Sampling: Errors

```
$ echo "Data Sciences Institute"
```

Learning Outcomes

How might your sampling and surveying approach cause inaccuracies in your data?

- Ability to identify sources of error in sampling and survey methodology.

- Ability to distinguish between different types of errors, such as variance and bias, sampling and non-sampling errors, etc.

Variance and Bias

Variance and Bias

Variance	Bias
- Random error - Describes variability of calculated quantities - Want to minimize	- Systematic error - Describes difference between calculated and true quantities (i.e. between sample statistics and population parameters) - Ideally, we want to eliminate bias. In practice, this is often not possible.

- Lohr, 2019, Figure 2.3

  Which of these displays high variance and which displays high bias?
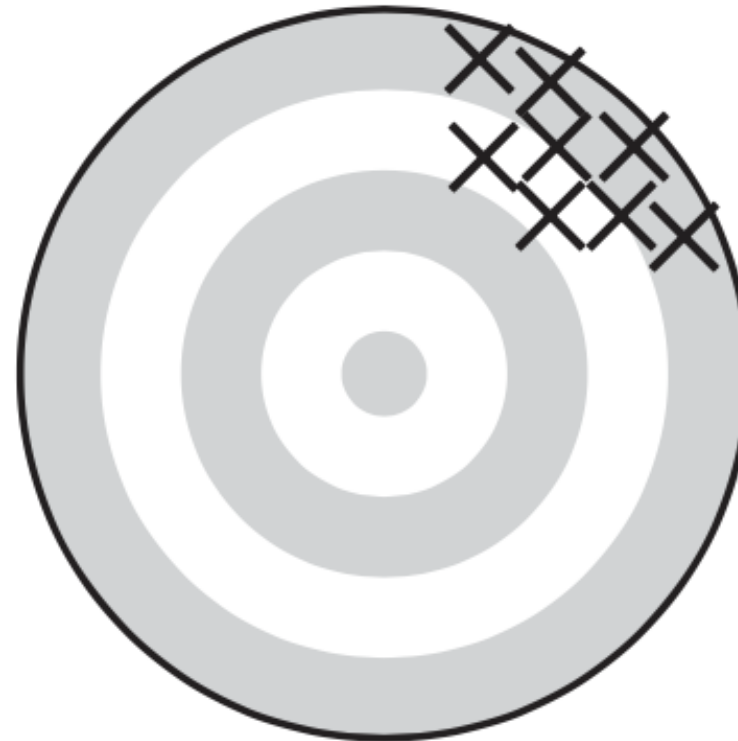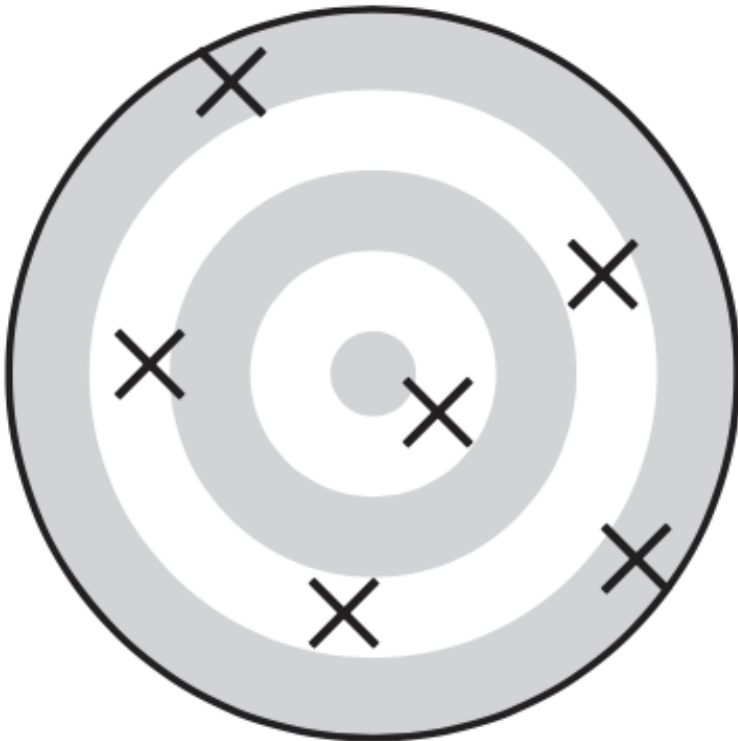
Types of Errors

