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

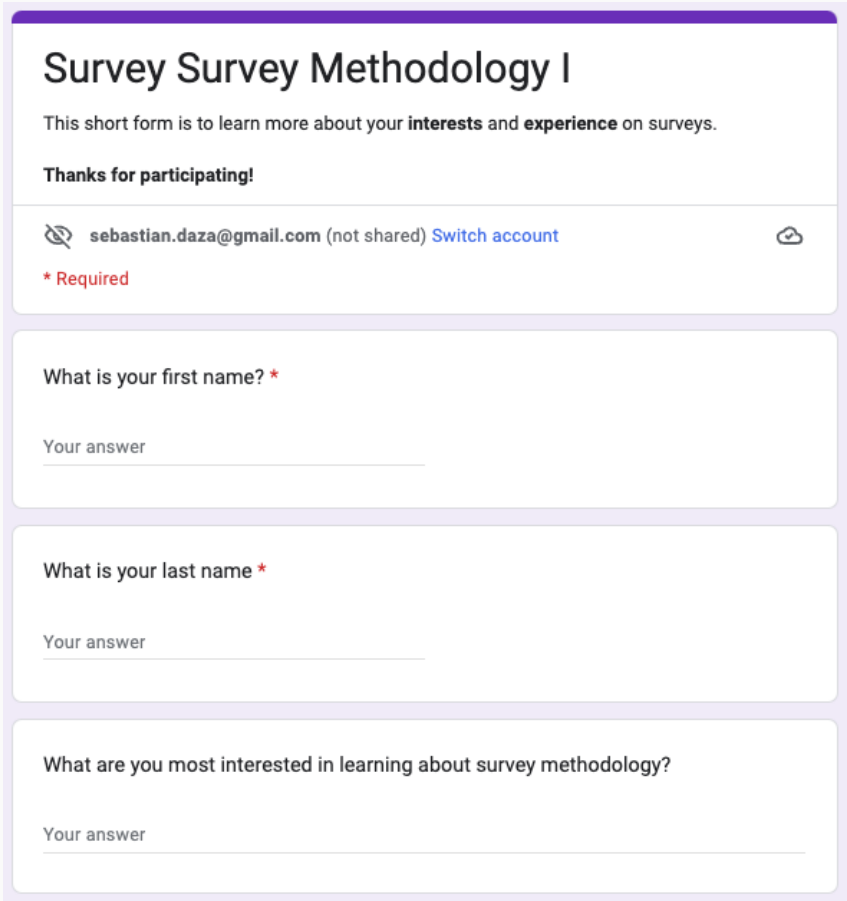
🥷 = Ninja level

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

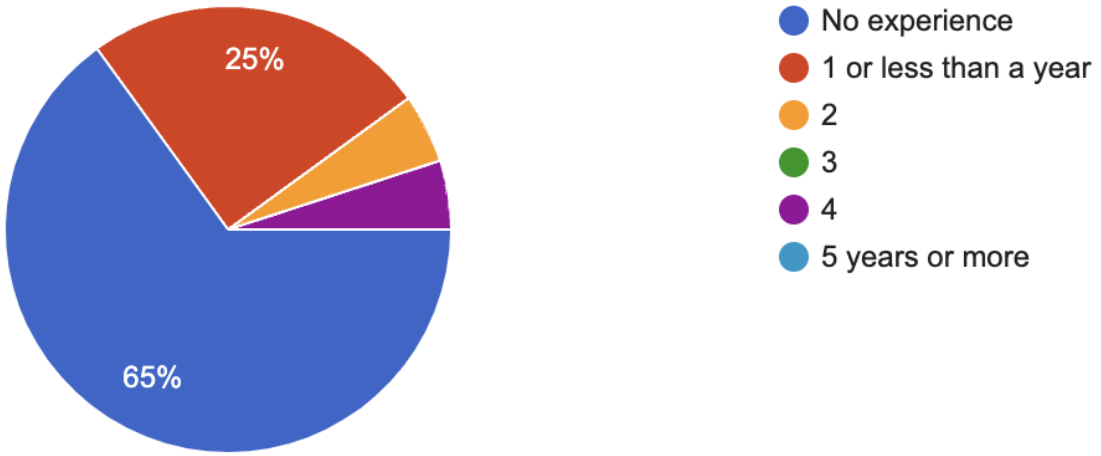
Analyzing and interpreting survey results

