed

  Use the data to guess the value of an *unknown number*

  depends on the **random** sample

  Create an **estimate** of the unknown quantity

Motivating Example

You want to know the true mean and variance of happiness in Bhutan.

- But you can’t ask everyone.
- You poll 200 random people.
- Your data looks like this:

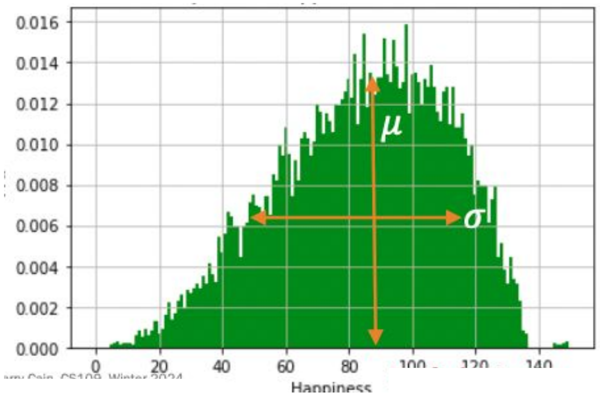
Happiness = {72, 85, 79, 91, 68, ..., 71}

- The mean of all these numbers is 83.

Is this the **true mean happiness** of Bhutanese people?

Course Scope for the Next Few Weeks: Statistical Inference

If we had a distribution F of our entire population, we could compute exact statistics about about happiness.



But we only have 200 people—or rather, a sample.

If we only have a single sample,

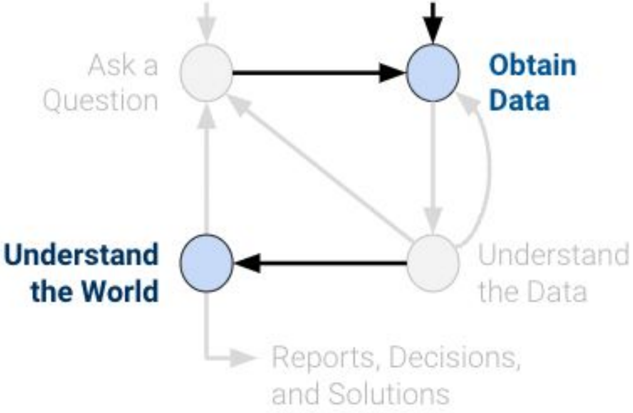
- How do we report **estimated** statistics?
  - We're careful to call them estimated mean and estimated variance, since they're based on samples (i.e., experiments)
- How do we report estimated errors on these estimates?
- How do we perform something called **hypothesis testing**? Oh, and what is it?

Course Scope for the Next Few Weeks: Statistical Inference

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?

Are there alternative ways to infer population statistics without polling entire populations?

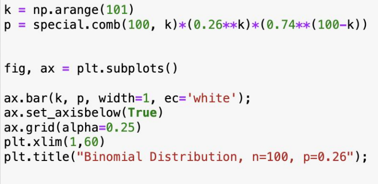


Review: Probability vs Empirical Distributions

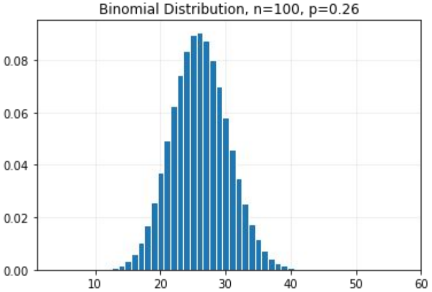
Any random variable has a distribution:

Probability (aka Population or Theoretical) Distribution

These are the distributions of random variables or the distribution of some feature of some population. We have focused on some common ones in the past few weeks)

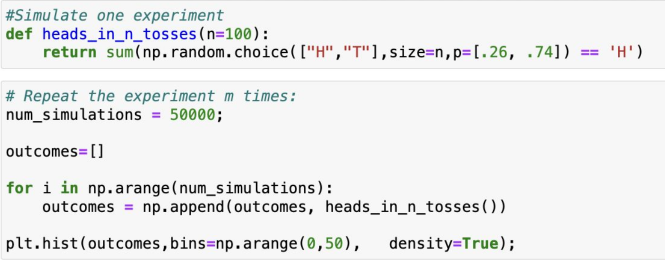


```
k = np.arange(101)
p = special.comb(100, k)*(0.26**k)*(0.74**(100-k))


fig, ax = plt.subplots()

ax.bar(k, p, width=1, ec='white');
ax.set_axisbelow(True)
ax.grid(alpha=0.25)
plt.xlim(1,60)
plt.title("Binomial Distribution, n=100, p=0.26");
```



