Data Science Concepts and Analysis

Week 2: Sampling, Missingness, and Bias

- Sampling and statistical bias
- Case study: voter fraud
- The missing data problem

This week: collect and acquaint

Objective: Enable you to critically assess data quality based on how it was collected.

- **Sampling and statistical bias**
    - Sampling terminology
    - Common sampling scenarios
    - Sampling mechanisms
    - Statistical bias
- **The missing data problem**
    - Missingness modulates sampling design
    - Types of missingness: MCAR, MAR, and MNAR
    - Pitfalls and simple fixes
- **Case study: voter fraud**
    - Steven Miller's analysis of 'Voter Integrity Fund' surveys
    - Sources of bias
    - Ethical issues

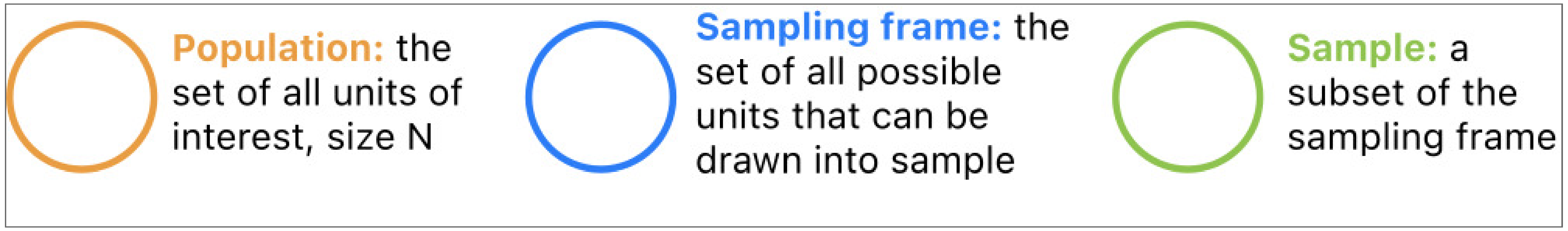
Sampling and statistical bias

- Sampling terminology
    - Population
    - Sampling frame
    - Sample
- Common sampling scenarios
    - Census
    - Simple random sample
    - 'Typical' sample
    - 'Administrative data'
- Sampling mechanisms
    - Census
    - Random sampling
    - Probability sampling
    - Nonrandom sampling
- Statistical bias
    - Definition
    - Sampling design and bias

Sampling terminology

Here we'll introduce standard statistical terminology to describe data collection.
All data are collected somehow. A **sampling design** is a ***way of selecting observational units for measurement***. It can be construed as a particular relationship between:

- a **population** (all entities of interest);
- a **sampling frame** (all entities that are possible to measure); and
- a **sample** (a specific collection of entities).



Population: the set of all units of interest, size N

Sampling frame: the set of all possible units that can be drawn into sample

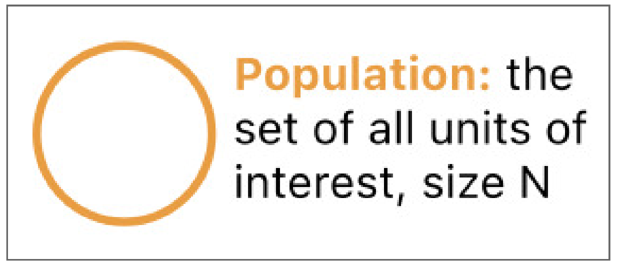
Sample: a subset of the sampling frame

Population

Last week, we introduced the terminology **observational unit** to mean ***a certain (usually physical) entity measured for a study*** -- using this terminology, datasets consist of observations made on observational units.
In less technical terms, all data are data *on* some kind of thing, such as countries, species, locations, and the like.

A statistical **population** is the ***collection of all units of interest***. For example:

- all countries (GDP data);
- all mammal species (Allison 1976);
- all babies born in the US (babynames data)
- all locations in a region (SB weather data);
- all adult U.S. residents (BRFSS data).



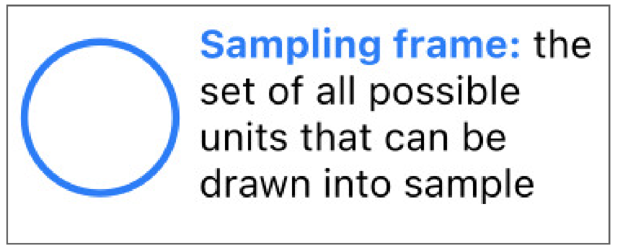
Population: the set of all units of interest, size N

Sampling frame

There are usually some units in a population that can't be measured due to practical constraints -- for instance, many adult U.S. residents don't have phones or addresses.

For this reason, it is useful to introduce the concept of a **sampling frame**, which refers to ***the collection of all units in a population that can be observed for a study***. For example:

- all countries reporting economic output between 1961 and 2019;
- all nonendagered mammals that die of natural causes in monitored areas;
- all babies with birth certificates from U.S. hospitals born between 1990 and 2018;
- all locations where it is possible to install weather monitors;
- all adult U.S. residents with phone numbers in 2019.