Sample design, validity and reliability

Experience: Design

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

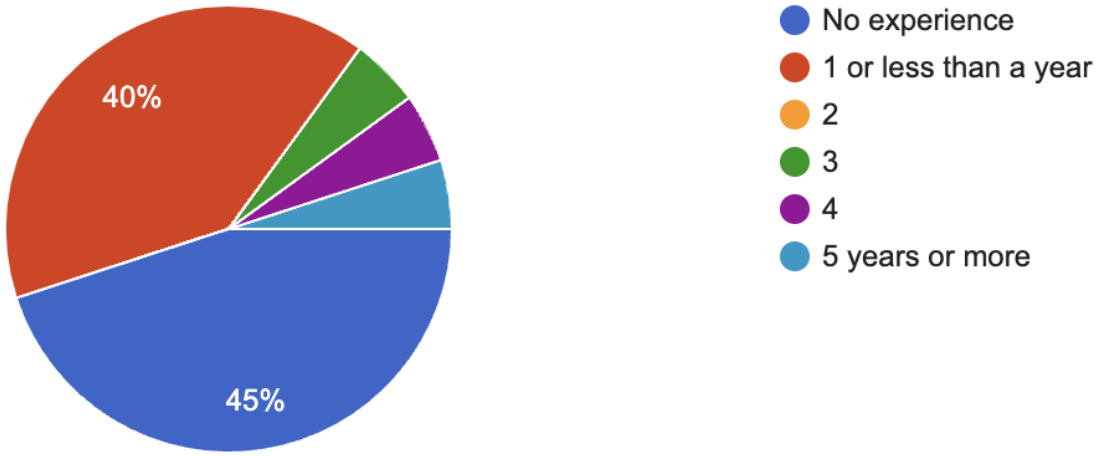
20 responses



Experience: Analysis

How much experience do you have in analyzing social surveys?

20 responses



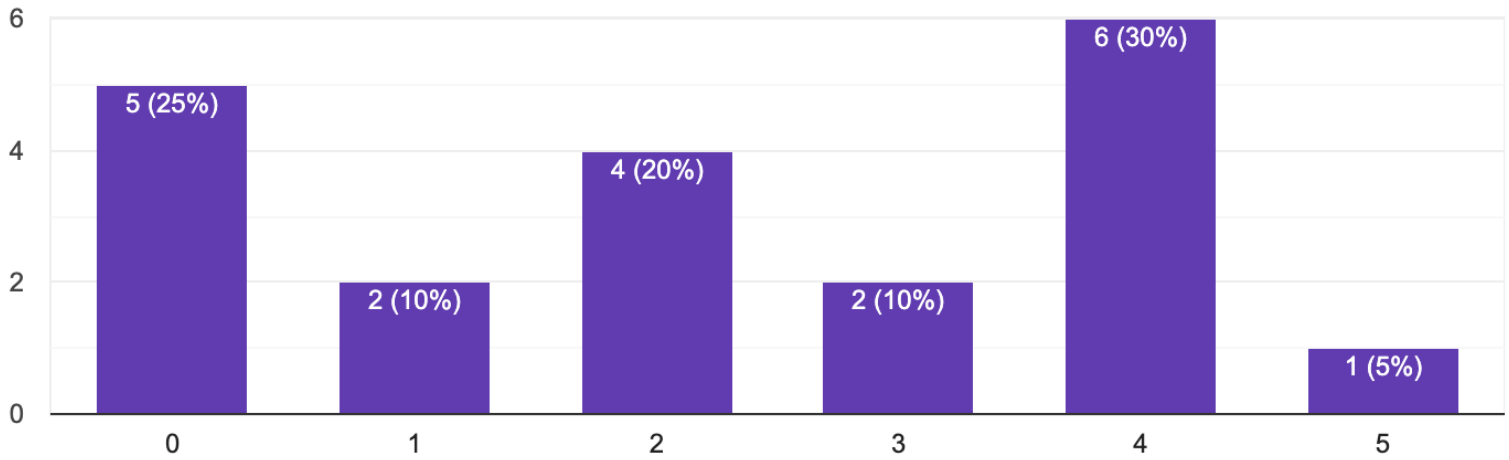
Statistical Power

Mean = 2.487

How familiar are you with the concept of statistical power?