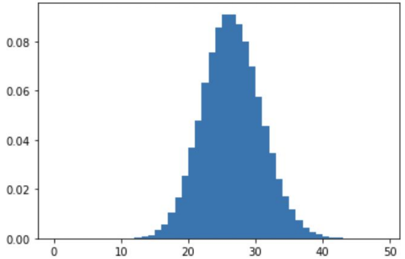
- **Empirical (aka Simulated or Sample) Distribution:**

  based on random samples (or simulations)
- Observations can be from **repetitions of an experiment or random samples from a population**
  - All observed values
  - The proportion of times each value appears



```
#Simulate one experiment
def heads_in_n_tosses(n=100):
    return sum(np.random.choice(["H","T"],size=n,p=[.26, .74]) == 'H')
```

```
# Repeat the experiment m times:
num_simulations = 50000;

outcomes=[]

for i in np.arange(num_simulations):
    outcomes = np.append(outcomes, heads_in_n_tosses())

plt.hist(outcomes,bins=np.arange(0,50),   density=True);
```



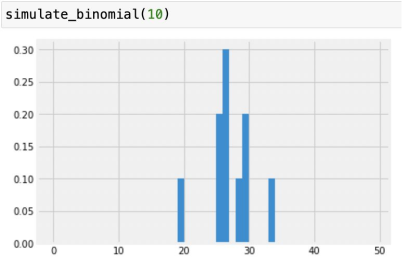
Law of Averages / Law of Large Numbers

If a chance experiment is **repeated many times**,

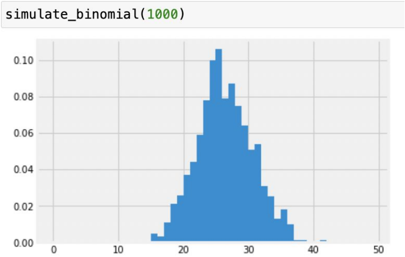
independently and under the **same conditions**,

then the **Empirical (Sample) Distribution** gets closer to the Theoretical **Probability Distribution.**

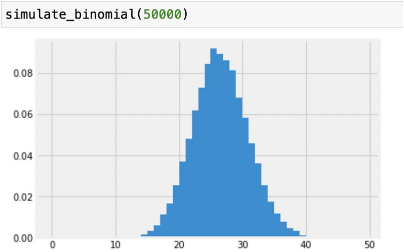
Ex: An experiment consists of flipping a coin 100 times and counting the number of heads, where the probability of heads is 0.26. You repeat this experiment and plot the distribution for the number of heads:



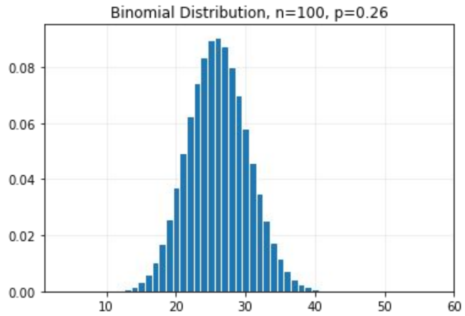
```
simulate_binomial(10)
```



```
simulate_binomial(1000)
```



```
simulate_binomial(50000)
```



As you increase the number of times you do this experiment, the empirical distribution gets closer to the theoretical probability distribution

