Survey Research Methodology I

Computational Social Science UC3M
Fall 2023-2024

Sebastian Daza

Overview

- Presentations
- Review of syllabus and structure of the course
  - Zotero
  - Github
- Results from student's survey
- Intro to survey methodology
- Total survey error framework

Syllabus

- https://github.com/sdaza/survey-methods
- All material will be there (so check it often)
    - Data
    - Code
    - Lectures
    - Assignments and final project
- If you are not familiar with `git`, try using GitHub Desktop

Note about code

- Examples and simulations
  - Jupyter notebooks
- I recommend you use Visual Studio Code
  - Integrated development environment (IDE)
  - https://colab.research.google.com
- Mostly R
  - `data.table`
  - More info here

Emojis!

🤔 = Expecting your participation and discussion

💻 = We will do some live coding

🤓 = Extra knowledge

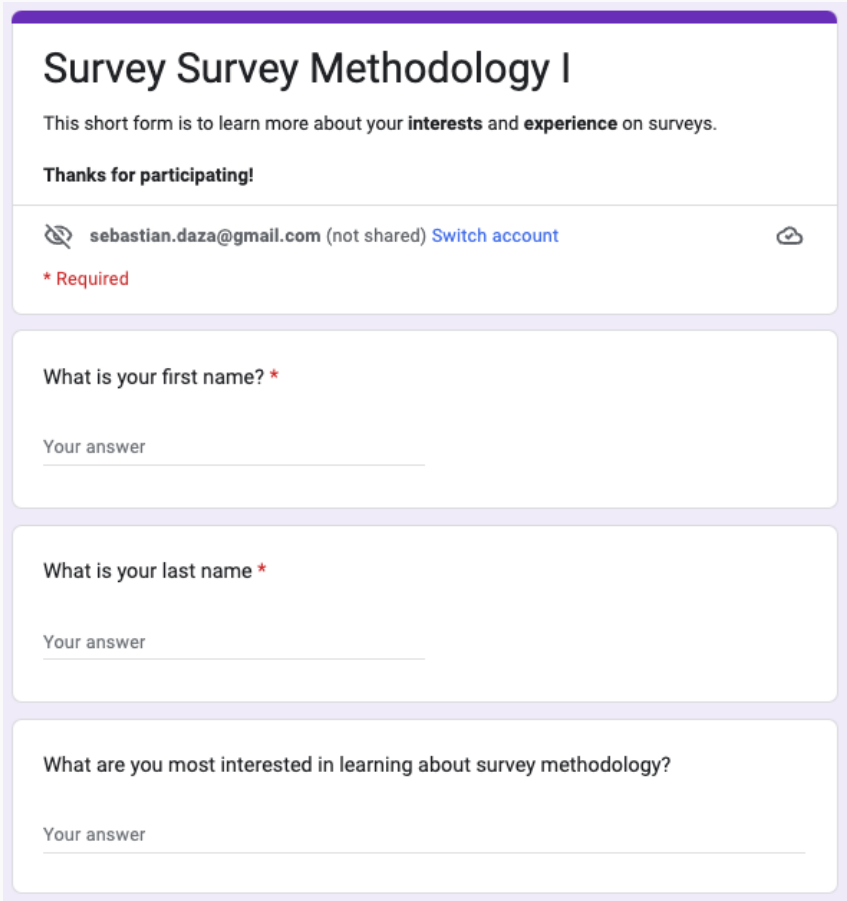
🥷 = Ninja level

Students’ Survey

Response rate

2023: 21 out of 31, 68%

2022: 15 out of 21, 71%



Survey Survey Methodology I

This short form is to learn more about your **interests** and **experience** on surveys.

Thanks for participating!

sebastian.daza@gmail.com (not shared) Switch account

* Required

What is your first name? *

Your answer

What is your last name *

Your answer

What are you most interested in learning about survey methodology?

Your answer

Interests on Survey Methodology

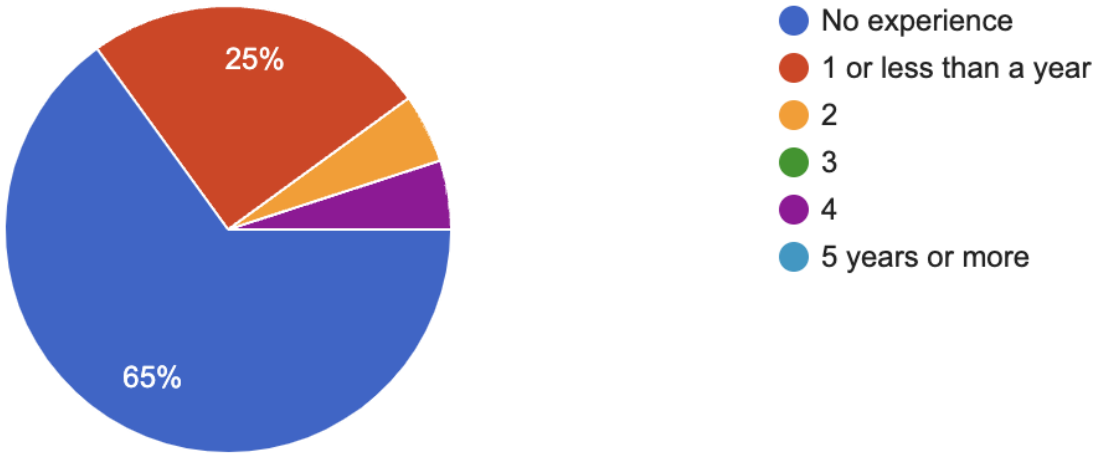
Analyzing and interpreting survey results

Sample design, validity and reliability

Experience: Design

How much experience do you have in conducting or designing social surveys?

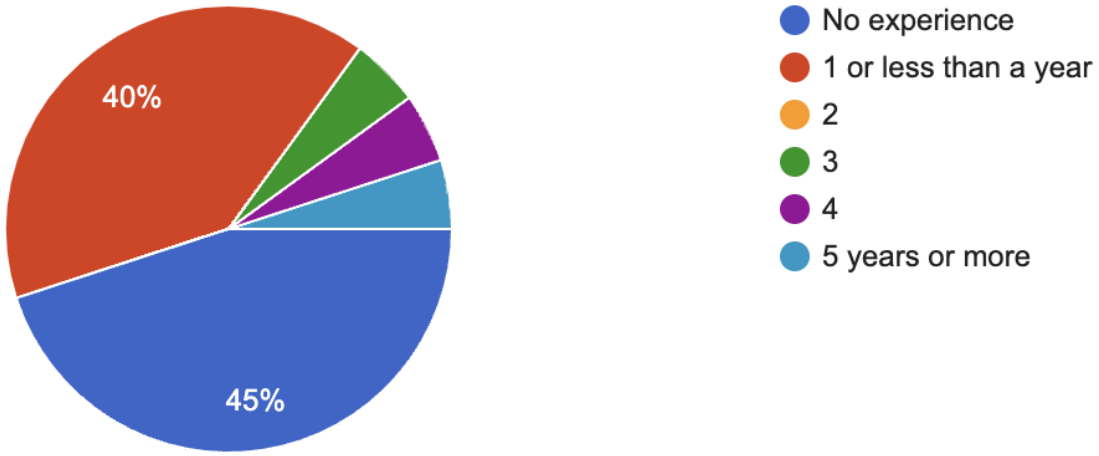
20 responses



Experience: Analysis

How much experience do you have in analyzing social surveys?

20 responses



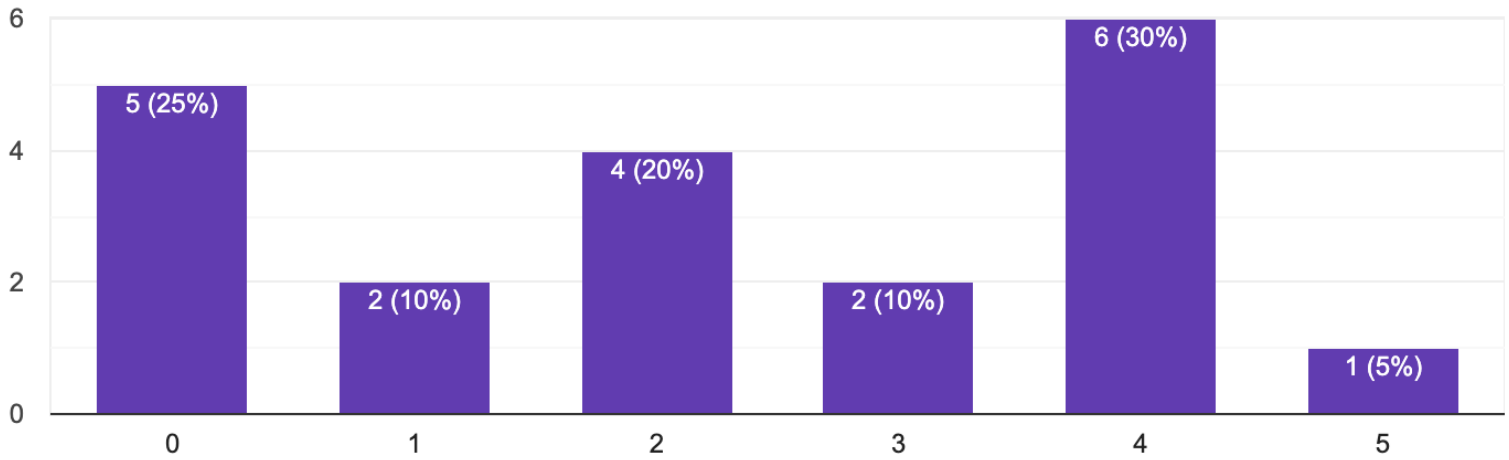
Statistical Power

Mean = 2.487

How familiar are you with the concept of statistical power?

Copy

20 responses

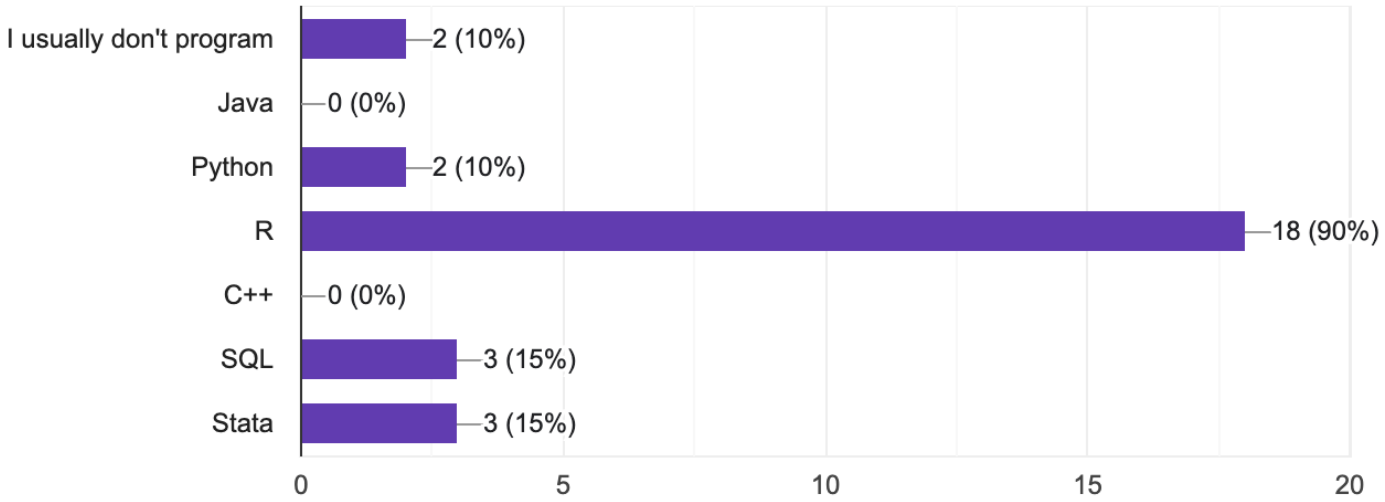


Programming

What languages do you usually use for programming?

Copy

20 responses

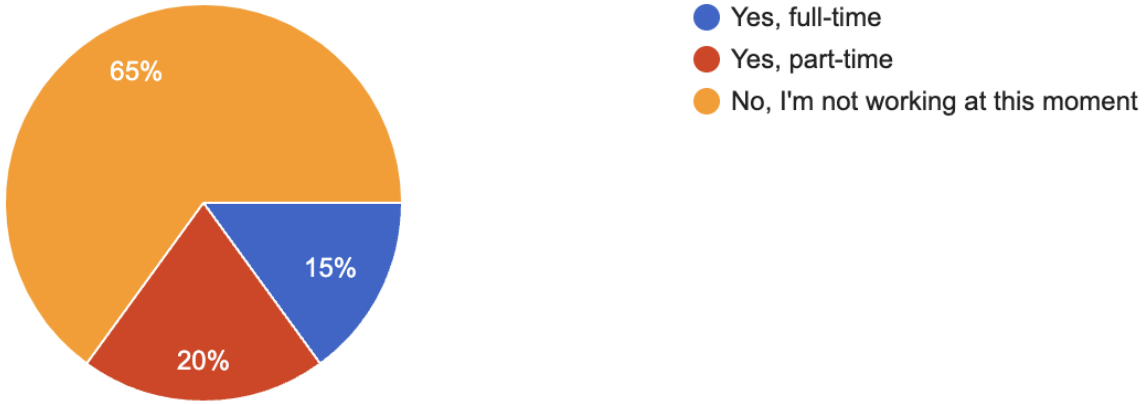


Working

Are you currently working?

20 responses

Copy



Future Position

Intro Survey Methodology

What is all this about?