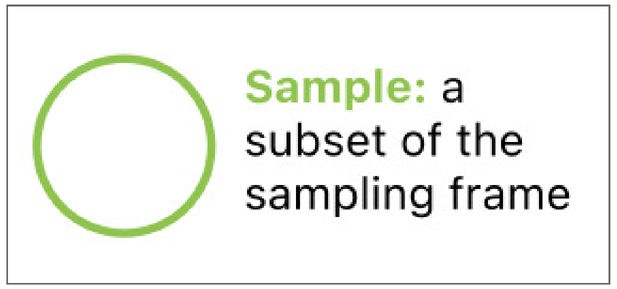
Sampling frame: the set of all possible units that can be drawn into sample

Sample

Finally, it's rarely feasible to measure every observable unit due to limited data collection resources -- for instance, states don't have the time or money to call every phone number every year.

A **sample** is ***a collection of units in the sampling frame actually selected for study***. For instance:

- 234 countries;
- 62 mammal species;
- 13,684,689 babies born in CA;
- 1 weather station location at SB airport;
- 418,268 adult U.S. residents.



Sample: a subset of the sampling frame

Sampling scenarios

We can now imagine a few common sampling scenarios by varying the relationship between population, frame, and sample.

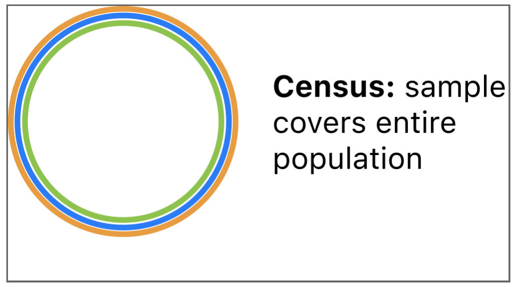
Let's introduce some notation. Denote an observational unit by U_i, and let:

$$\begin{aligned}
&\mathcal{U} = \{U_i\}_{i \in I} && \text{(universe)} \\
&P = \{U_1, \dots, U_N\} && \text{(population)} \\
&F = \{U_j : j \in J \subset I\} && \text{(frame)} \\
&S \subseteq F && \text{(sample)}
\end{aligned}$$

Sampling scenarios: population census

Perhaps the simplest scenario is a **population census**, where the entire population is observed. In this case:

$$S = F = P$$



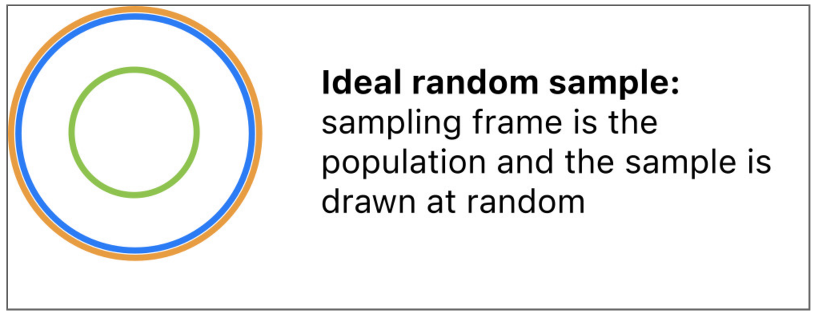
*From a census, all properties of the population are definitevely **known**.*

- No need to model census data!

Sampling scenarios: simple random sample

The statistical gold standard is the **simple random sample** (SRS) in which units are selected at random from the population. In this case:

$$S \subset F = P$$



Ideal random sample: sampling frame is the population and the sample is drawn at random

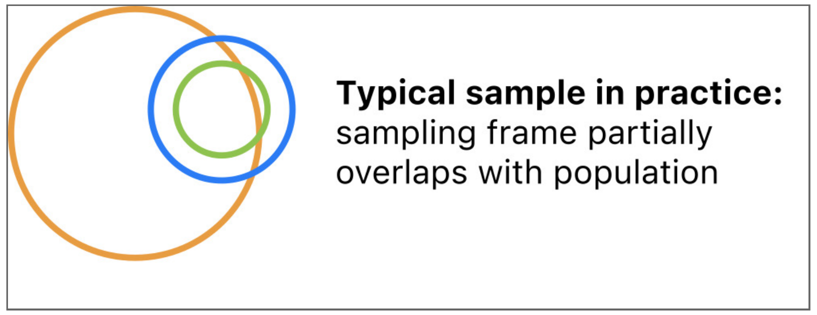
*From a SRS, sample properties are **reflective** of population properties.*

- Can safely extrapolate from the sample to the population.