Sampling: Definitions

CSCI 3022

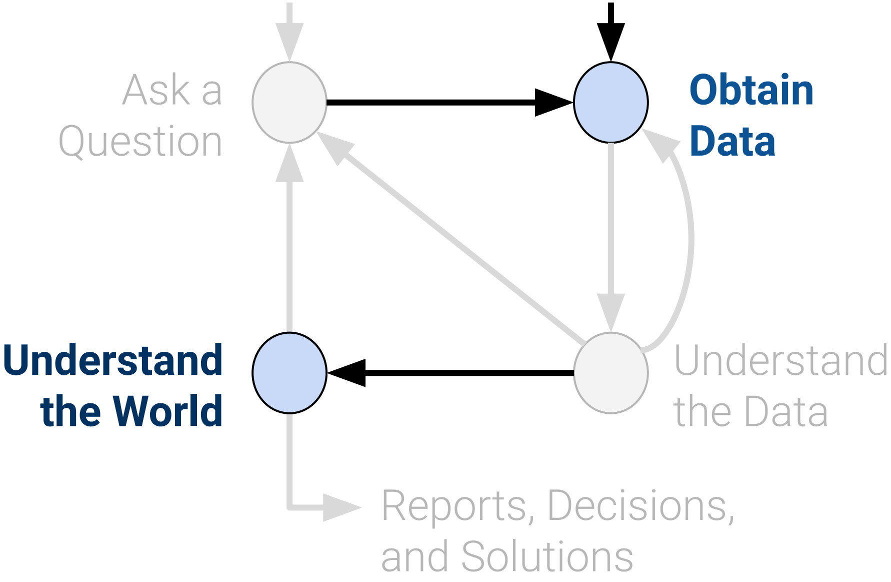
- Sampling: Definitions
- Sampling Bias: A Case Study
- Probability Samples
-

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

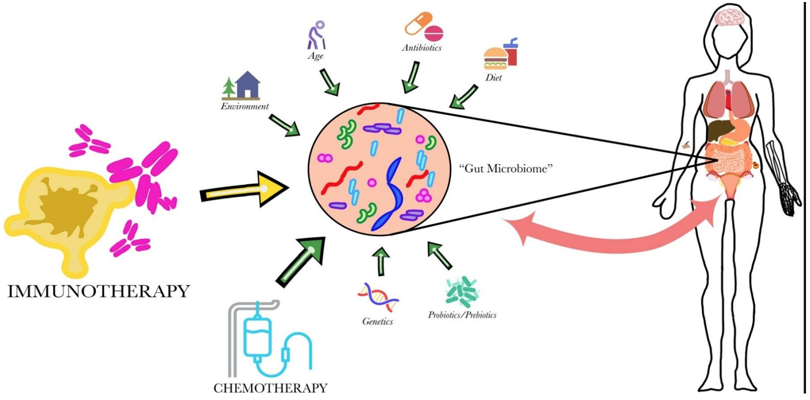
This is a **population**.

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

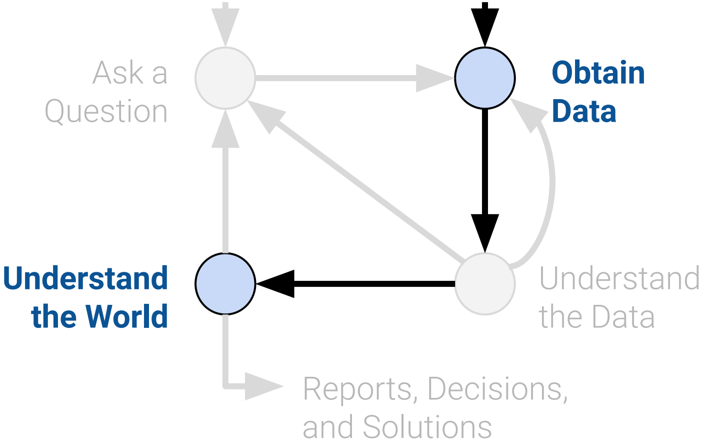
A census is great, but expensive and difficult to execute.

- 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.

- Samples are often used to make **inferences about the population**.
- How you draw the sample will affect your accuracy.
- Two sources of error in a sample:
  - **chance error**: random samples can vary from what is expected, in any direction.
  - **bias**: a systematic error in one direction.
    - Could come from our sampling scheme, and survey methods.

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



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.