> *Seeks to identify principles about the design, collection, processing, and analysis of surveys that are linked to the cost and quality of survey estimates (Groves et al. 2009)*

- A scientific field and profession
- Multidisciplinary

Some history

Four basic developments form the core of the modern sample survey method

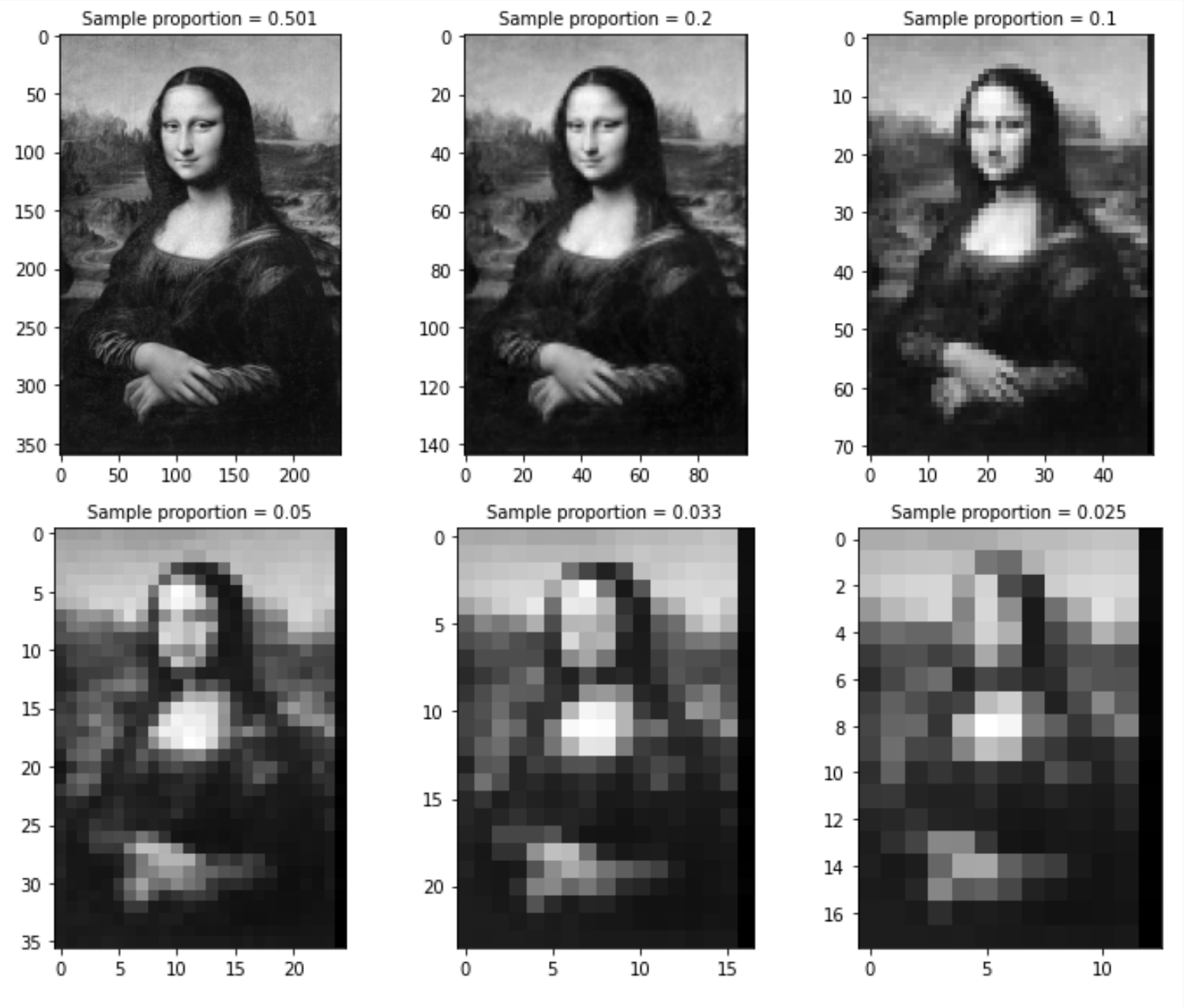
- Sampling: from samples $\rightarrow$ unbiased estimates
- Inference: statistics, margins of errors
- Measurement: the art of asking questions
- Analysis: multivariate, complex survey designs

Some history

Before the development of theories of probability, researchers had no basis for generalizing sample data to estimate population characteristics, so they tended to study the entire population (censuses)

The closer to complete enumeration one could come, the better!

Sampling intuition



Sampling theory adoption

- Neyman's seminal paper (1934): foundation of sampling theory
  - Adoption of Neyman's ideas, however, was slow
- Market research (1920's) operated on a completely different model that did not imitate censuses
  - Early product testing asked a convenient set of consumers (samples) to express preferences
- Political polls began to appear (1930s) and try to solve the sampling problem with quota methods
- Survey research did not begin to enter universities until the late 1930s.

Turning point

Without question, the turning point came in 1936, when Gallup's preelection polls, based on carefully drawn but relatively small (quota) samples of the US population ($\sim$ 5000 respondents), correctly predicted Roosevelt's victory, while the Literary Digest poll, based on millions of straw ballots mailed to known phone subscribers and Literary Digest subscribers, forecasted a victory for Republican Alf Landon. This David versus Goliath contest showed that a carefully implemented age, sex, and region quota sample was superior to a low-return (about 15%) mail survey covering better-off households.

Setback to move forward

Political polls drew renewed attention when they failed to predict the outcome of the 1948 election pitting the incumbent Truman against popular New York governor Dewey. They did show evidence of a late Truman surge, but even the final Gallup and Crossley polls forecast Dewey as the victor, albeit by a steadily decreasing margin. Investigation into what had gone wrong concluded that the quota sampling approach was partly to blame. Multistage area probability samples (with a random selection of respondents within households) developed at the Census Bureau, became the sampling method of choice, and remain so now.

Evolution

- By the late 1960s, the sample survey had become well-established as the method of choice for much data collection in social sciences
- Many reputable departments have local survey research centers
- American Association of Public Opinion Research (AAPOR)
    - *Public Opinion Quarterly*
    - Integrate commercial/polling and academic research
- Inter-University Consortium for Political and Social Research (ICPSR)
- Council of American Survey Research Organizations (CASRO) $\sim$ Market research

Landscape (in the US)

- Academia
  - NORC (Chicago)
  - SRC (Michigan)
- Private
  - RTI International (North Carolina)
  - Westat (Maryland)
  - RAND (California)
  - Pew Research Center
- Government
  - US Bureau of the Census
  - Bureau of Labor Statistics
  - National Center for Health Statistics
- Media polls
  - Major political polls
  - ABC News, IPSOS, CBS News, Fox News, NYTimes

Would you like to conduct a survey?

Many decisions

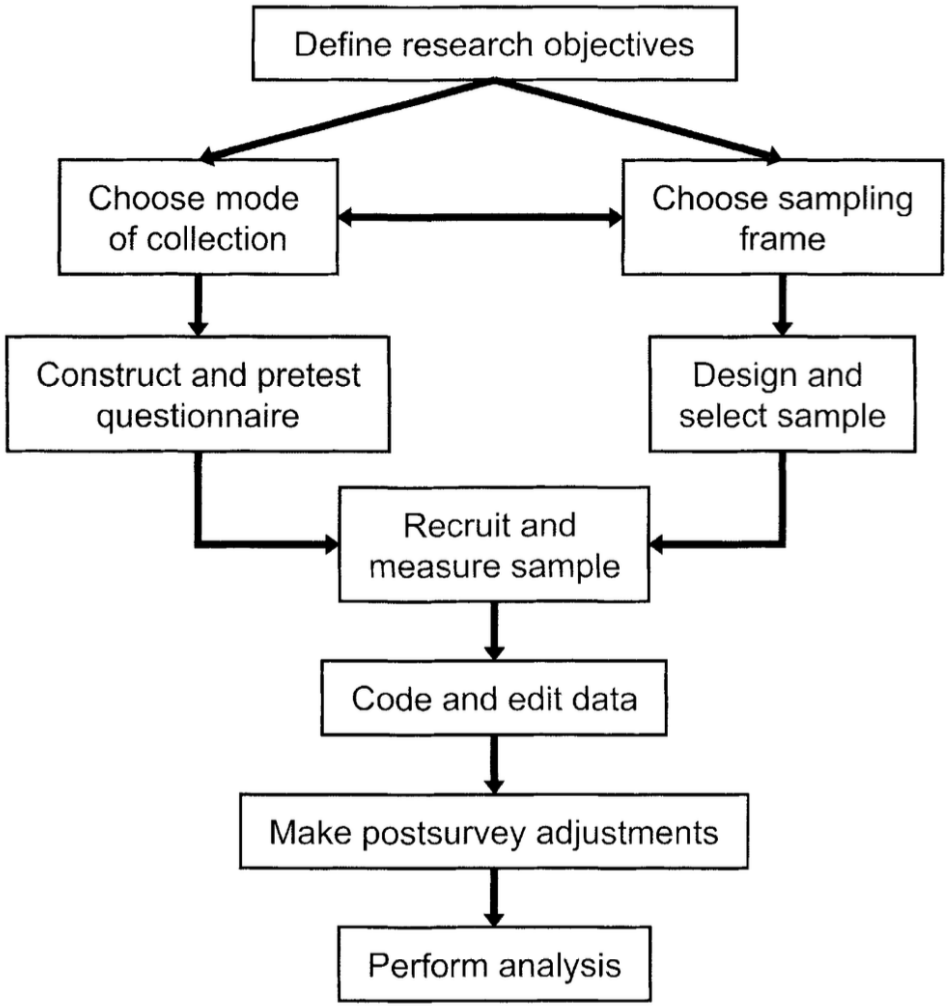
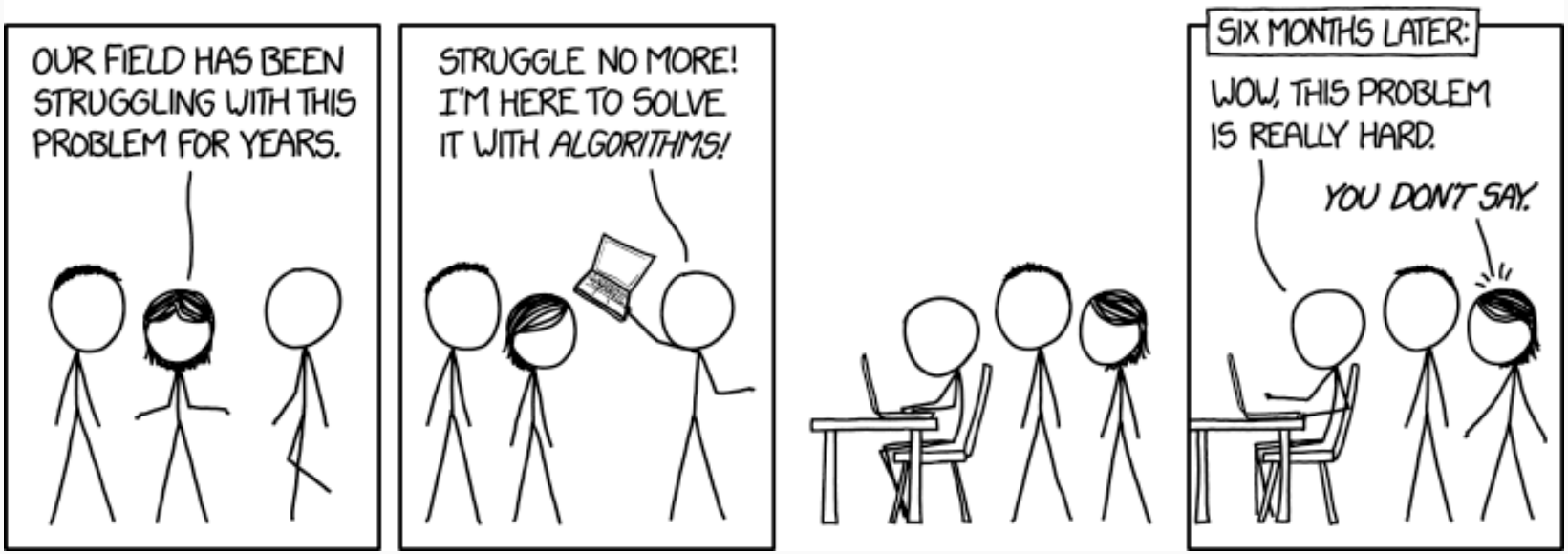


Figure 2.4 A survey from a process perspective.

Many Decisions

- Making a broad range of decisions regarding various aspects of a survey.
- These decisions can potentially influence the accuracy of estimates derived from surveys.
- Surveys, conducted in unregulated environments of the real world, can be influenced by these settings.

Complexity of social problems



The problems of social science are hard not just for social scientists but for everyone, even physicists

Key Challenges

- Optimize the use of available resources.
- Balance investments across all survey components to maximize the value of the resulting data.
- Instead of focusing on only a few elements of a survey, consider all elements as a whole - a total survey error approach.
- Be aware of numerous trade-offs.