Sampling scenarios: typical sample

More common in practice is a SRS from a sampling frame that overlaps but does not cover the population. In this case:

$$S \subset F \quad \text{and} \quad F \cap P \neq \emptyset$$



Typical sample in practice: sampling frame partially overlaps with population

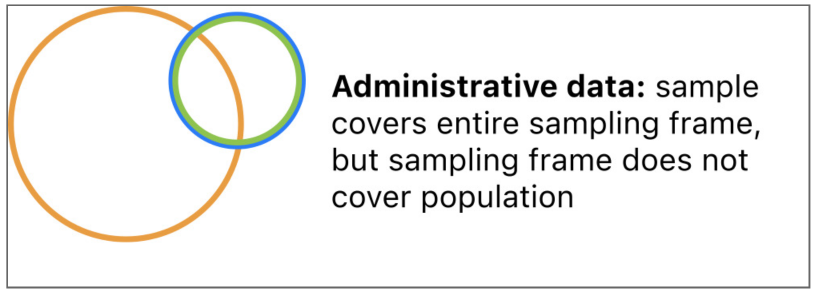
*In this scenario, sample properties are **reflective of the frame**.*

- Can extrapolate to a subpopulation, but not the full population.

Sampling scenarios: administrative data

Also common is **administrative data** in which all units are selected from a convenient frame that partly covers the population. In this case:

$$S = F \quad \text{and} \quad F \cap P \neq \emptyset$$



Administrative data: sample covers entire sampling frame, but sampling frame does not cover population

*Administrative data are **singular**; they do not represent any broader group.*

- No reliable extrapolation is possible.

Extrapolation: generalizing from samples

The relationships among the population, frame, and sample determine the **scope of inference**: the ***extent to which conclusions based on the sample are generalizable***.

A good sampling design can ensure that the statistical properties of the sample are expected to match those of the population. If so, it is sound to generalize:

- the sample is said to be *representative* of the population
- and the scope of inference is *broad*.

A poor sampling design will produce samples that distort the statistical properties of the population. If so, it is not sound to generalize:

- sample statistics are subjet to bias
- and the scope of inference is *narrow*.

Characterizing sampling designs

The sampling scenarios above can be differentiated along two key attributes:

1. The overlap between the sampling frame and the population.
   - frame $=$ population → census
   - frame $\subset$ population →
   - frame $\cap$ population $\neq \emptyset$ → typical, admin

2. ~~1~~. The mechanism of obtaining a sample from the sampling frame. "sampling mechanism"
   - census
   - random sample
   - probability sample
   - nonrandom (convenience) sample

If you can articulate these two points, you have fully characterized the sampling design.

Steps, given an unfamiliar dataset:

1. define Population, frame, sample
2. determine overlap $P \cap F$
3. determine mechanism

Inclusion probabilities

In order to describe sampling mechanisms precisely, we need a little terminology.
For any way of drawing a sample from a frame, each unit has some **inclusion probability** -- ***the probability of being included in the sample***.

Let's suppose that the frame F comprises N units, and denote the inclusion probabilities by:

$$p_i = P(\text{unit } i \text{ is included in the sample}) \quad *$$

The inclusion probability of each unit is usually determined by the physical procedure of collecting data, rather than fixed *a priori*.

* represents a physical process

Sampling mechanisms

Sampling mechanisms are ***methods of drawing samples*** and are categorized into four types based on inclusion probabilities.

- in a **census** every unit is included
  - $p_i = 1$ for every unit $i = 1, \ldots, N$
- in a **random sample** every unit is equally likely to be included
  - $p_i \propto \frac{1}{N}$ for every unit $i = 1, \ldots, N$
- in a **probability sample** units have different inclusion probabilities
  - $p_i \neq p_j$ for at least one $i \neq j$
- in a **nonrandom sample** inclusion probabilities are indeterminate
  - $p_i = ?$

typically arises from ad-hoc collection

Revisiting example datasets

Let's characterize the sampling designs of some of our example datasets.

GDP data

Annual observations of GDP growth for 234 countries from 1961 - 2018.

- Population: all countries existing between 1961-2019.
- Frame: all countries reporting economic output for some year between 1961 and 2019.
- Sample: equal to frame.

So:

1. Overlap: frame partly overlaps population.
2. Mechanism: sample is a census of the frame.

This is administrative data.
Scope of inference: none.

Revisiting example datasets

Let's characterize the sampling designs of some of our example datasets.

Mammal data

Observations of average brain and body weights for 62 mammal species.

- Population: all mammals.
- Frame: all mammal species that die of natural causes in monitorerd areas.
- Sample: individuals from 62 species sampled opportunistically.

So:

1. Overlap: frame is a subset of population.
2. Mechanism: convenience sampling.

We didn't give this a name, but let's call it 'convenience data'.
Scope of inference: none.

Revisiting example datasets

Let's characterize the sampling designs of some of our example datasets.

BRFSS data

Phone surveys of 418K U.S. residents in 2019.

- Population: all U.S. residents.
- Frame: all adult U.S. residents with phone numbers.
- Sample: 418K adult U.S. residents with phone numbers.

So:

1. Overlap: frame is a subset of the population.
2. Mechanism: probability sample.
    - Randomly selected phone numbers were dialed in each state.

This is a typical sample.
Scope of inference: adult residents with phone numbers.

Revisiting example datasets

Let's characterize the sampling designs of some of our example datasets.

Your turn: baby names

Records of given names of babies in CA from 1990 - 2018.