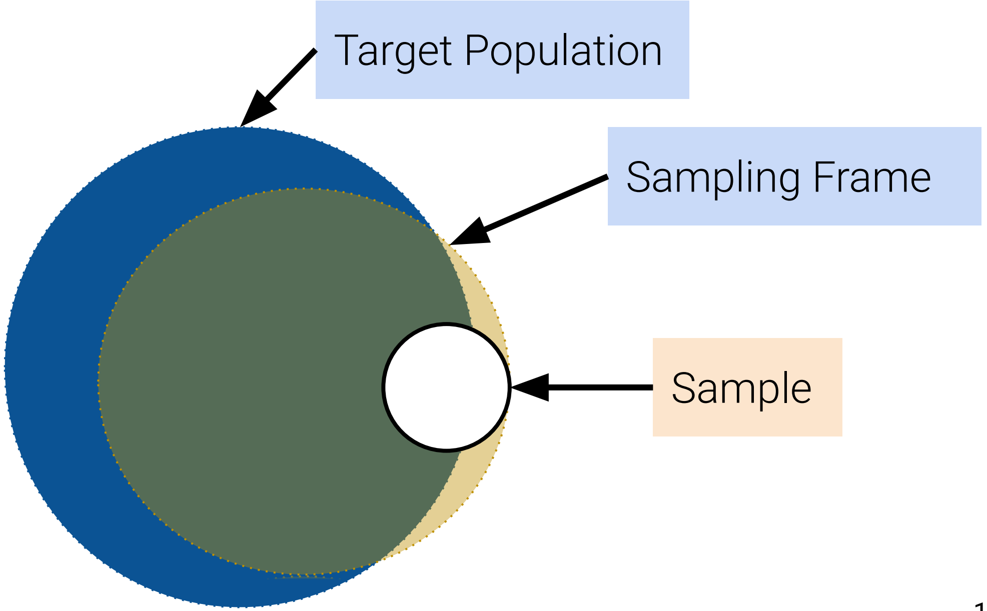
- If you're sampling people, the sampling frame is the set of all people that could possibly end up in your sample.

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.



Target Population

Sampling Frame

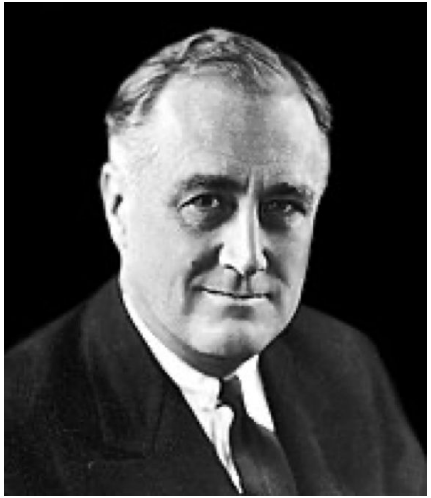
Sample

Bias: A Case Study

CSCI 3022

- Censuses and Surveys
- Sampling: Definitions
- **Sampling Bias: A Case Study**
- Probability Samples

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?

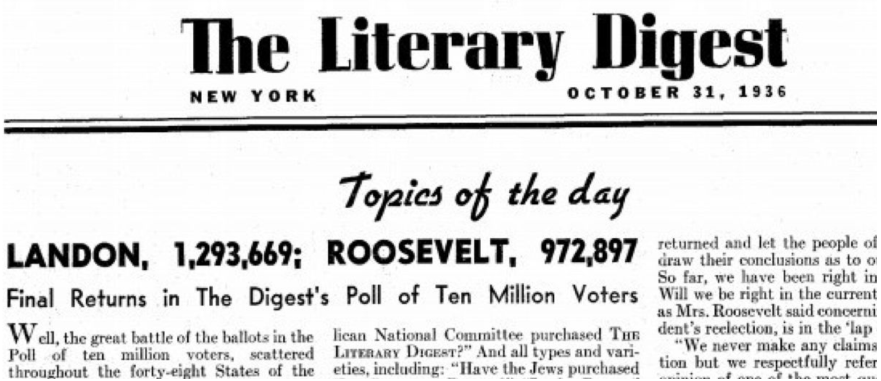
(1) The Literary Digest sample was **not representative** of the population.

- 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!)

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!

	% 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

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**: Examine the sampling frame and the method of sampling.

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.