All surveys are not created equal

Total survey error

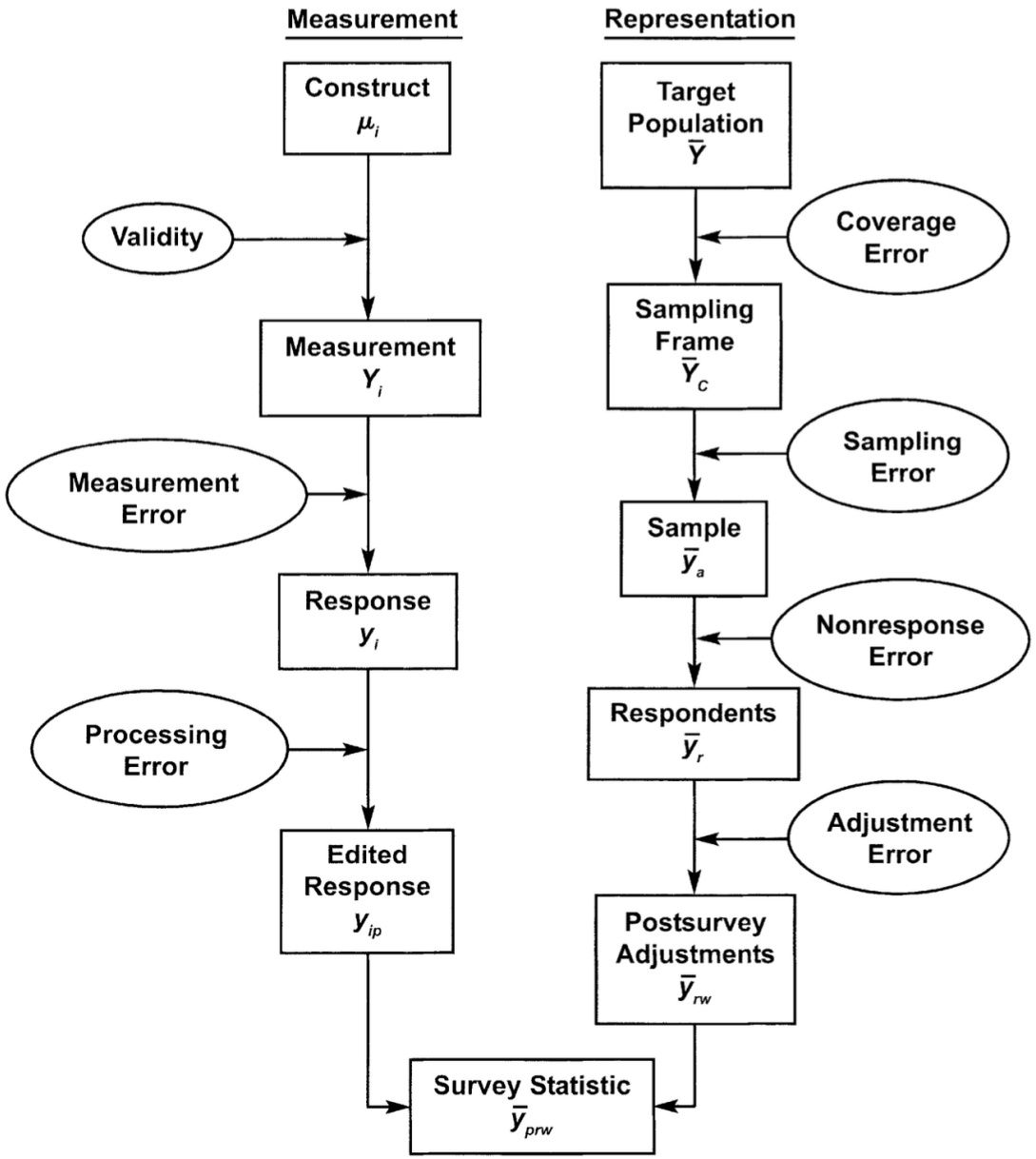
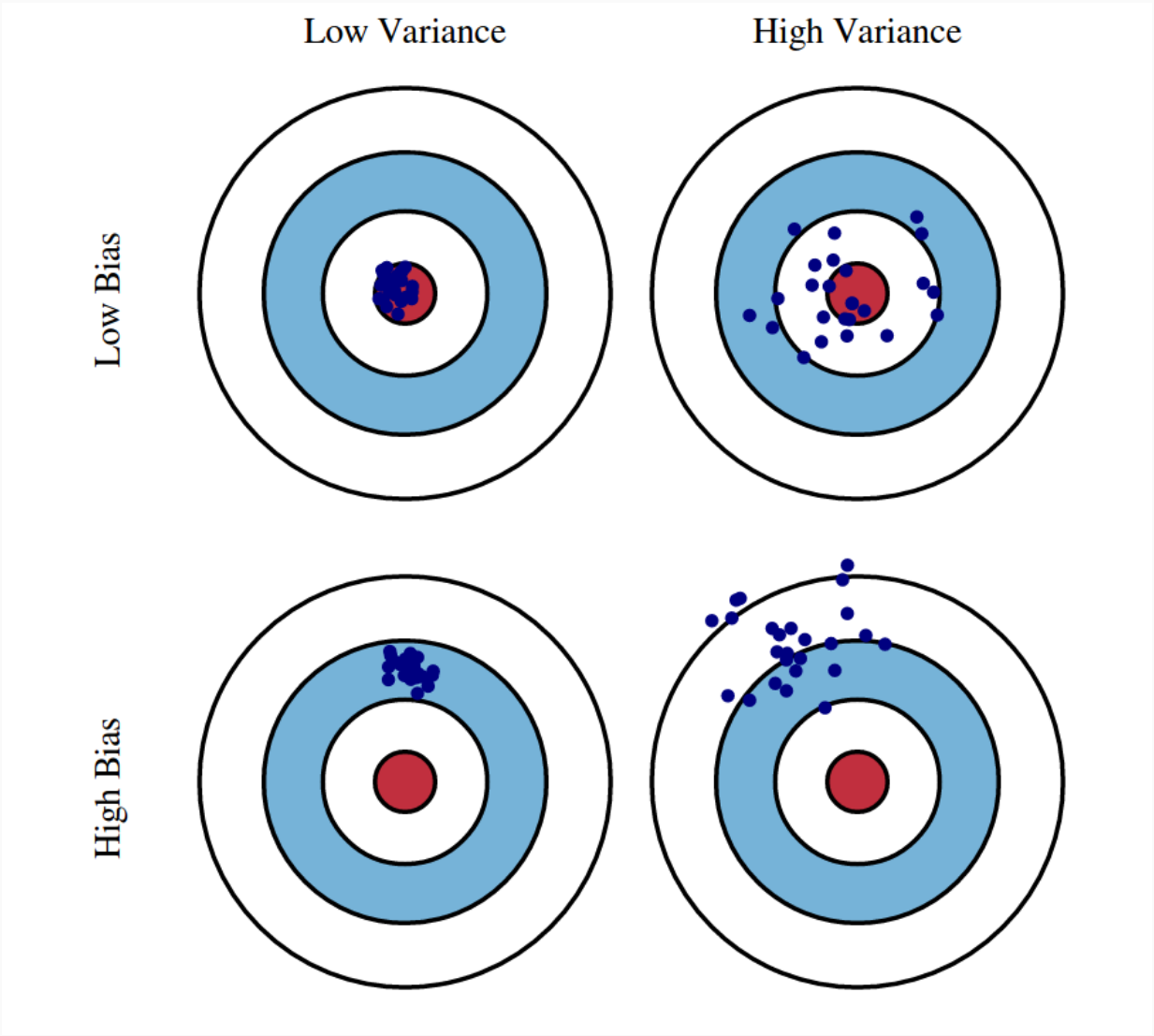
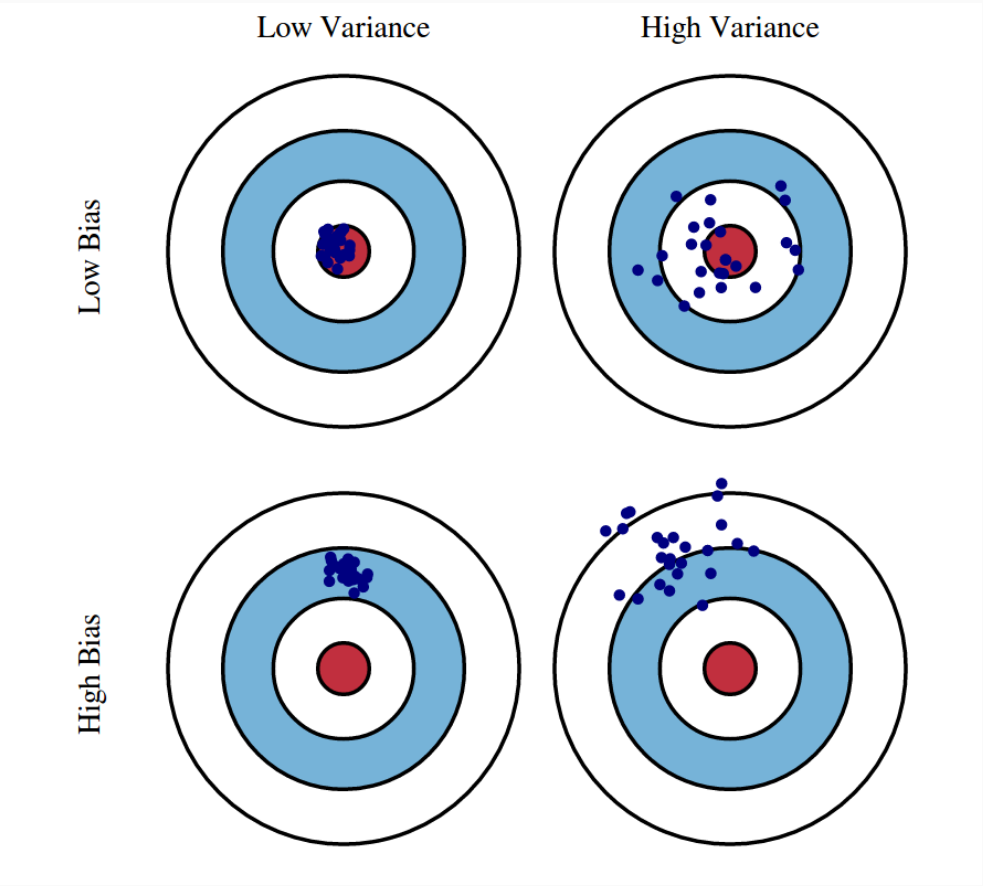


Figure 2.5 Survey life cycle from a quality perspective.

What do we mean by error? 🤔



What do we mean by error?



- **Bias:** the difference between expected value (e.g., measures) and the true value being estimated
- **Variance:** how variable your measures are

What kind of error?

$$y_i = \mu_i + \epsilon_i$$

- y_i: the observation for characteristic y for unit i
- μ_i: true value of the characteristic of interest
- ϵ_i: observation error (which may be positive for some individuals and negative for others)

💻 Let's move to **R** a bit...

Error and quality perspective

We can focus on total error and accuracy, but there are also other dimensions...

Survey quality is a complex and multidimensional concept

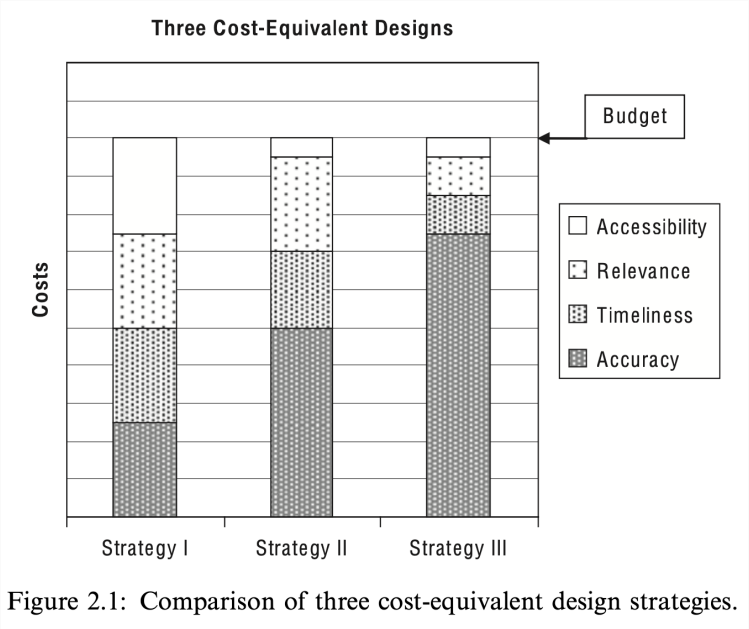


Figure 2.1: Comparison of three cost-equivalent design strategies.

The goal is to minimize total survey error subject to cost constraints while accommodating other user-specified quality dimensions

Real request 🤔

The government of your country, by direct order of the president or prime minister, asks you to conduct a weekly opinion survey that evaluates the government's approval, as well as the opinion of citizens on current issues and public policy. The results report should be on the president's desk by 5 pm every Sunday.

What design would you propose?

What criteria would you prioritize?

How would you verify that the survey works well?

How much would it cost?