- Population: babies born in CA, 1990-2018
- Frame: babies w/ birth certificates from CA, 1990-2018
- Sample: 13M babies, equal to frame

So:

1. Overlap: partial, or maybe subset : $P \supset F$ or $F \cap P \neq \emptyset$
2. Mechanism: Census

This is * *data.* * administrative

Scope of inference: none

Revisiting example datasets

Let's characterize the sampling designs of some of our example datasets.

Your turn: SB weather data

Daily records of min/max temperatures at SB airport from January to March 2021.

- Population: all locations (region? state? US? world?)
- Frame: all locations possible to monitor
- Sample: 1 location

So:

1. Overlap: $F \subset P$
2. Mechanism: convenience sampling

This is * data. *convenience

Scope of inference: none

Statistical bias

Statistical **bias** is the average difference between a sample property and a population property across all possible samples under a particular sampling design.

- In short, arises from systematic over- or under-representation of units.

Bias is therefore determined by sampling design, *i.e.*, (i) the overlap between the frame and the population and (ii) the sampling mechanism.

When the frame is well-chosen and the mechanism involves a random selection process -- simple random samples or probability samples -- it is possible to quantify and correct bias.

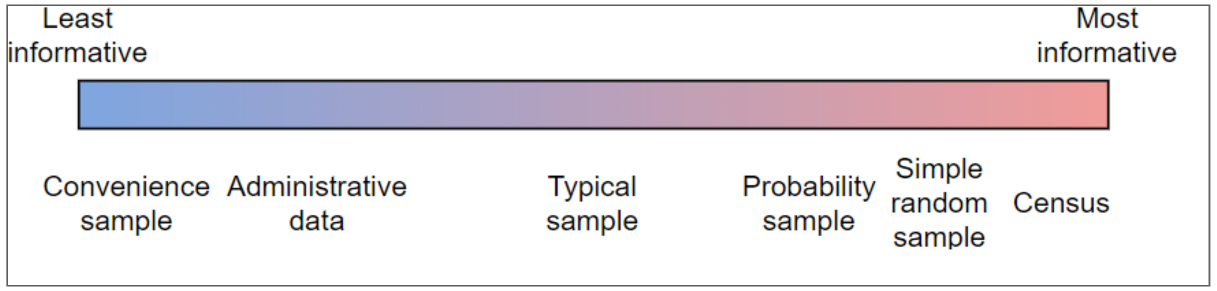
*You will explore this in **Lab 2** in the context of estimating population means using sample means under different sampling mechanisms.*

Data quality

It's easy to fall into value judgements about a dataset based on scope of inference and bias, but there's nothing inherently *better* or *worse* about the scenarios we've considered.

- Our goal in all this is to create a framework for understanding the limitations of a given dataset so that we can make pragmatic choices about how to use it that align with the information it has to offer.

It may help to map the scenarios we've considered onto an 'informativeness' spectrum:



The missing data problem

- Missingness modulates sampling design
- Types of missingness: MCAR, MAR, and MNAR
- Pitfalls and simple fixes

Missingness

> *The best-laid plans of mice and men often go awry.*

Missing data arise when ***one or more variable measurements fail for a subset of observations***.

This can happen for a variety of reasons, but is very common in pratice due to, for instance:

- equipment failure;
- sample contamination or loss;
- respondents leaving questions blank;
- attrition of study participants (dropping out).

Missingness gets a bad rap.

- It is often derided, downplayed, minimized, and overlooked.
- There is publication bias against studies with lots of missing data.
- Many researchers and data scientists ignore it by simply deleting affected observations.

But it's a reality in most datasets, and deserves attention.

Missing representations

It is standard practice to record observations with missingness but enter a special symbol (`..`, `-`, `NA`, etcetera) for missing values.

In python, missing values are mapped to a special float:

```
In [2]:
```

```python
float('nan')
```

```
Out[2]:
```

```
nan
```

Pandas has the ability to map specified entries to NaN when parsing data files.

Missing representations

Here is some made-up data with two missing values:

In [3]:

```
pd.read_csv('data/some_data.csv', index_col = 'obs')
```

Out[3]:

	value
obs	
0	-0.9286936933427271
1	-0.3088381742999848
2	-
3	-1.4345064041945543
4	0.03958917896644836
5	-
6	-0.5316890502224456
7	1.4734842645335422

Missing representations

Pandas has the ability to map specified entries to NaN when parsing data files:

In [4]:

```
some_data = pd.read_csv('data/some_data.csv', index_col = 'obs', na_values = '-')
some_data
```

Out[4]:

	value
obs	
0	-0.928694
1	-0.308838
2	NaN
3	-1.434506
4	0.039589
5	NaN
6	-0.531689
7	1.473484

Calculations with NaNs

NaNs halt calculations on numpy arrays.

In [5]:

```
# mean in numpy -- halt
some_data.values.mean()
```

Out[5]:

```
nan
```

However, the default behavior in pandas is to ignore the NaN's, which allows the computation to proceed:

In [6]:

```
# mean in pandas -- ignore
some_data.mean()
```

Out[6]:

```
value   -0.281776
dtype: float64
```

But here's the rub: those missing values could have been anything, and ignoring them changes the result from what it would have been!