Probability Samples

CSCI 3022

- Censuses and Surveys
- Sampling: Definitions
- Bias: A Case Study
- **Probability Samples**
- IID 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 Homework 6, 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**,

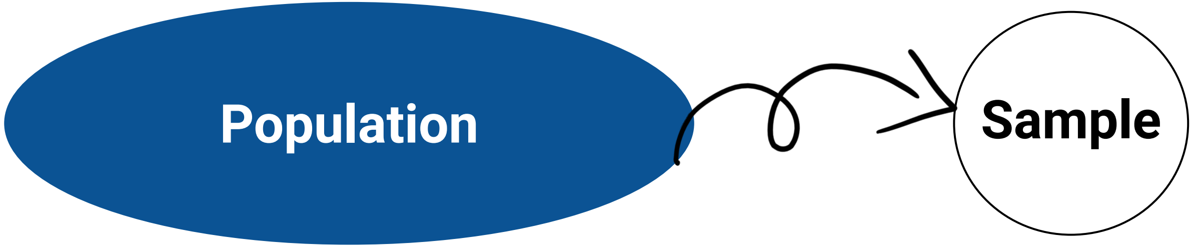
- We have to be able to provide the **chance** that any specified **set** of individuals will be in the sample.
- All individuals in the population **need not** have the same chance of being selected.
- Because we know all the probabilities, we will be able to **measure the errors**.

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

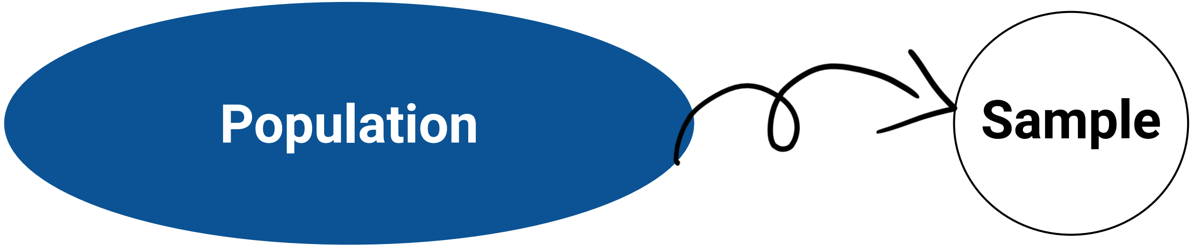


If we have a probability sample (aka a random sample):

- 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

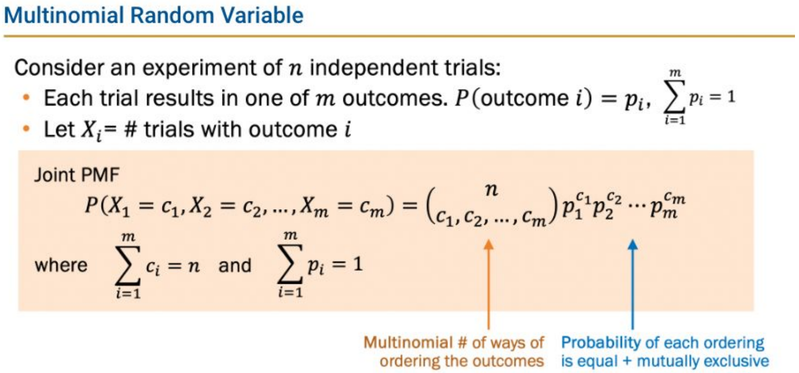


If we have a probability sample:

- 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!

Special case: Random sampling with replacement of a **categorical population** produces **Multinomial Probabilities**.



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

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

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 (aka probability 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 raffle could use either sampling scheme, depending on if winners are eligible for multiple prizes.

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 **systematic sample:** Order the sample frame. Choose an integer *k*. Sample every *kth* unit in the sample frame.

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!

An IID sample, mathematically

CSCI 3022

- Censuses and Surveys
- Sampling: Definitions
- Bias: A Case Study
- **Probability Samples**
- **An IID sample, mathematically**

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.

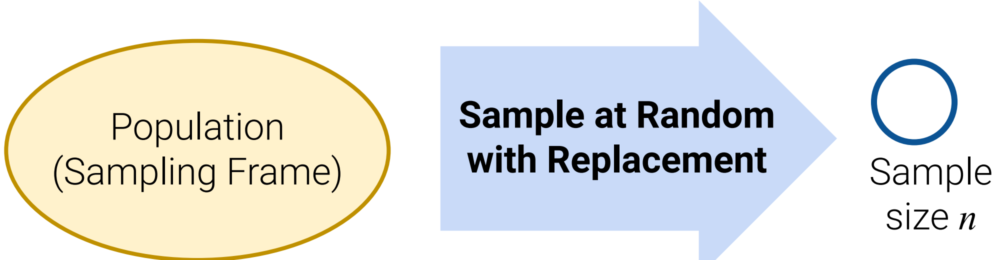
However, in Data Science, we often collect **samples**.