Total survey error

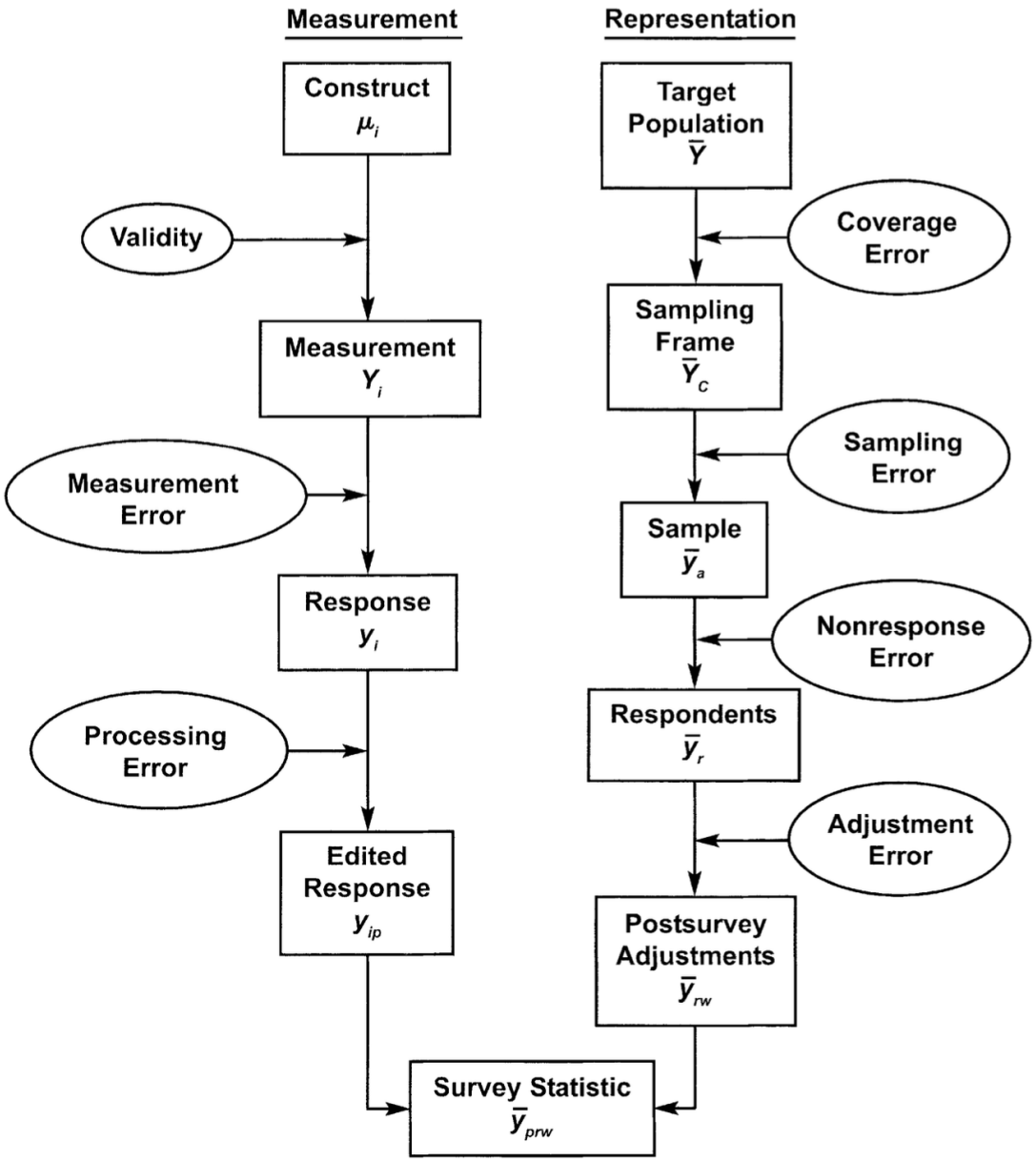
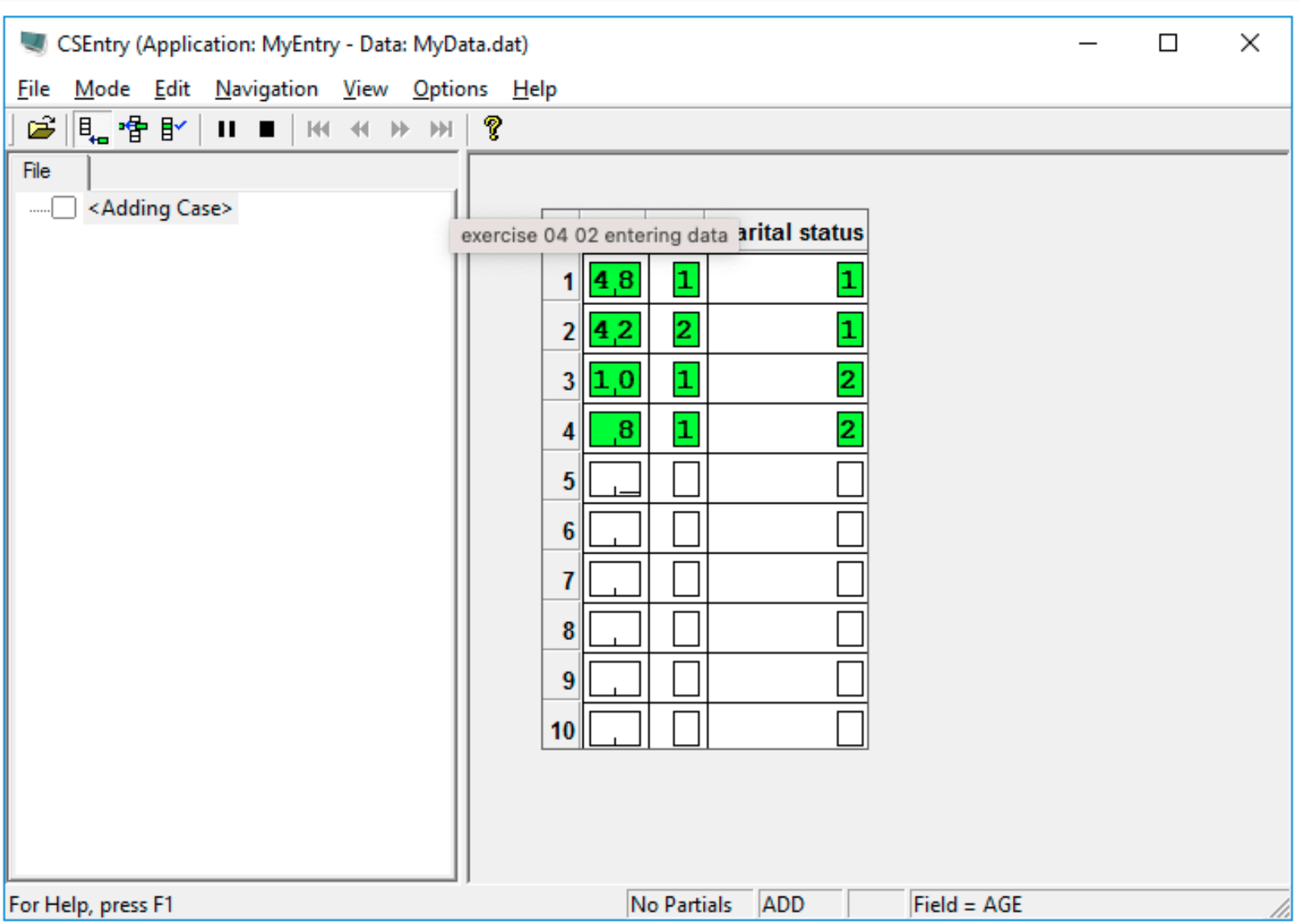


Figure 2.5 Survey life cycle from a quality perspective.

Processing error

Census and Survey Processing System (CSPro)



Coverage error 🤔

Systematic vs random distortion?

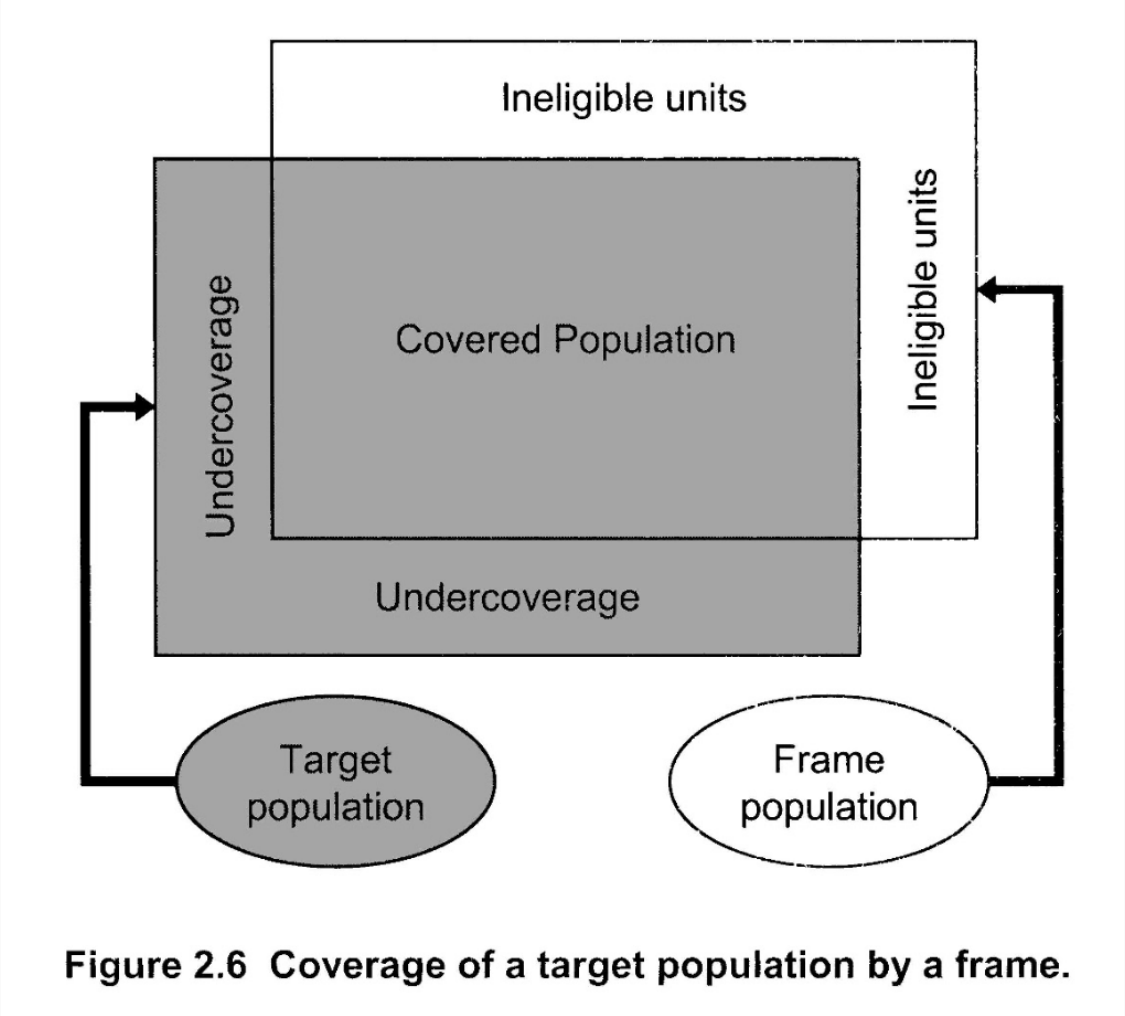


Figure 2.6 Coverage of a target population by a frame.

Sampling error 🤔

Systematic vs random distortion?

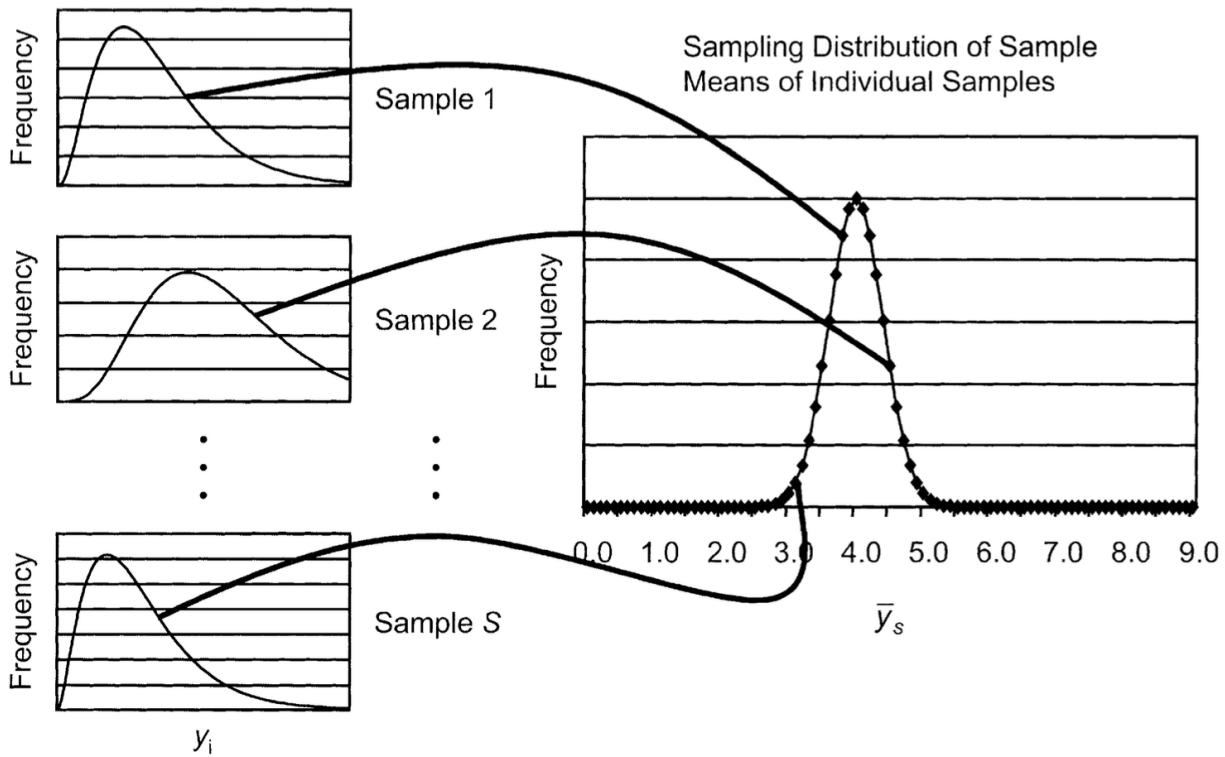


Figure 2.7 Samples and the sampling distribution of the mean.

💻 Let's move to R...

Non-response error 🤔

Systematic vs random distortion?