Sampling Errors

- **Sampling error** is the error that results from taking a given sample instead of measuring the entire frame population (error from variance)

- Different samples will likely produce different sample statistics, and these sample statistics will likely be different from the true population parameter

- Sampling error can only be avoided if the frame population and sample population are exactly the same

- ⚠️ **Generally**: Smaller sample size produces greater sampling error ⚠️

Non-Sampling Error

- **Non-sampling error** is error that does not occur as the result of variability between different samples

- Often systemic (i.e. a result of study design or pre-existing characteristics of the population)

- Examples: selection bias, measurement error, nonresponse

Selection Bias

- **Selection bias** occurs when some population units are unintentionally excluded from the sample population

- Results in a non-representative sample

- Sources:
  - Using a sample selection procedure that, unknown to the investigators, depends on some characteristic associated with the properties of interest
  - Substituting a convenient member of a population for a designated member who is not readily available
  - Allowing the sample to consist entirely of volunteers
  - Coverage error

Selection Bias produces Coverage Error

- **Coverage error** is when the sampling frame does not match the target population
  - *Undercoverage* = Failing to include all the target population in the sample frame
  - *Overcoverage* = Including units in the sample frame that are not in the target population
- **Coverage bias** is when coverage error makes sample estimates differ from the population value

How do we measure coverage?

- It is difficult to measure coverage, since if missing populations were easily identifiable and accessible they would already be included in the frame.

- Ways to assess coverage:
  - Compare estimates of demographic characteristics to known values from the population
    - For example, if your frame contains 75% men and 25% women, it is likely that women are undercovered
  - Compare coverage rate or estimates with an external study or data source
    - For example, coverage of households with infants could be assessed by comparing the sampling frame with recent birth records in the area of interest

Measurement Error

- **Measurement error** occurs when survey responses tend to differ from the true population value

- **Measurement bias** occurs when measurement errors tends to occur consistently in one direction

- Sources:
  - Respondent untruth, misunderstanding, or forgetfulness
  - Trying to impress an interviewer
  - Presentation (question order, interviewer persona)

Non-Response

- Sampled individuals who participate in a survey or study are called **respondents**

- **Non-response error** occurs when members of the sample do not respond to the survey or study (i.e. when the respondents to a survey are only a subset of the sample)

- **Non-response bias** occurs when respondents are systematically different from those who didn't respond

Types of Non-Response

- **Unit nonresponse** = When an entire sampling unit is missing (e.g. a person does not respond)

- **Item nonresponse** = When a specific measurement or variable for a given sampling unit is missing (e.g. a person responds but does not answer a particular question)

Non-response can lead to...

- Non-representative samples

- Bias!

- Misallocation of resources

What is an acceptable response rate?

- Depends on the nature of nonresponse
  - If nonresponse bias is low, a lower response rate is fine
  - If nonresponse bias is high, any rate/amount of respondents may still produce invalid results
- Regardless of its value, it is important that the response rate is calculated consistently and reported with survey results.

Processing Error

- Errors in data entry or editing

- Sources:
  - Incorrect transcription
  - Typos in data entry
  - Open-ended question (e.g. trying to code multiple responses as a single response)

Total Survey Error Framework

Total Survey Error = Representation Errors + Measurement Errors + Processing Errors

- The total survey error framework states that there are two types of error – **bias** and variance – and there are three sources of these errors – **representation** (e.g. coverage, sampling error, nonresponse), **measurement**, and processing.