- 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.

The **big assumption** we make in modeling/inference:

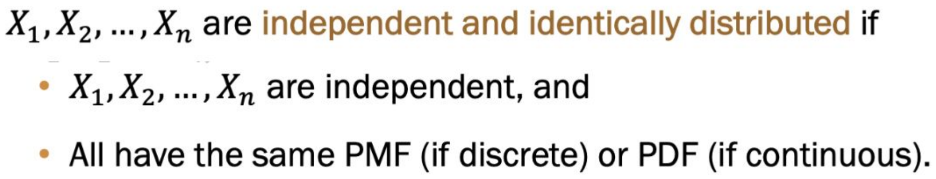
Our random sample data points are **INDEPENDENT and IDENTICALLY DISTRIBUTED (IID)**

> We can safely make this assumption anytime we sample at random with replacement (OR when we use a simple random sample and our sample size < 10% of the population size)



IID Random Variables

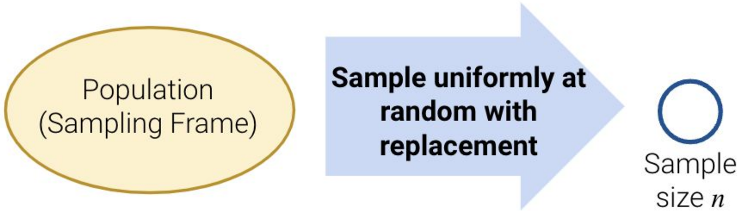
Recall:



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

- $X_1, X_2, \dots, X_n$ 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!

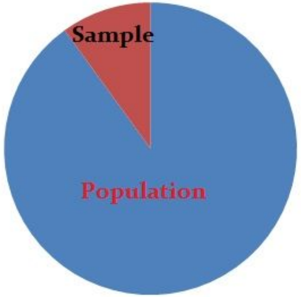
10% Rule for gathering an IID sample

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)

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

Random Sample With Replacement

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

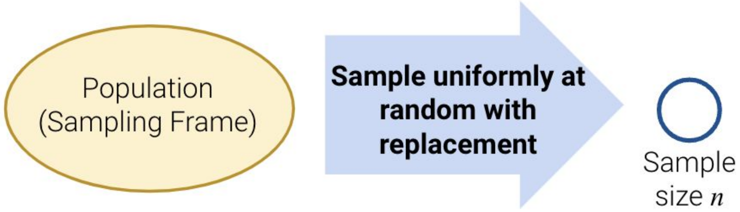
20< 0.10*(10000)

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 a **sample** from distribution F if:

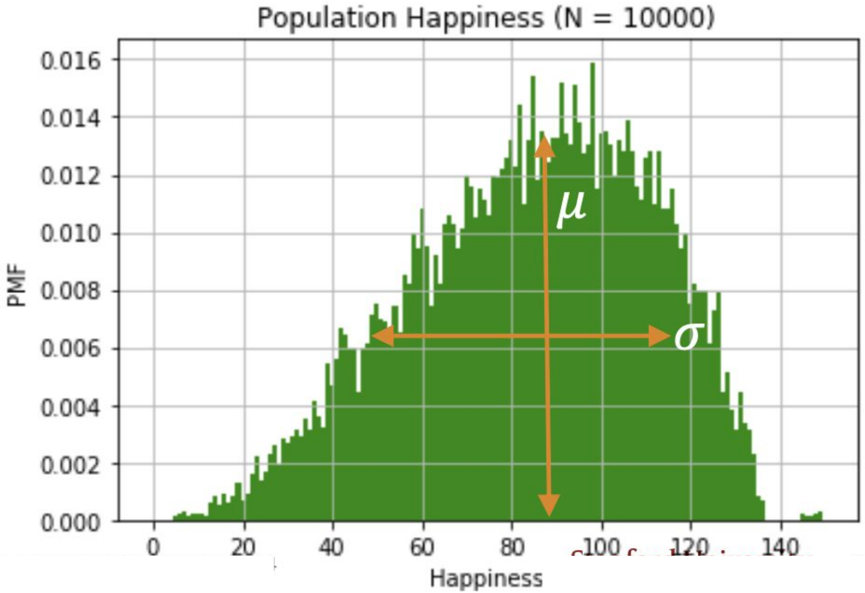
- X_i are all independent and identically distributed (iid)
- X_i all have same distribution function F (the **underlying distribution**), where $E[X_i] = \mu$, $\text{Var}(X_i) = \sigma^2$