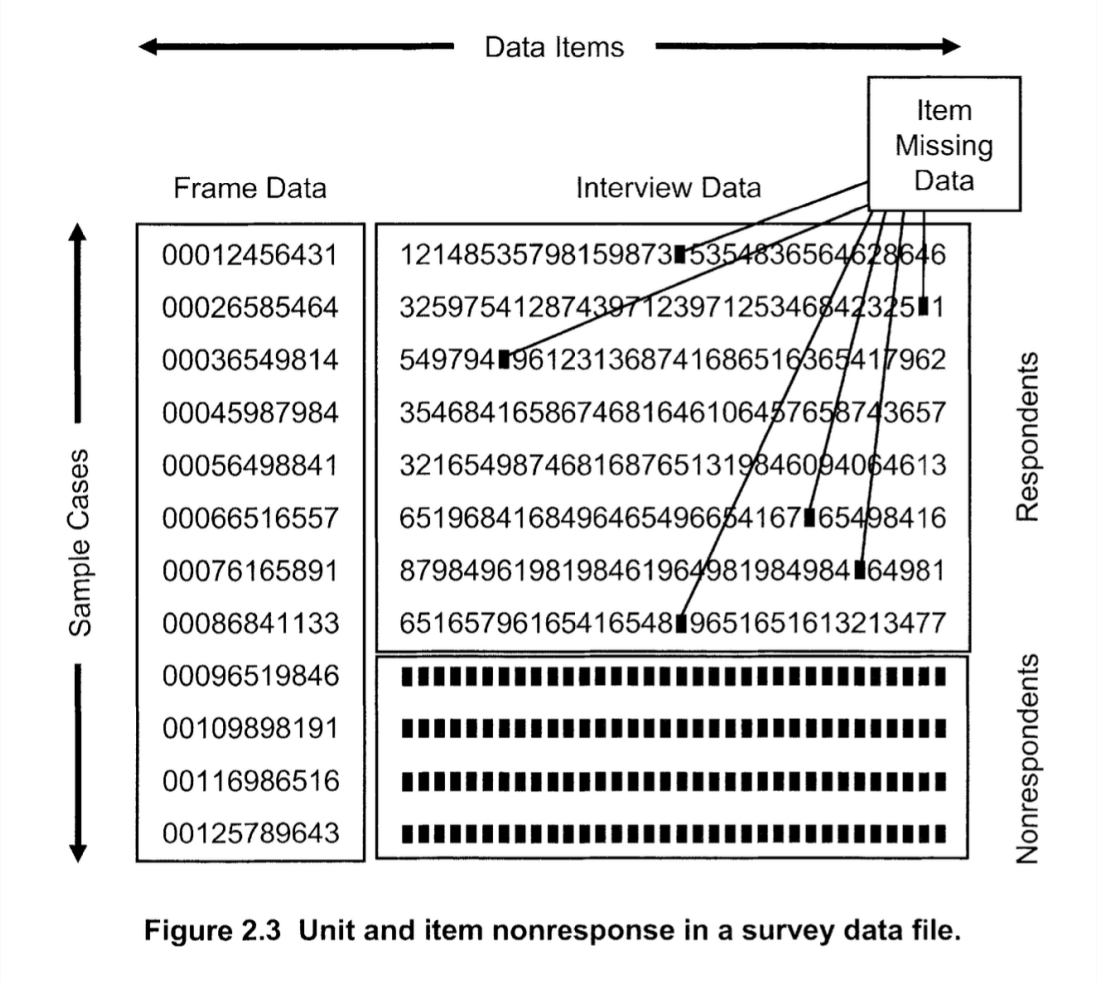
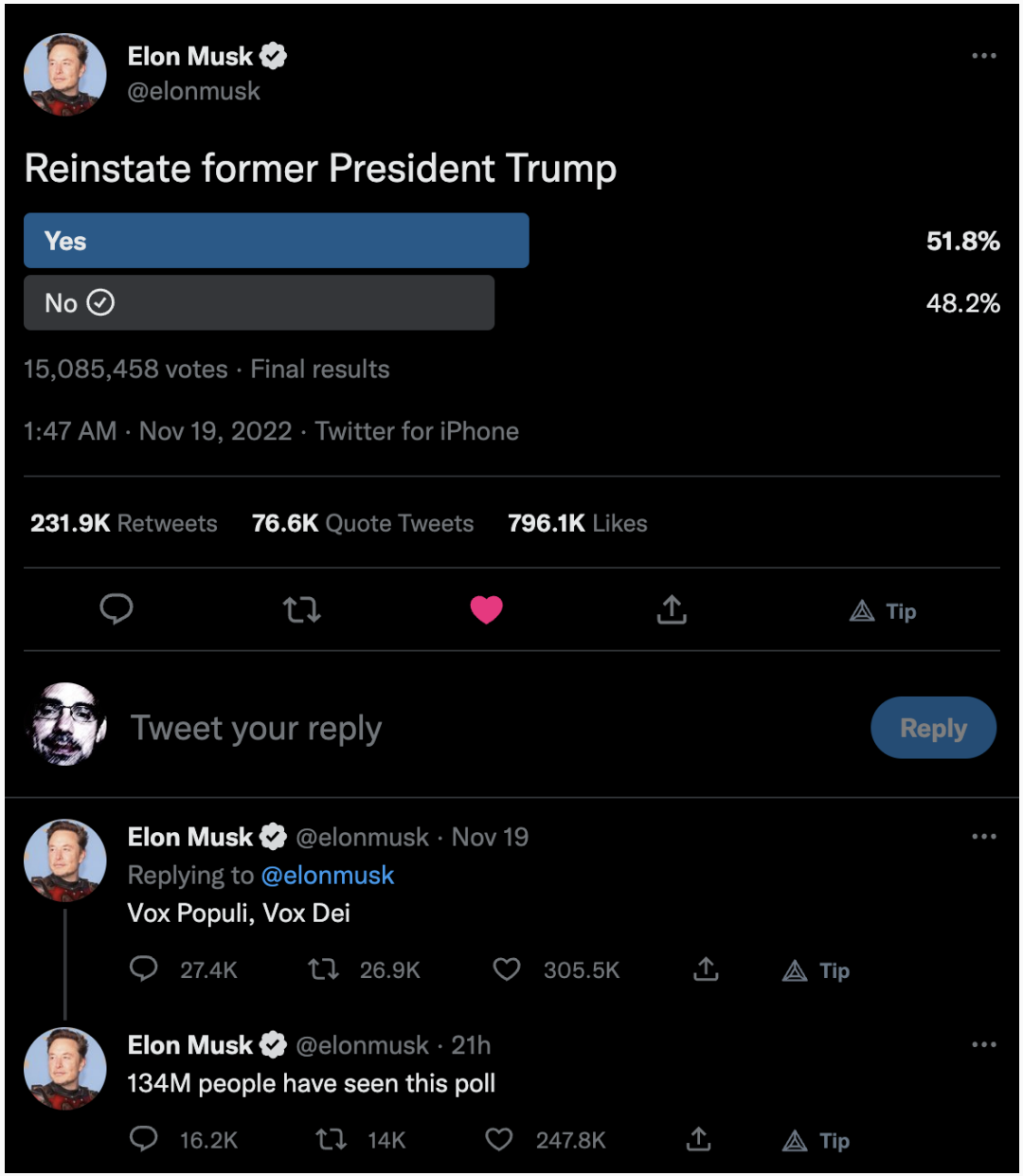


Figure 2.3 Unit and item nonresponse in a survey data file.

Short presentations

Twitter


Elon Musk
@elonmusk
Reinstate former President Trump
Yes
51.8%
No
48.2%
15,085,458 votes · Final results
1:47 AM · Nov 19, 2022 · Twitter for iPhone
231.9K Retweets
76.6K Quote Tweets
796.1K Likes
Tip
Tweet your reply
Reply
Elon Musk @elonmusk · Nov 19
Replying to @elonmusk
Vox Populi, Vox Dei
27.4K
26.9K
305.5K
Tip
Elon Musk @elonmusk · 21h
134M people have seen this poll
16.2K
14K
247.8K
Tip

Twitter

Michael Saylor ⚡ @saylor · 2h

Replying to @elonmusk

With 116.6 million followers, your polls are starting to become statistically significant. What if Twitter had an "All Users" poll that you could push to every single twitter account to find out what the entire network is thinking, with no particular adverse selection? 🤔

706 902 17.8K

Elon Musk @elonmusk · 1h

When polls are about a significant question, even those who don't follow me tend to hear about it. That said, I agree with the idea of an all-user poll. Should also be an all-user by country poll.

2,580 1,870 34.4K

Twitter Error

- Coverage error
- Query error
- Interpretation error

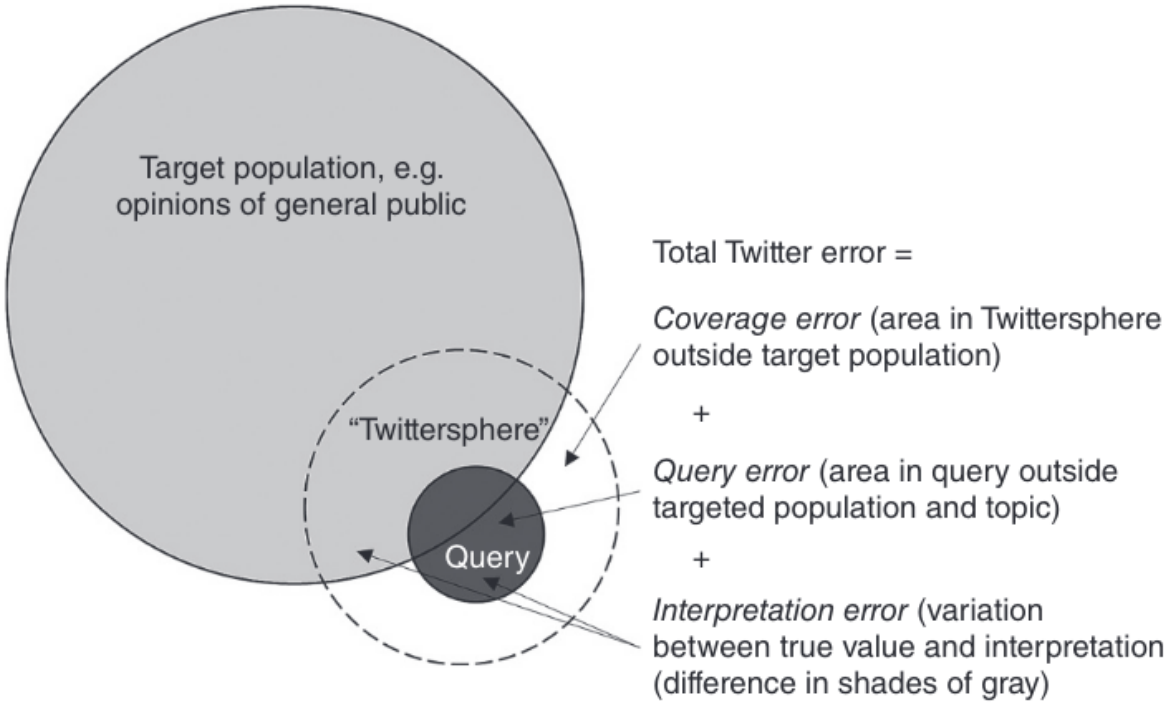


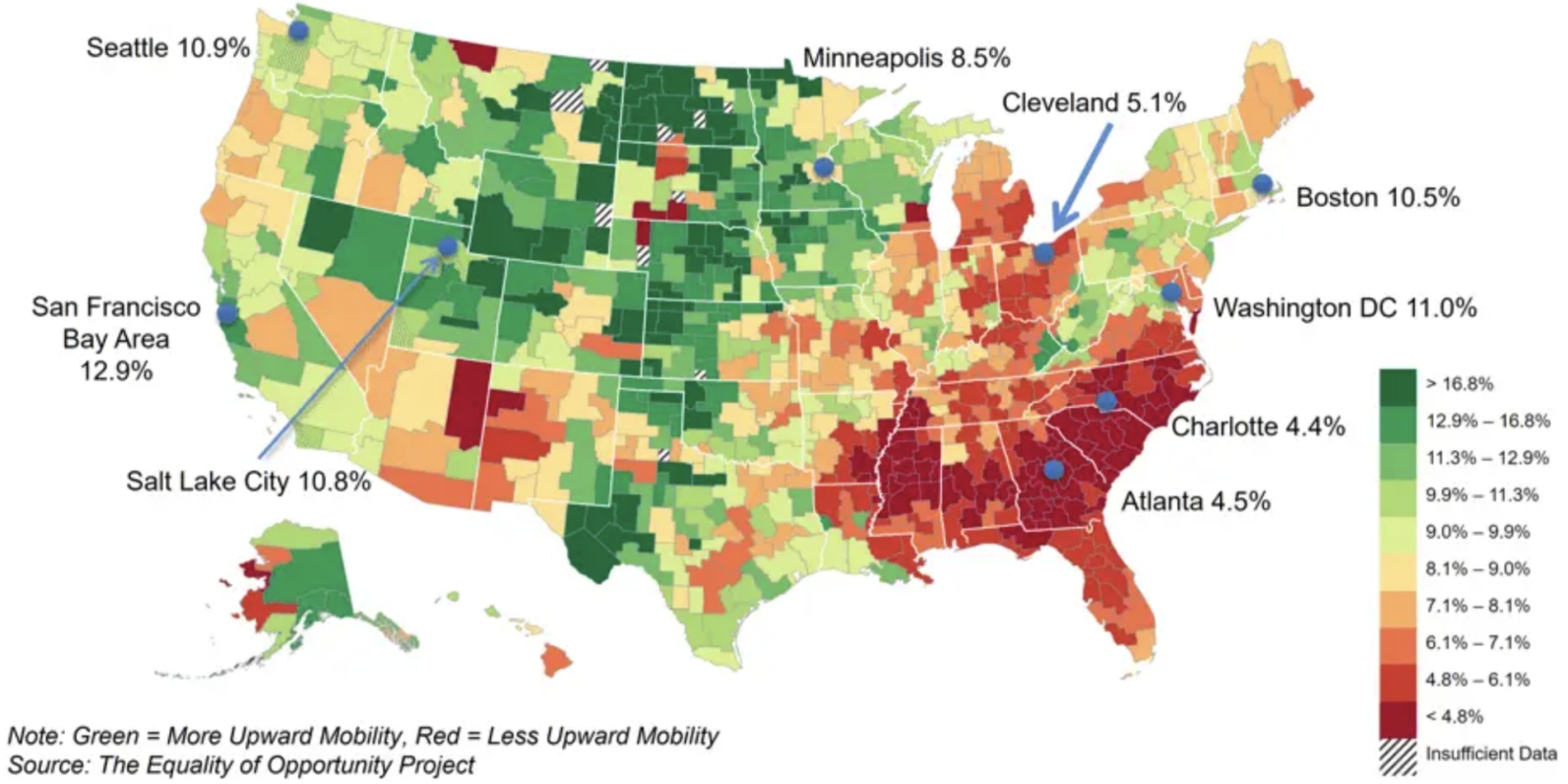
Figure 2.1 Theoretical spaces of Twitter data error.

Big data

- Volume
- Variety $\sim$ unstructured data (images, text)
- Velocity
- Veracity (tricky)
- Variability (models, meaning)
- Value
- Visualization

Big data alone is not enough!

Linkage (enhancing survey data)



Raj Chetty

Total Error Wrap-up

Total survey error

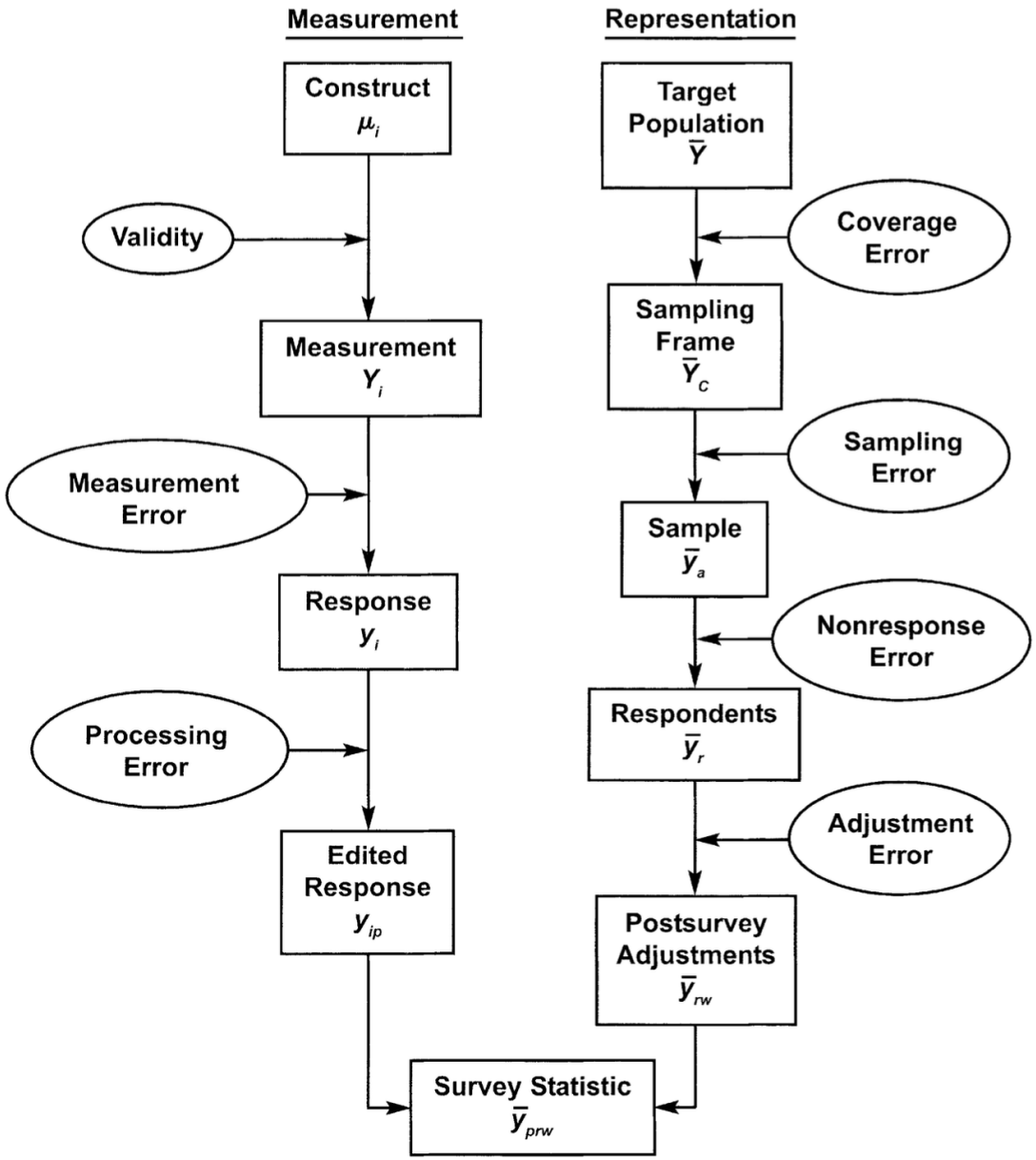


Figure 2.5 Survey life cycle from a quality perspective.