In [7]:

```
# one counterfactual scenario
some_data.loc[[2, 5], 'value'] = [5, 6]
some_data.mean()
```

Out[7]:

```
value    1.163668
dtype: float64
```

The missing data problem

In a nutshell, the **missing data problem** is: ***how should missing values be handled in a data analysis?***

> *Getting the software to run is one thing, but this alone does not address the challenges posed by the missing data. Unless the analyst, or the software vendor, provides some way to work around the missing values, the analysis cannot continue because calculations on missing values are not possible. There are many approaches to circumvent this problem. Each of these affects the end result in a different way. (Stef van Buuren, 2018)*

There's no universal approach to the missing data problem. The choice of method depends on:

- the analysis objective;
- the *missing data mechanism*.

We will talk briefly about the latter.

Missing data in PSTAT100

We won't go too far into this topic in PSTAT 100. Our goal will be awareness-raising, specifically:

- characterizing types of missingness (missing data mechanisms);
- understanding missingness as a potential source of bias;
- basic do's and don't's when it comes to missingness.

If you are interested in the topic, **Stef van Buuren's *Flexible Imputation of Missing Data*** (the source of one of your readings this week) provides an excellent introduction.

Missing data mechanisms

The standard framework for understanding missingness is much like that for understanding sampling: just as every unit has a probability of being selected in a sample, every observation has some probablitiy of going missing.

A **missing data mechanism** is a ***process causing missingness***.

Suppose we have a dataset $\mathbf{x}$ (tidy) consisting of n rows/observations and p columns/variables, and define:

$$q_{ij} = P(x_{ij} \text{ is missing})$$

Missing data mechanisms are classified into three categories based on these probabilities:

1. Missing completely at random (MCAR)
2. Missing at random (MAR)
3. Missing not at random (MNAR)

MCAR

Data are **missing completely at random** (MCAR) if the ***probabilities of missing entries are uniformly equal***.

$$q_{ij} = q \quad \text{for all} \quad i, j$$

This implies that the cause of missingness is unrelated to the data: missing values can be ignored.

This is the easiest scenario to handle.

MAR

Data are **missing at random** (MAR) if the ***probabilities of missing entries depend on observed data***.

$$q_{ij} = f(\mathbf{x}_i)$$

This implies that information about the cause of missingness is captured within the dataset: it is possible to model the missing data.

Missing data methods typically address this scenario.

MNAR

Data are **missing not at random** (MNAR) if the ***probabilities of missing entries depend on unobserved data***.

$$q_{ij} = ?$$

This implies that information about the cause of missingness is unavailable.

This is the most complicated scenario.

Example

In the GDP growth data, growth measurements are missing for many countries before a certain year.

We might be able to hypothesize about why -- perhaps a country didn't exist or didn't keep reliable records for a period of time.

However, the data as they are contain no additional information that might explain the cause of missingness. So these data are MNAR.

Simple fixes

The easiest approach to missing data is to drop observations with missing values: `df.dropna()`.

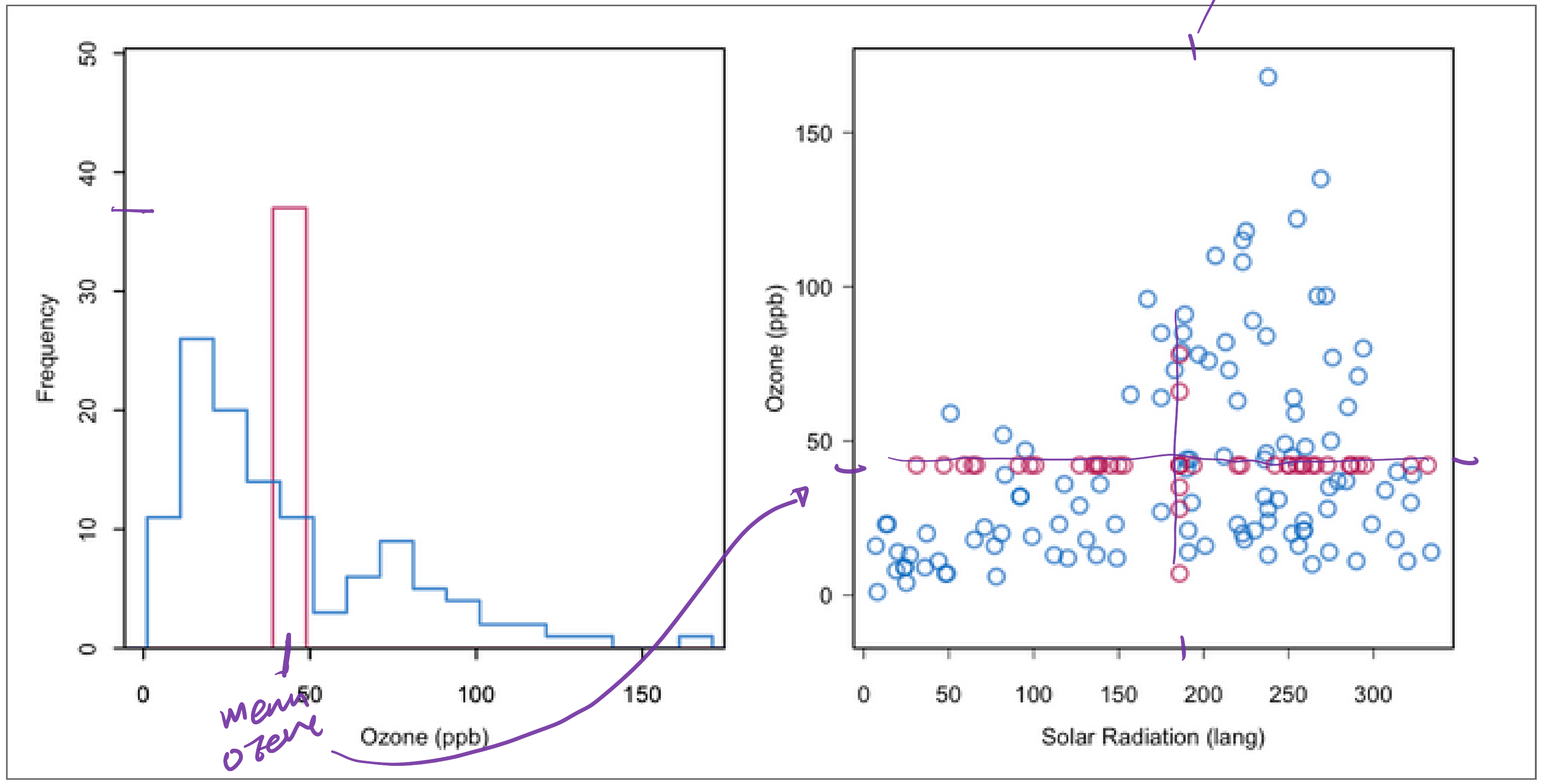
- Induces information loss, but is otherwise appropriate if data are MCAR.
- Induces bias if data are MAR or MNAR.