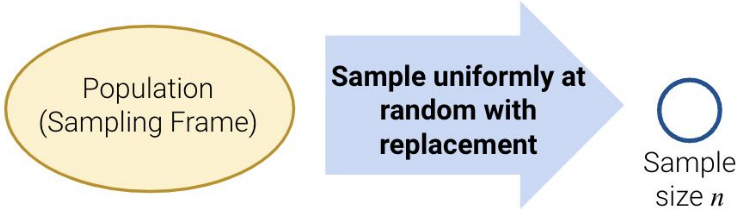
Population (Sampling Frame)

Sample uniformly at random with replacement

Sample size n



An IID sample, mathematically



Population (Sampling Frame)

Sample uniformly at random with replacement

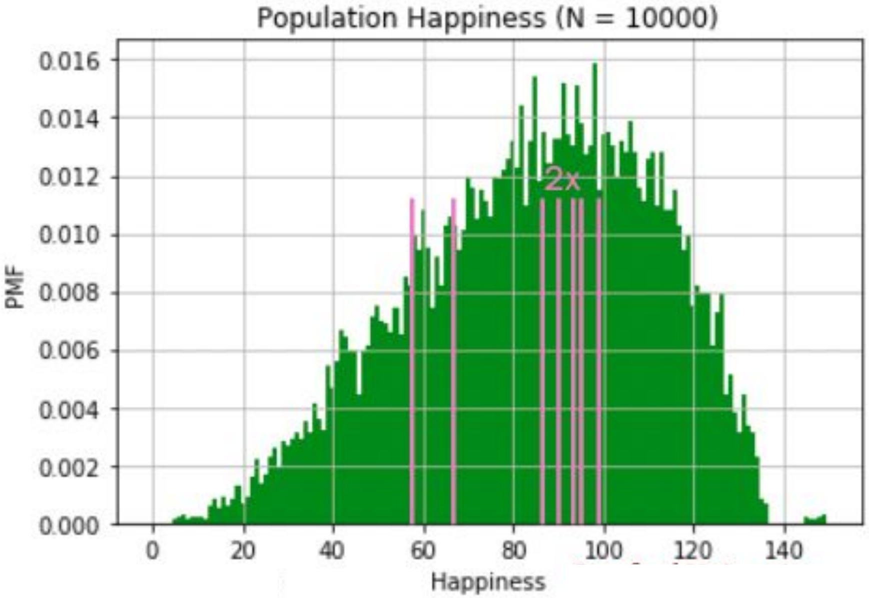
Sample size n

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:

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



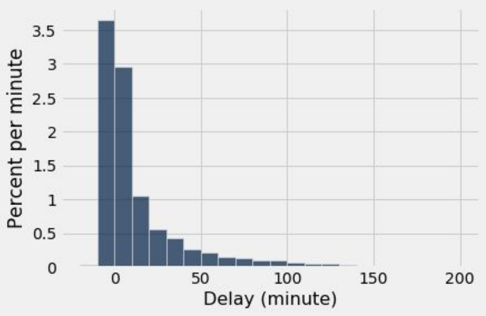
Empirical Distribution of an IID Sample

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

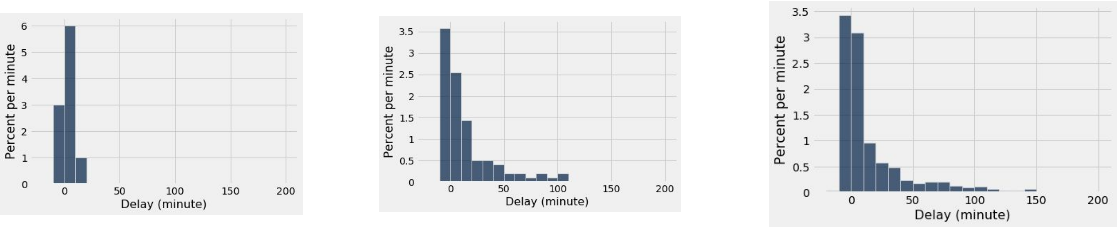
the empirical distribution of a sample (specifically a random sample with replacement)