Non-sampling errors

Table 2.1: Five major sources of nonsampling error and their potential causes.

Specification error	Frame error	Nonresponse error	Measurement error	Processing error
• Data elements do not align with objectives • Invalidity • Questions lack relevance for the research purposes	• Omissions • Erroneous inclusions • Duplications • Faulty information	• Whole unit • Within unit • Item • Incomplete information	• Information system • Setting • Mode of data collection • Respondent • Interview • Instrument	• Editing • Data entry • Coding • Weighting • Tabulation

Systematic versus random errors

Table 2.2: The risk of random errors and systematic errors by major error source.

MSE component	Risk of random error	Risk of systematic error
Specification error	Low	High
Frame error	Low	High
Nonresponse error	Low	High
Measurement error	High	High
Data Processing error	High	High
Sampling error	High	Low

Which errors could be impacted? 🤔

Identify which error sources might be affected

> *Include or exclude institutionalized persons (e.g., residents of hospitals, prisons, and military group headquarters) from the sampling frame in a survey of the prevalence of physical disabilities in Spain*

Which errors could be impacted? 🤔

Identify which error sources might be affected

> *To use self-administration of a mailed questionnaire for a survey of elderly Social Security beneficiaries regarding their housing situation*

Which errors could be impacted?🤔

Identify which error sources might be affected

> *Reduce interview costs by using existing office personnel to interview a sample of patients of a health maintenance organization (HMO), and thus increase the sample size of the survey. The topic of the survey is satisfaction with the medical care they receive.*

HMO = Medical insurance group that provides health services for a fixed annual fee

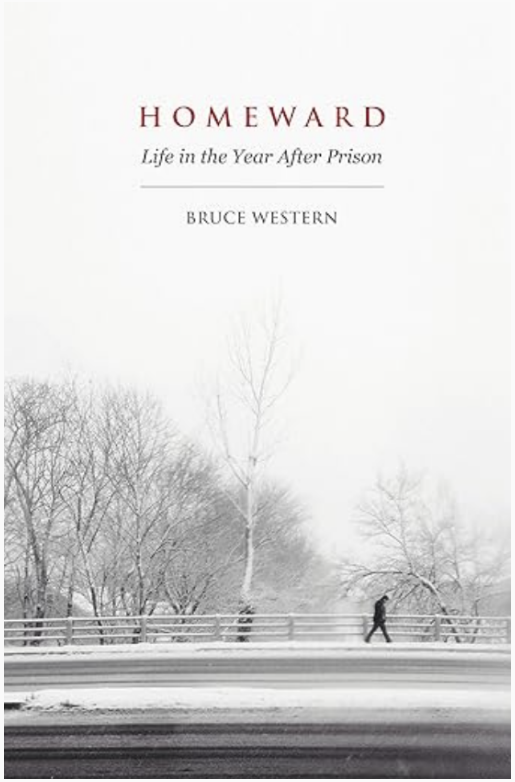
Error sources from design decisions 🤔

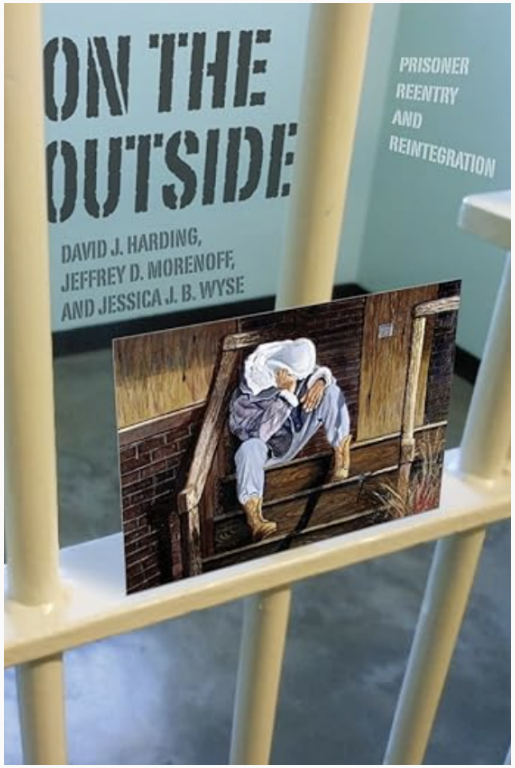
Identify which error sources might be affected

> *Extend interviewing on a survey of the use of childcare facilities by parents of young children from the originally scheduled period of January 1-May 1, to the new schedule of January 1-August 1*

Example: Chilean Women Reentry Study

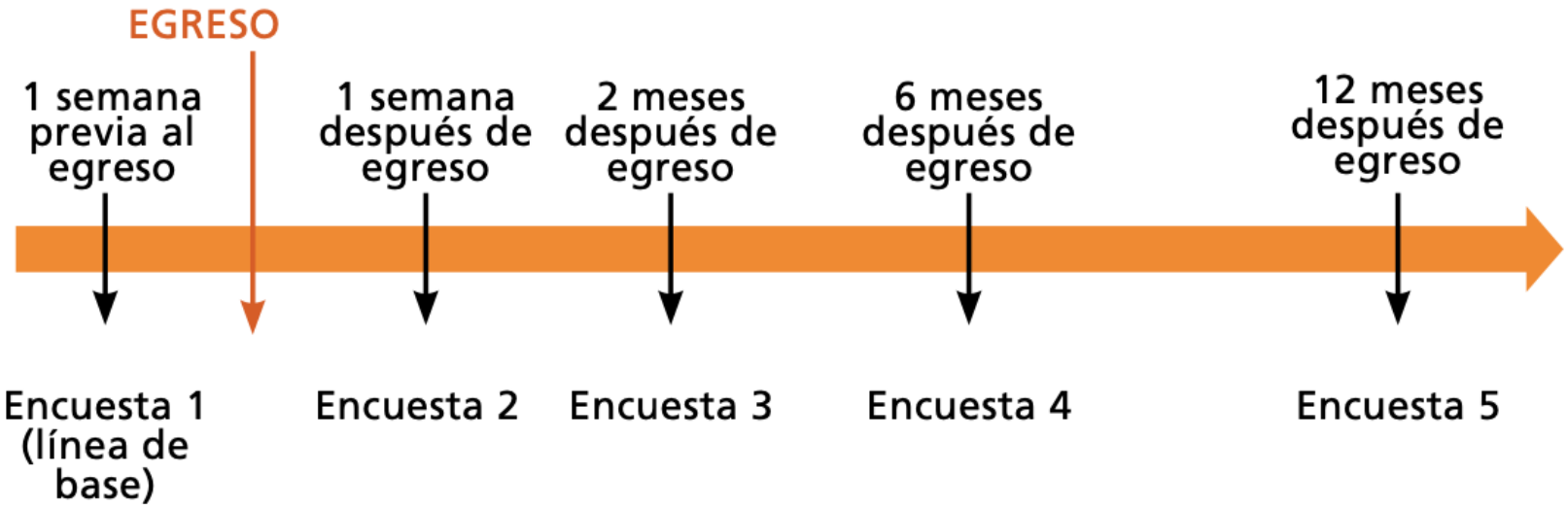
Process that individuals go through when transitioning from incarceration back into the community





Study Design

- 225 women released 2016-2017 (parole or complete sentence)
- **Chilean women** who served custodial sentences of at least 30 days
- Baseline + 4 measurements in a year since release
- Life calendars (12 months)
- Mixed methods



Study Design

- Focus on reducing attrition
- Why is this relevant?
- Other studies:
  - *Returning Home*: retention between 32% and 69%
  - *Boston Reentry Study*: retention of 91%

Tabla 3.1: Tasa de respuesta

	Línea Base	Primera Semana	Dos Meses	Seis Meses	Doce Meses
Número de entrevistas	225	181	177	197	200
Sin contacto	-	26	31	19	21
Contactada sin encuesta	-	18	17	9	4
Casos perdidos*	-	10	8	3	4
Tasa de respuesta (%)	-	80,4	78,7	87,6	88,9

**Casos perdidos definidos como aquellos que no participan en la entrevista actual y siguientes.*

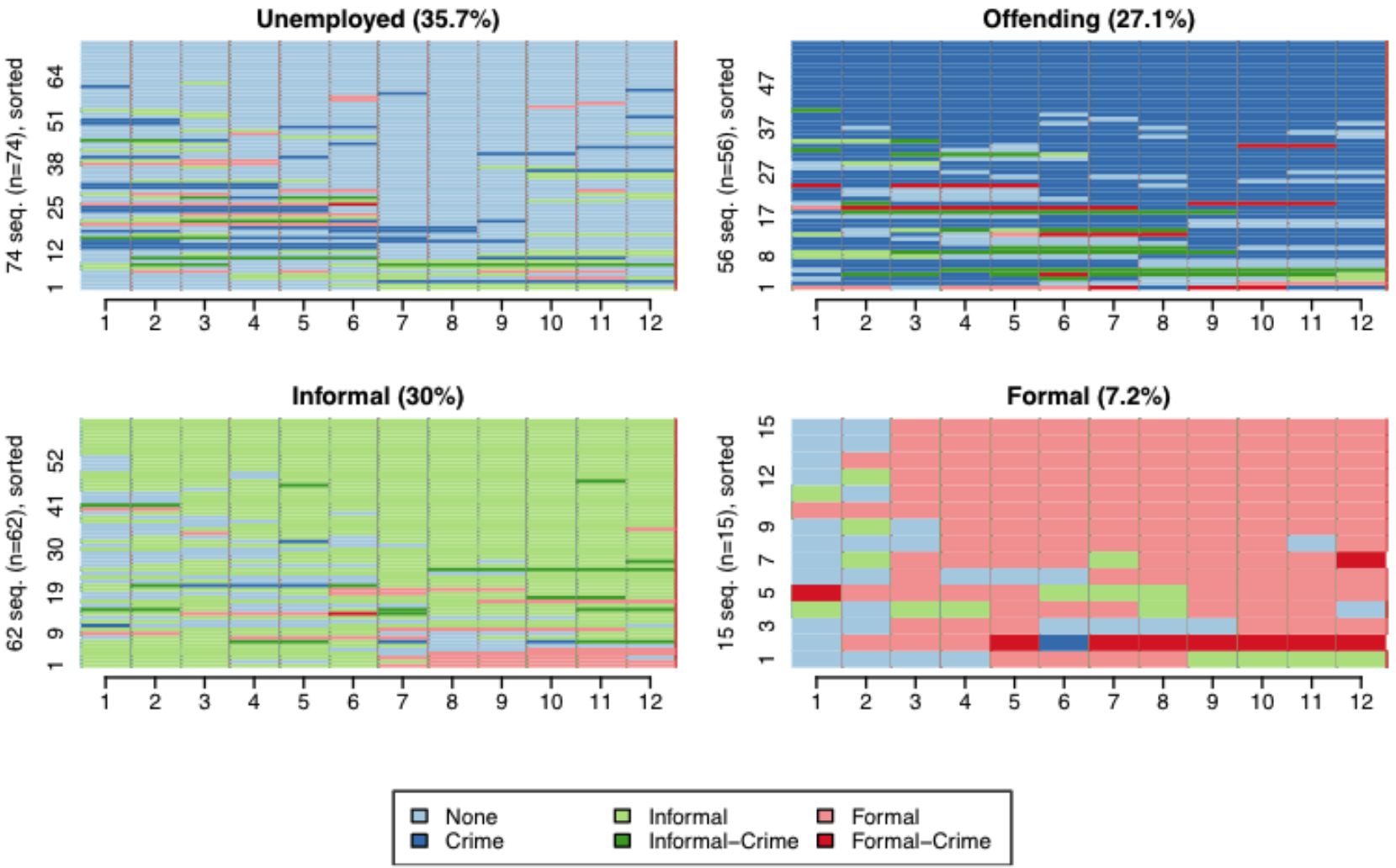
Study Design

- Face-to-face survey
  - Individual characteristics prior to prison, experience in prison, post-release events/experiences
  - Focus on (re)integration topics: housing, work, family and partner relationships, motherhood, criminal activity and drug use, mental health, program participation
  - The application lasted between **1 and 2 hours** and was conducted by trained interviewers
- Qualitative interviews
- Administrative data

Let's look at the questionnaire...