Another simple fix is *mean imputation*, filling in missing values with the mean of the corresponding variable: `df.fillna()`.

- Only a good idea if a very small proportion of values are missing.
- Induces bias if data are MAR or MNAR.

Perils of mean imputation

mean solar radiation



Imputing too many missing values distorts the distribution of sample values.

Do’s and don’t’s

Do:

1. Always check for missing values *upon import*.
    - Tabulate the proportion of observations with missingness
    - Tabulate the proportion of values for each variable that are missing
2. Take time to find out the reasons data are missing.
    - Determine which outcomes are coded as missing.
    - Investigate the physical mechanisms involved.
3. Report missing data if they are present.

Don't:

1. Use software defaults for handling missing values blindly.
2. Drop missing values if data are not MCAR.

Case study: voter fraud

- Steven Miller's analysis of 'Voter Integrity Fund' surveys
- Sources of bias
- Ethical issues

The Miller case

On November 21, 2020, a professor at Williams College, Steven Miller, filed an affidavit alleging that an analysis of phone surveys showed that among registered republican voters in PA:

- ~40K mail ballots were fraudlently requested;
- ~48K mail ballots were not counted.

> *President Donald J. Trump amplified the statement in a tweet, the Chairman of the Federal Elections Commission (FEC) referenced the statement as indicative of fraud, and a conservative group prominently featured it in a legal brief seeking to overturn the Pennsylvania election results. (Samuel Wolf, Williams Record, 11/25/20)*

The Miller affidavit was criticized by statisticians as incorrect, irresponsible, and unethical.

The flawed assumption

On a purely mathematical level, Miller's calculations were standard. The key issue was a single flawed assumption:

> *The analysis is predicated on the assumption that the responders are a* ***representative sample*** *of the population of registered Republicans in Pennsylvania for whom a mail-in ballot was requested but not counted, and responded accurately to the questions during the phone calls. (Miller affidavit)*

Essentially, Miller made two critical mistakes *in the analysis*:

1. Failure to critically assess the sampling design and scope of inference.
2. Ignored missing data.

We will conduct a *post mortem* and examine these issues.

Miller is a number theorist, not a trained survey statistician, so on some level his mistakes were understandable, but they did a lot of damage.

Sampling design

> *There were 165,412 unreturned mail ballots requested by registered republicans in PA.*

Those voters were surveyed by phone by Matt Braynard's private firm External Affairs on behalf of the Voter Integrity Fund.