"Representation is about making inferences from your respondents to your target population" – Salganik, 2018
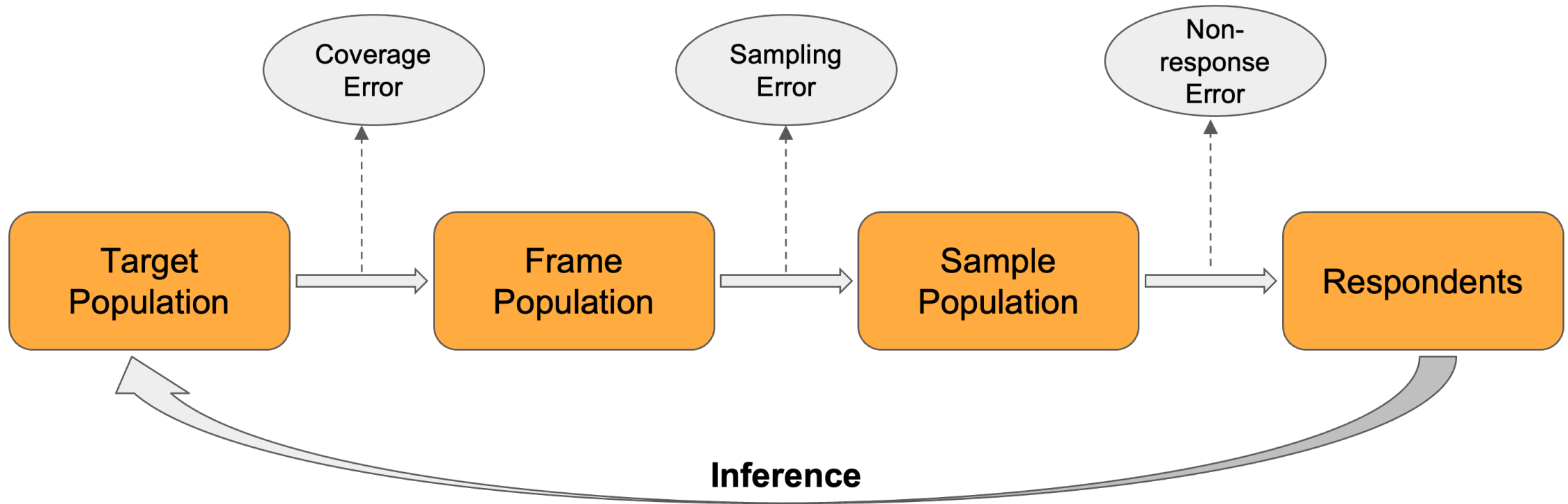
Activity: Evaluating Sampling Strategies

Suppose U of T has 10,000 part-time students (the population). We are interested in the average money a part-time student spends on books. We take two different samples. First, we use convenience sampling and survey ten students from an organic chemistry class. Many of these students are taking first term calculus. The amount of money they spend on books is as follows: $128, $87, $173, $116, $130, $204, $147, $189, $93, $153 The second sample is taken using a list of senior citizens who take P.E. classes and taking every fifth senior citizen on the list, for a total of ten senior citizens. They spend: $50, $40, $36, $15, $50, $100, $40 $53, $22, $22. It is unlikely that any student is in both samples.

Do you think that either of these samples is representative of the entire 10,000 part-time student population? Why?

Activity: Evaluating Sampling Strategies

A local radio station has a fanbase of 20,000 listeners. The station wants to know if its audience would prefer more music or more talk shows. Asking all 20,000 listeners is an almost impossible task. The station uses convenience sampling and surveys the first 200 people they meet at one of the station's music concert events. 24 people said they'd prefer more talk shows, and 176 people said they'd prefer more music.

Do you think that this sample is representative of the entire 20,000 listener population? Why?

Activity: Evaluating Sampling Strategies

Airline companies are interested in the consistency of the number of babies on each flight, so that they have adequate safety equipment. Suppose an airline conducts a survey. Over Thanksgiving weekend, it surveys six flights from Toronto to Montreal to determine the number of babies on the flights. It determines the amount of safety equipment needed by the result of that study.

1. List three potential sources of error in how the survey was conducted.
2. List three ways that you would improve the survey if it were to be repeated.

Next

- Survey Quality

- Questionnaire Design

- Ethics