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

Appendix

CSCI 3022

Demo: Selection Bias
Practice: Probability Samples

- S

Demo

Barbie vs. Oppenheimer

On July 21st, two highly anticipated movies arrive in theaters: **Barbie** and **Oppenheimer**.

We want to know which movie will prevail on opening day, in Berkeley.

NY Times, GQ

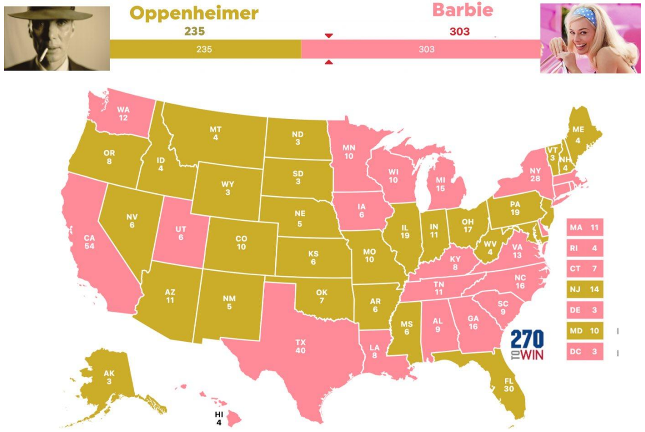
Imaginary Barbie Land Pollster

Suppose we took a sample of Berkeley residents to predict the box office outcomes.

- We poll all **retirees** for their preference.
- Even if they answer truthfully, this is a **convenience sample**.

Then, suppose July 21st has passed (it has!).

- How “off” is our sample estimate from the actual outcome?
- How would a random sample with replacement have performed?



Twitter

Demo

Answer: Barbie won!

- Opening weekend tally was:
    - **Barbie**: $155M
    - **Oppenheimer**: $82.4M

Demo

Example Scheme 1: Probability Sample

Suppose I have 3 TA's (**A**lan, **B**ennett, **C**eline):
I decide to sample 2 of them as follows:

- I choose **A** with probability 1.0
- I choose either **B** or **C**, each with probability 0.5.

All subsets of 2:	{**A**, **B**}	{**A**, **C**}	{**B**, **C**}
Probabilities:	0.5	0.5	0

This is a **probability sample** (though not a great one).

- Of the 3 people in the population, I know the chance of getting each subset.
- Suppose I'm measuring the average distance TA's live from campus.
  - This scheme does not see the entire population!
  - My estimate using the single sample I take has some **chance error** depending on if I see AB or AC.
  - This scheme **biases** towards A's response

Example Scheme 2: Simple Random Sample?

We have the following sampling scheme:

- A class roster has 1100 students listed alphabetically.
- Pick one of the first 10 students on the list at random (e.g. Student 8).
- To create your sample, take that student and every 10th student listed after that (e.g. Students 8, 18, 28, 38, etc).

1. Is this a probability sample?

2. Does each student have the same probability of being selected?

3. Is this a simple random sample?

Example Scheme 2: Simple Random Sample?

Consider the following sampling scheme:

- A class roster has 1100 students listed alphabetically.
- Pick one of the first 10 students on the list at random (e.g. Student 8).
- To create your sample, take that student and every 10th student listed after that (e.g. Students 8, 18, 28, 38, etc).

1. Is this a probability sample?

Yes.

For a sample [n, n + 10, n + 20, …, n + 1090], where 1 <= n <= 10, the probability of that sample is 1/10.

Otherwise, the probability is 0.

Only 10 possible samples!

2. Does each student have the same probability of being selected?

Yes.

Each student is chosen with probability 1/10.

3. Is this a simple random sample?

No.

The chance of selecting (8, 18) is 1/10; the chance of selecting (8, 9) is 0.

This method is called a **systematic sample**