We don't really know how they obtained and selected phone numbers or exactly what the survey procedure was, but here's what we do know:

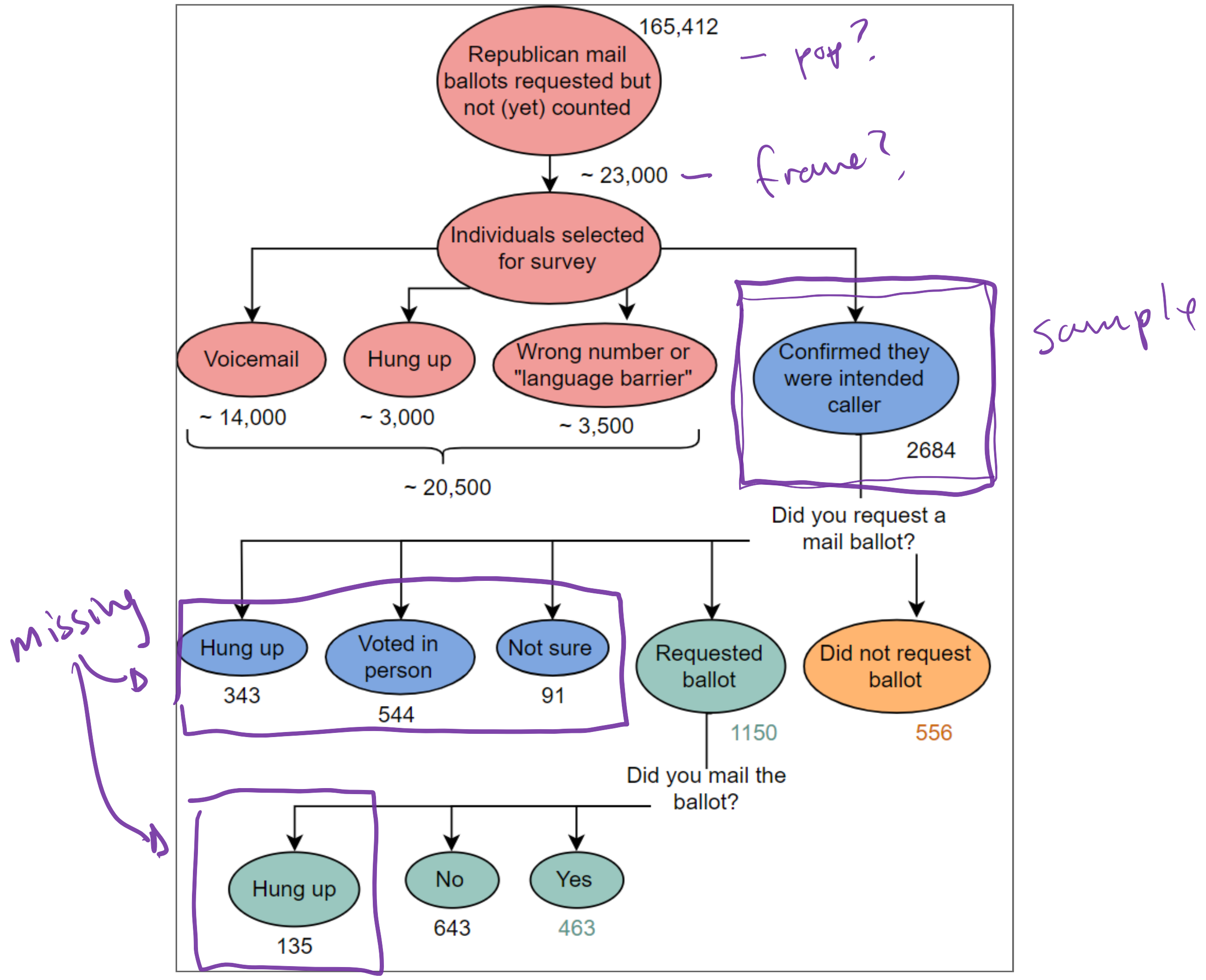
1. ~23K individuals were called on Nov. 9-10.
2. The ~2.5K who answered were asked if they were the registered voter or a family member.
3. If they said yes, they were asked if they requested a ballot.
4. Those who requested a ballot were asked if they mailed it.

Spot any immediate issues?

We'll look in greater detail at these steps, but can you spot any obvious fishiness?

- ~23K individuals were called on Nov. 9-10
    - How did they pick who to call?
    - Narrow snapshot in time.
    - 9th and 10th were a Monday and Tuesday.
    - Mail ballots were still being counted; don't actually know whether returned ballots were ultimately counted or not by this time.
- The ~2.5K who answered were asked if they were the registered voter or a family member.
    - Family members could answer on behalf of one another.
- If they said yes, they were asked if they requested a ballot.
    - Misleading question: there's a registration checkbox; you don't have to file an explicit request in Pennsylvania.
- Those who requested a ballot were asked if they mailed it.
    - What about voters who claimed not to request a ballot? Did they receive one, and if so, did they mail it?

Survey schematic



non response : $\frac{20500}{23000} = \frac{205}{230} = 0.891$

missing $\frac{343+544+91}{2684} = \frac{978}{2684} = 0.364$

missing $\frac{135}{1150} = 0.117$

Sampling design

Population: republicans registered to vote in PA who had mail ballots officially requested that hadn't been returned or counted by November 9?

Sampling frame: unknown; source of phone numbers unspecified.

Sample: 2684 registered republicans or family members of registered repbulicans who had a mail ballot officially requested in PA and answered survey calls on Nov. 9 or 10.

Sampling mechanism: nonrandom; depends on availability during calling hours on Monday and Tuesday, language spoken, and willingness to talk.

This is not a representative sample of any meaningful population.