Work trajectories across women 🤓

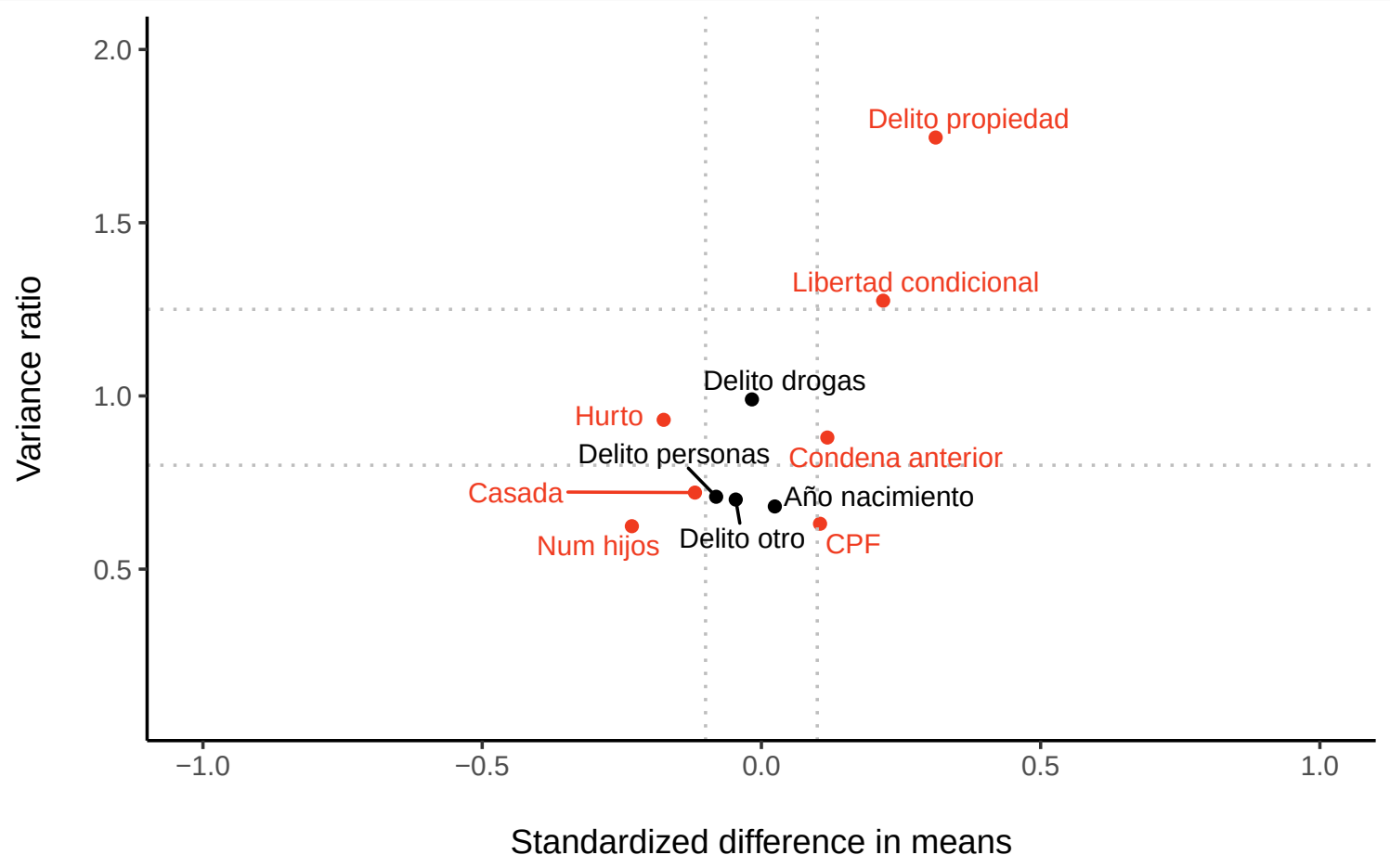
Figure S2: Sequences job-crime categories of women inmates during the first 12 months following their release by four employment-crime clusters (N = 207)



Total survey error assessment 🤔

- What areas of error do you think are the most important in this study?
- What would you have done differently?

Nonresponse error (baseline)



Using survey data and “big data” 🤔

Lookiero: Would users keep more than 5 items?

We need to simulate scenarios and cannot do AB testing (too expensive and slow)

- **Issues:** Uncertainly and not much data...
- **But wait!** Marketing conducted a survey in France, about 500 users
  - **Question:** **Would you like more than 5 items? Would you keep them?**
  - I got the answers from the marketing team

Using survey data and “big data” 🤓

A simple model to predict answers of users (survey), and simulate scenarios

- **Random Forest**
  - Predict if users will keep more than 5 items versus not
  - Features (covariates) = historical data (15)
  - Cross-validation (k-fold) + Balanced class
  - Small sample size = 449 users
- **Performance metrics**
  - **Accuracy** = 61%
  - **Precision** = 68%
  - **Recall** = 74%

Using survey data and “big data” 🤓

A simple model to predict answers of users (survey), and simulate scenarios

- **Simulation**
  - Impute the values for all french users (coarse solution)
  - Binomial model with over-dispersion
    - Probability of keeping items
  - We simulate the behavior of users based on the estimated probability

What are the limitations of this approach?

Introduction to sampling

Probabilistic sampling

We know the probability of selection of each unit of the sampling frame

- Simple random sampling (SRS)
- Stratified sampling
- Cluster sampling
- Multi-stage design

Non-probabilistic sampling

We DO NOT KNOW the probability of selection of each unit of the sampling frame

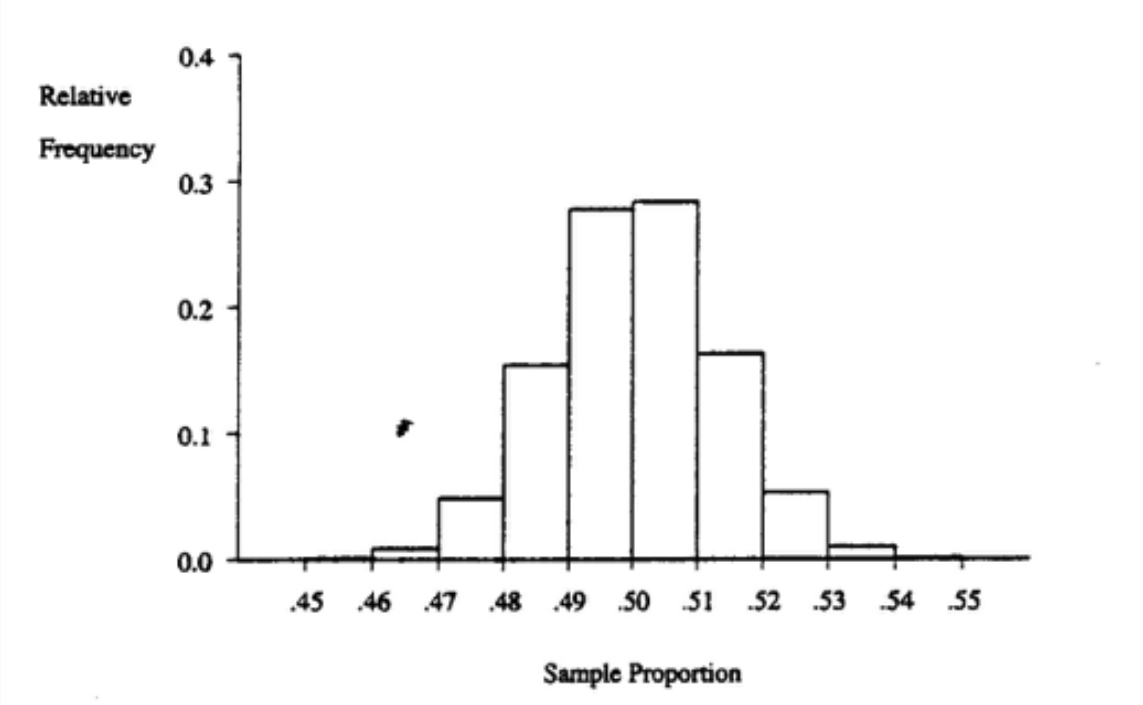
- Convenience sampling
- Quota sampling
- Judgmental or purposive sampling
- Snowball sampling

For instance, street corner interviews will be...

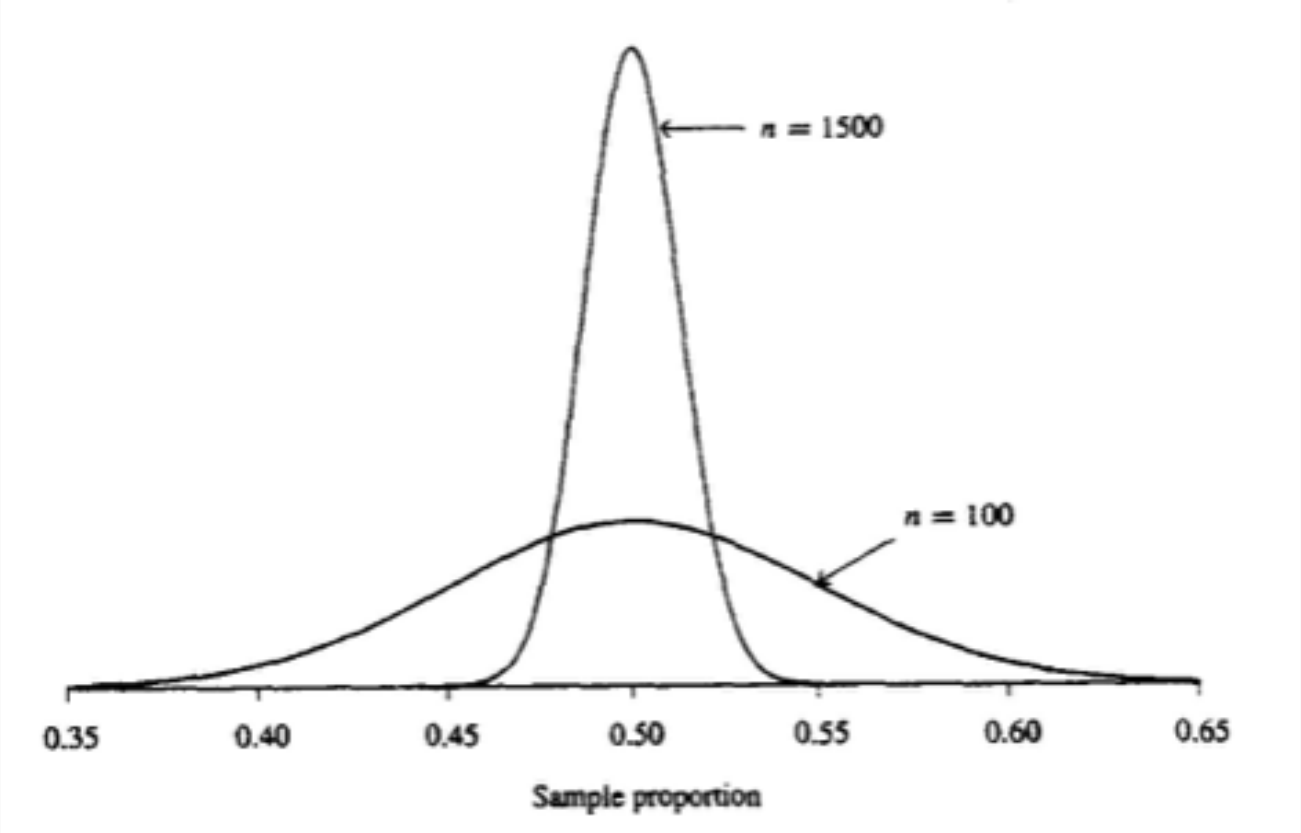
Frequentist inference refreshment

Sampling distribution

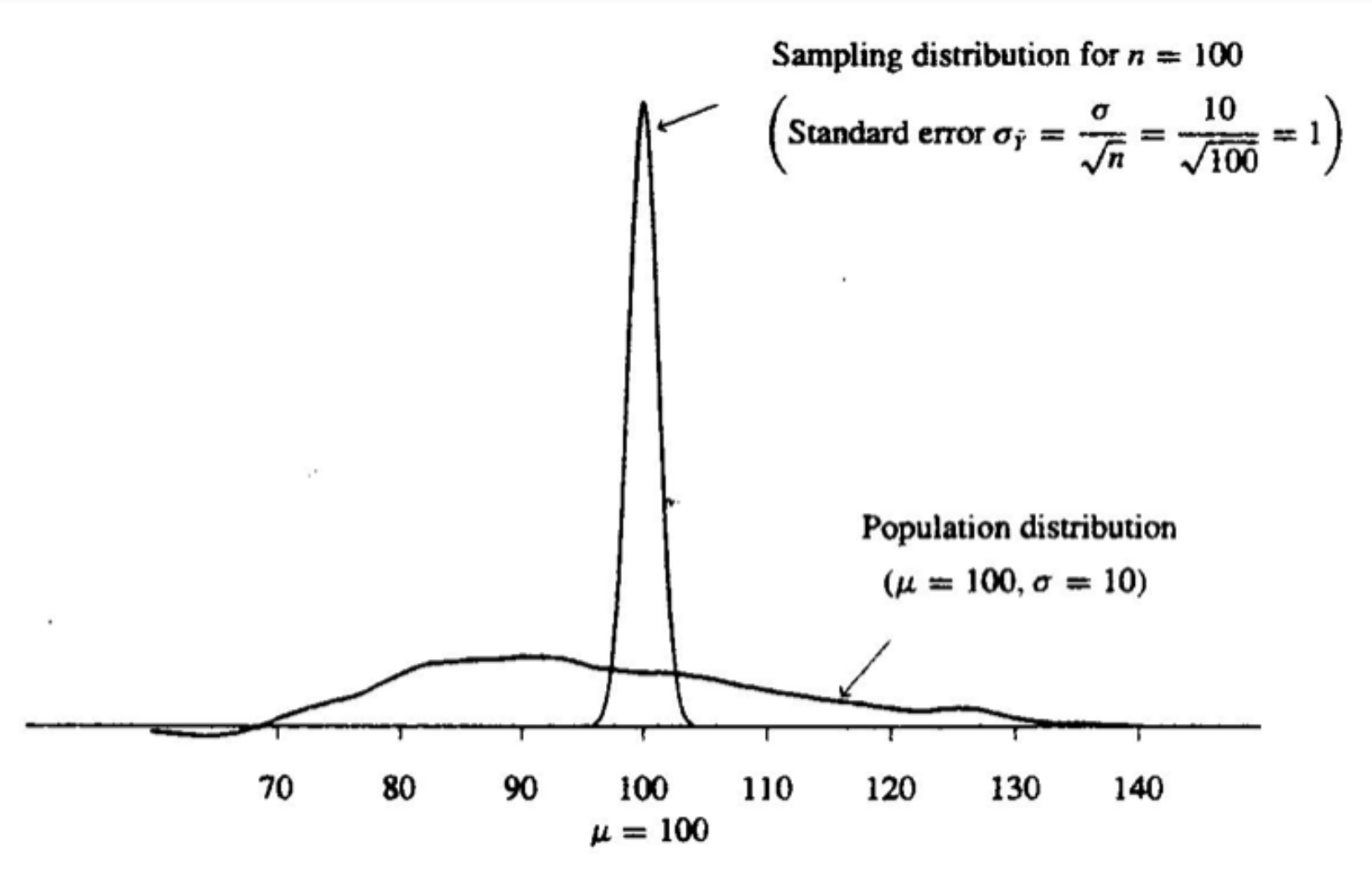
We repeat (random) sampling to get a theoretical set of possible values.
But in practice, we only get one sample.
We rest on statistical theory and the properties of repeating sampling infinitely