Missingness

Respondents hung up at every stage of the survey.
This is probably not at random -- individuals who do not believe voter fraud occurred are more likely to hang up.
However, we don't have any information about whether respondents think fraud occurred.
So data are MNAR, and likely over-represent people more likely to claim they never requested a ballot.

The analysis

Miller first calculated:

- The proportion of respondents who reported not requesting ballots among those who either voted in person, didn't request a ballot, or did request a ballot.
    - Ignored those who weren't sure and those who hung up.
    - Claimed that the estimated number of fraudulent requests was:

$$\left(\frac{556}{1150+556+544}\right) \times 165,412 = 0.2471 \times 165,412 = 40,875$$

Simulation

It's not too tricky to envision sources of bias that would affect Miller's results. *How much bias might there be?* This is an oversimplification, but if we are willing to assume that

1. respondents all know whether they actually requested a ballot and tell the truth,
2. respondents who didn't request a ballot are more likely to be reached, and
3. respondents who did request a ballot are more likely to hang up during the interview,

then we can show through a simple simulation that an actual fraud rate of under 1% will be estimated at over 20% almost all the time.

Simulated population

First let's generate a population of 150K voters.

In [8]:

```
np.random.seed(41021)

# proportion of fraudlent requests
true_prop = 0.009

# generate population of 100K; 100 of 100K did not request a ballot
N = 150000
population = pd.DataFrame(data = {'requested': np.ones(N)})
num_nrequest = round(N*true_prop) - 1
population.iloc[0:num_nrequest, 0] = 0
```

Simulated sample

Then let's introduce sampling weights based on the conditional probability that an individual will talk with the interviewer given whether they requested a ballot or not.

In [9]:

```
# assume respondents tell the truth
p_request = 1 - true_prop
p_nrequest = true_prop

# assume respondents who claim no request are 15x more likely to talk
talk_factor = 15

# observed nonresponse rate
p_talk = 0.09

# conditional probability of talking given claimed request or not
p_talk_request = p_talk/(p_request + talk_factor*p_nrequest)
p_talk_nrequest = talk_factor*p_talk_request

# draw sample weighted by conditional probabilities
np.random.seed(41021)
population.loc[population.requested == 1, 'sample_weight'] = p_talk_request
population.loc[population.requested == 0, 'sample_weight'] = p_talk_nrequest
samp = population.sample(n = 2500, replace = False, weights = 'sample_weight')
```

Simulated missing mechanism

Then let's introduce missing values at different rates for respondents who requested a ballot and respondents who didn't.

In [10]:

```
# assume respondents who affirm requesting are 4x more likely to hang up or deflect
missing_factor = 4

# observed missing/unsure rate
p_missing = 0.25

# conditional probabilities of missing given request status
p_missing_nrequest = p_missing/(0.8 + missing_factor*0.2)
p_missing_request = missing_factor*p_missing_nrequest

# input missing values
np.random.seed(41021)
samp.loc[samp.requested == 1, 'missing_weight'] = p_missing_request
samp.loc[samp.requested == 0, 'missing_weight'] = p_missing_nrequest
samp['missing'] = np.random.binomial(n = 1, p = samp.missing_weight.values)
samp.loc[samp.missing == 1, 'requested'] = float('nan')
```

Simulated result

If we then drop all the missing values and calculate the proportion of respondents who didn't request a ballot, we get:

In [11]:

```
# compute mean after dropping missing values
1 - samp.requested.mean()
```

Out[11]:

```
0.21206743566992015
```

So Miller's result is *expected* if the sampling and missing mechanisms introduce bias, even if the true rate of fraudulent requests is under 1% -- on the order of 1,000 ballots.

Takeaways

The main mistakes were ignoring the sampling design and missing data -- in other words, proceeding to analyze the data without first getting well-acquainted. We should assume these were honest mistakes.

After the affidavit was filed, a colleague spoke with Miller; he recanted and acknowledged his mistakes, but this received far less attention than the conclusions in the affidavit.

Professional ethics and social responsibility

The American Statistical Association publishes **ethical guidelines for statistical practice**. The Miller case violated a large number of these, most prominently, that an ethical practitioner:

- Reports the sources and assessed adequacy of the data, accounts for all data considered in a study, and explains the sample(s) actually used.
- In publications and reports, conveys the findings in ways that are both honest and meaningful to the user/reader. This includes tables, models, and graphics.
- In publications or testimony, identifies the ultimate financial sponsor of the study, the stated purpose, and the intended use of the study results.
- When reporting analyses of volunteer data or other data that may not be representative of a defined population, includes appropriate disclaimers and, if used, appropriate weighting.

Summary

This week we've touched on sampling design and missing data.

Terminology: population, sampling frame, sample, sampling mechanism, missing data mechanism.

- Sampling mechanisms: census; random sample, probability sample; convenience (nonrandom) sample.
- Missing data mechanisms: completely at random (MCAR); at random (MAR); not at random (MNAR).

Key concepts:

- Sampling design and missing data handling determine the reliability and scope of inferences.
- Either poor design or inappropriate handling of missing data can induce bias and generate misleading results.

The Miller case study illustrated how easy it is to overlook these issues, and the potential social impact of statistical malpractice.