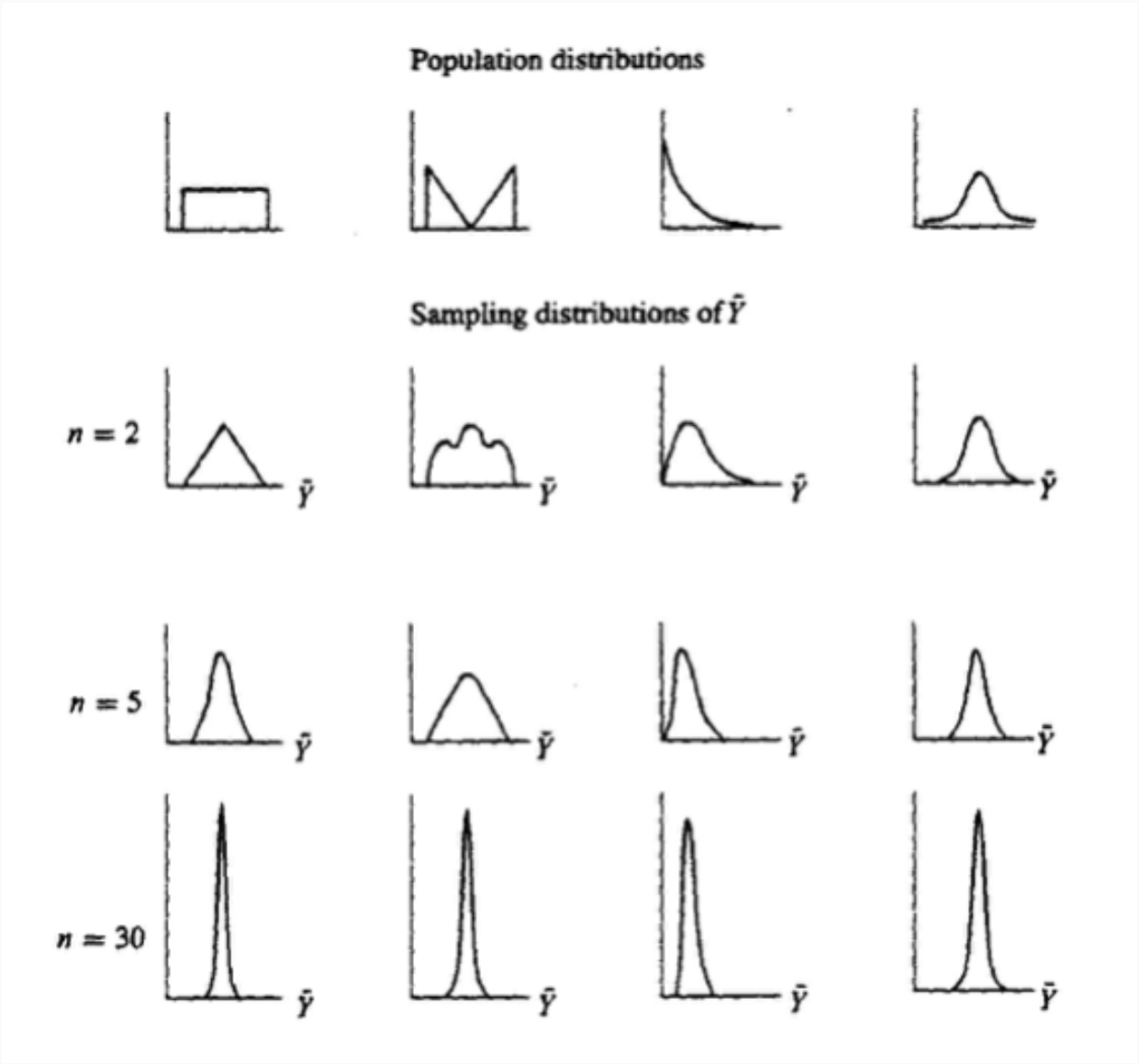
Sampling distribution



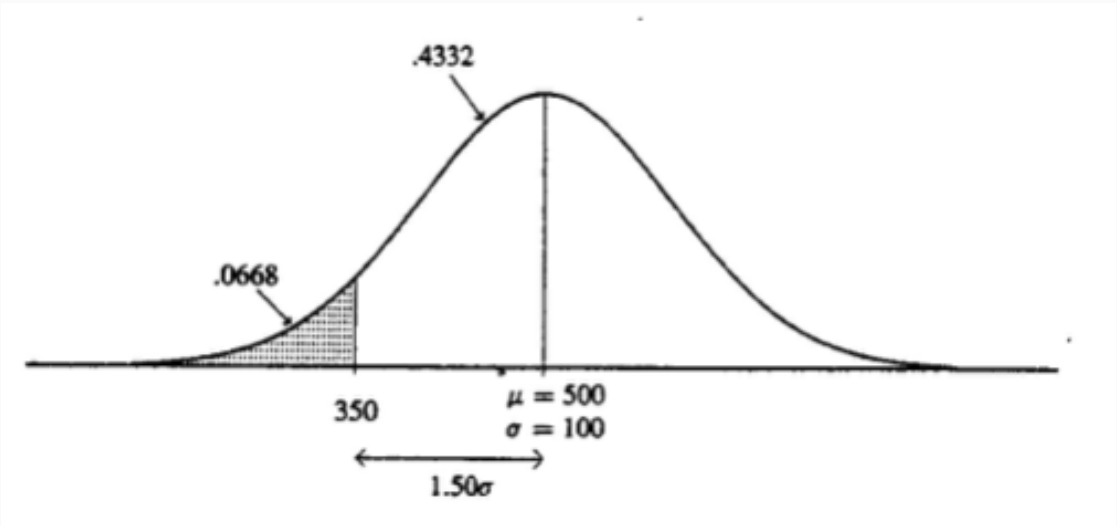
Central limit theorem



Central limit theorem



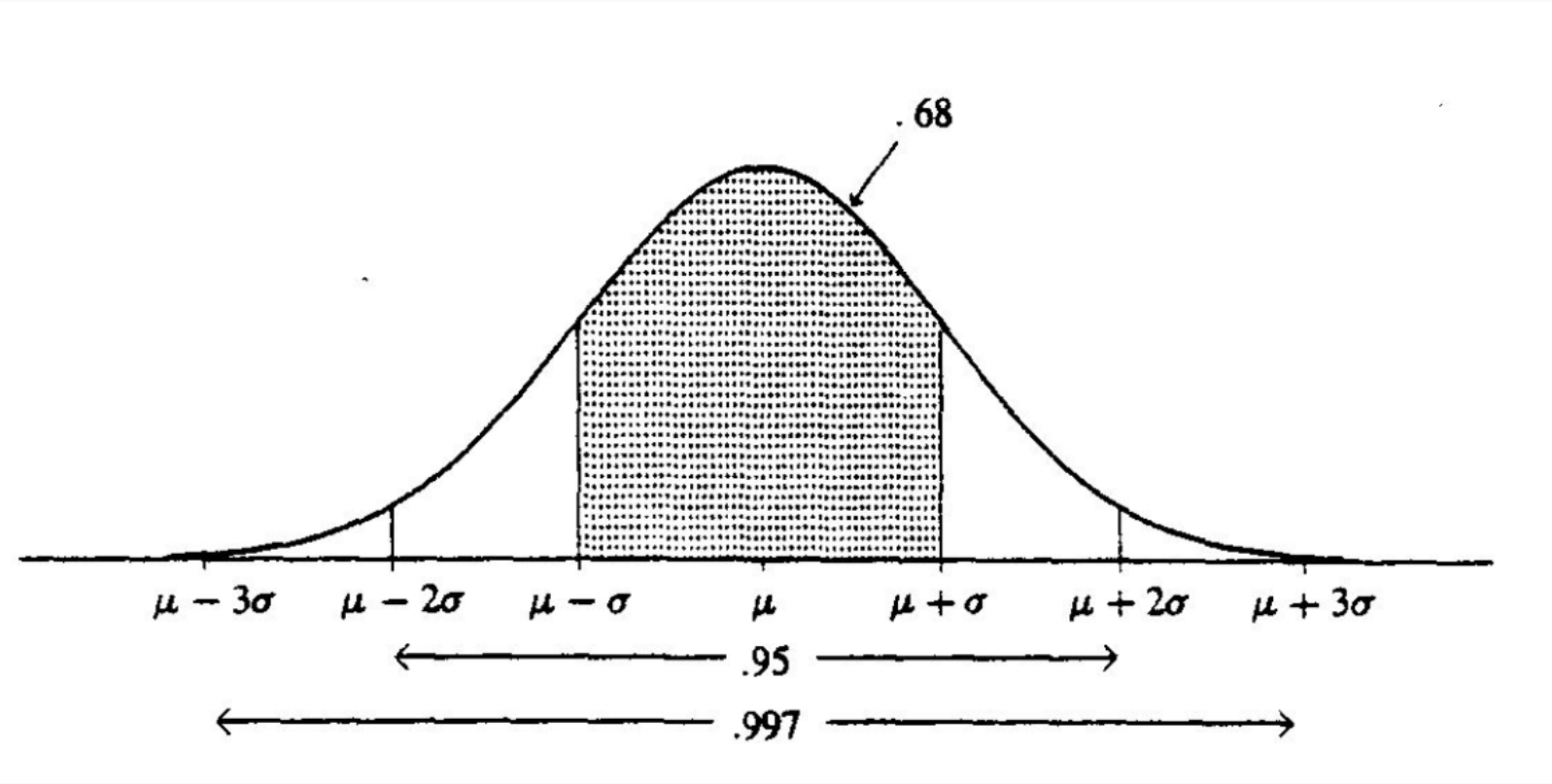
Central limit theorem



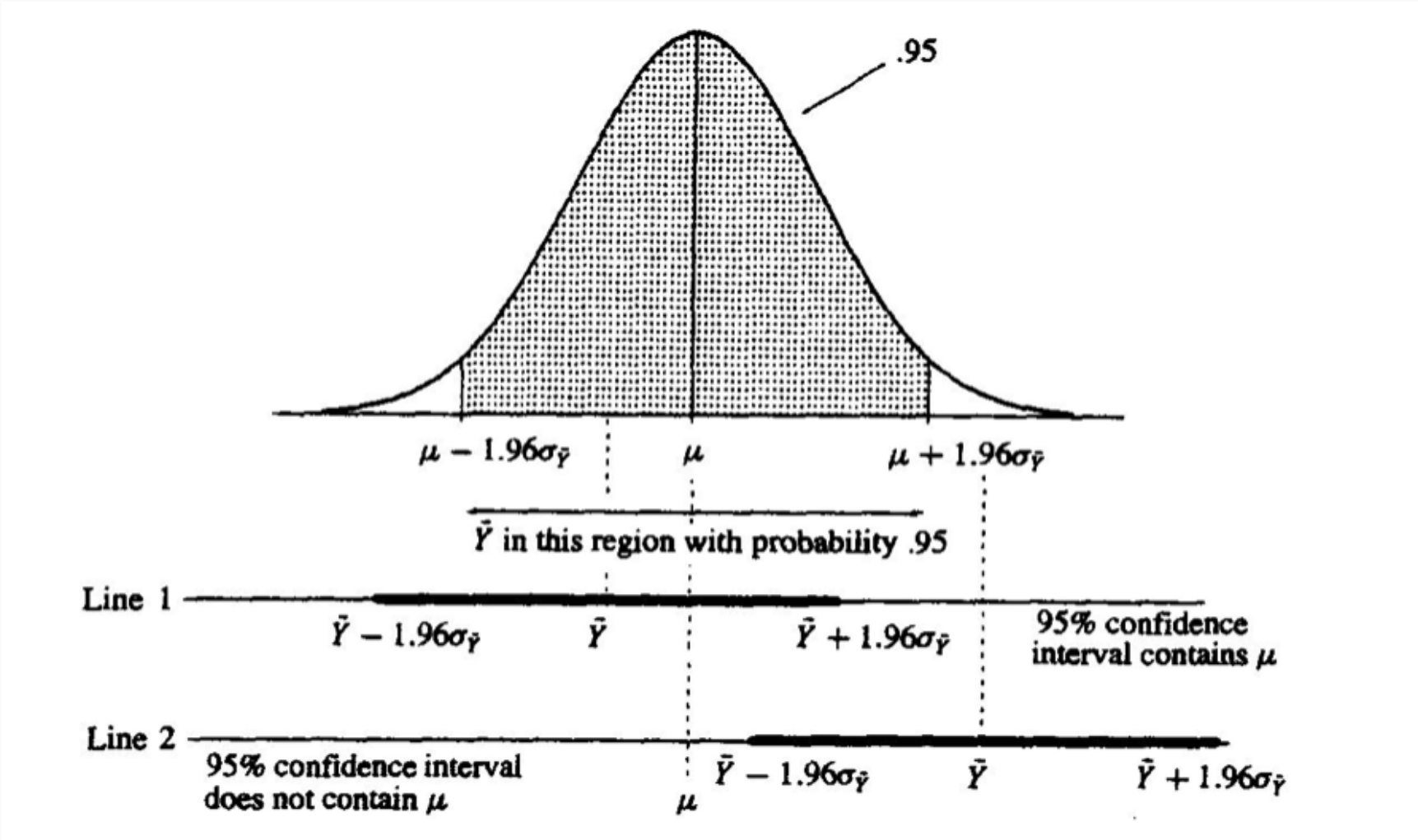

```
z = (350-500)/100
print(z)
[1] -1.5

pnorm(z, lower.tail=TRUE)
[1] 0.0668072
```

Normal distribution



Confidence interval



Margin of error (MOE)

$$MOE = z * \sqrt{\frac{\sigma^2}{n}}$$

$$MOE = z * SE$$

What is the z-value for 95% confidence?

```
(1 - 0.95)/2
[1] 0.025

qnorm(0.975)
[1] 1.959964
```

💻 Let's move to R...

https://github.com/sdaza/survey-methods

sebastian.daza@gmail.com