Copy

20 responses

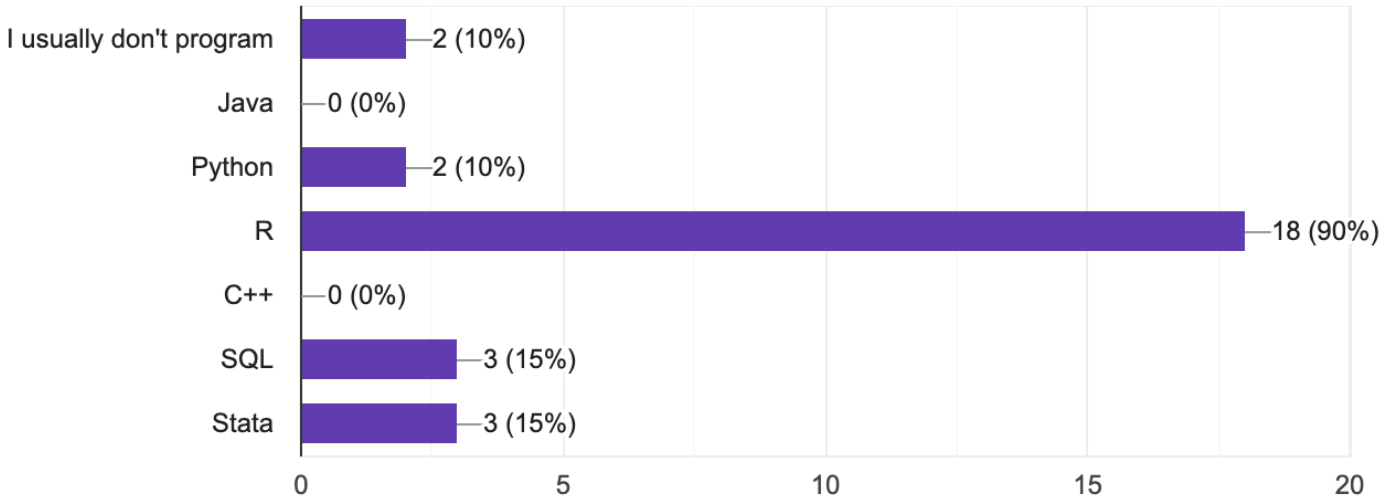


Programming

What languages do you usually use for programming?

Copy

20 responses

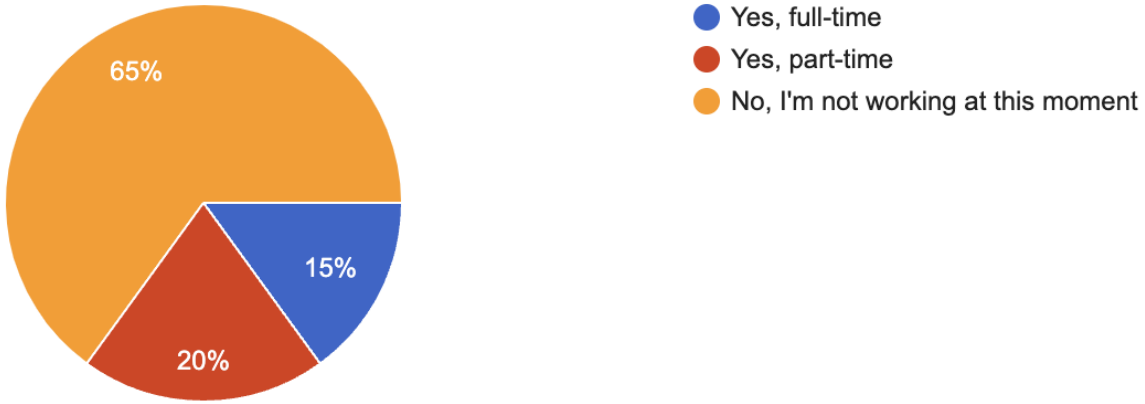


Working

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

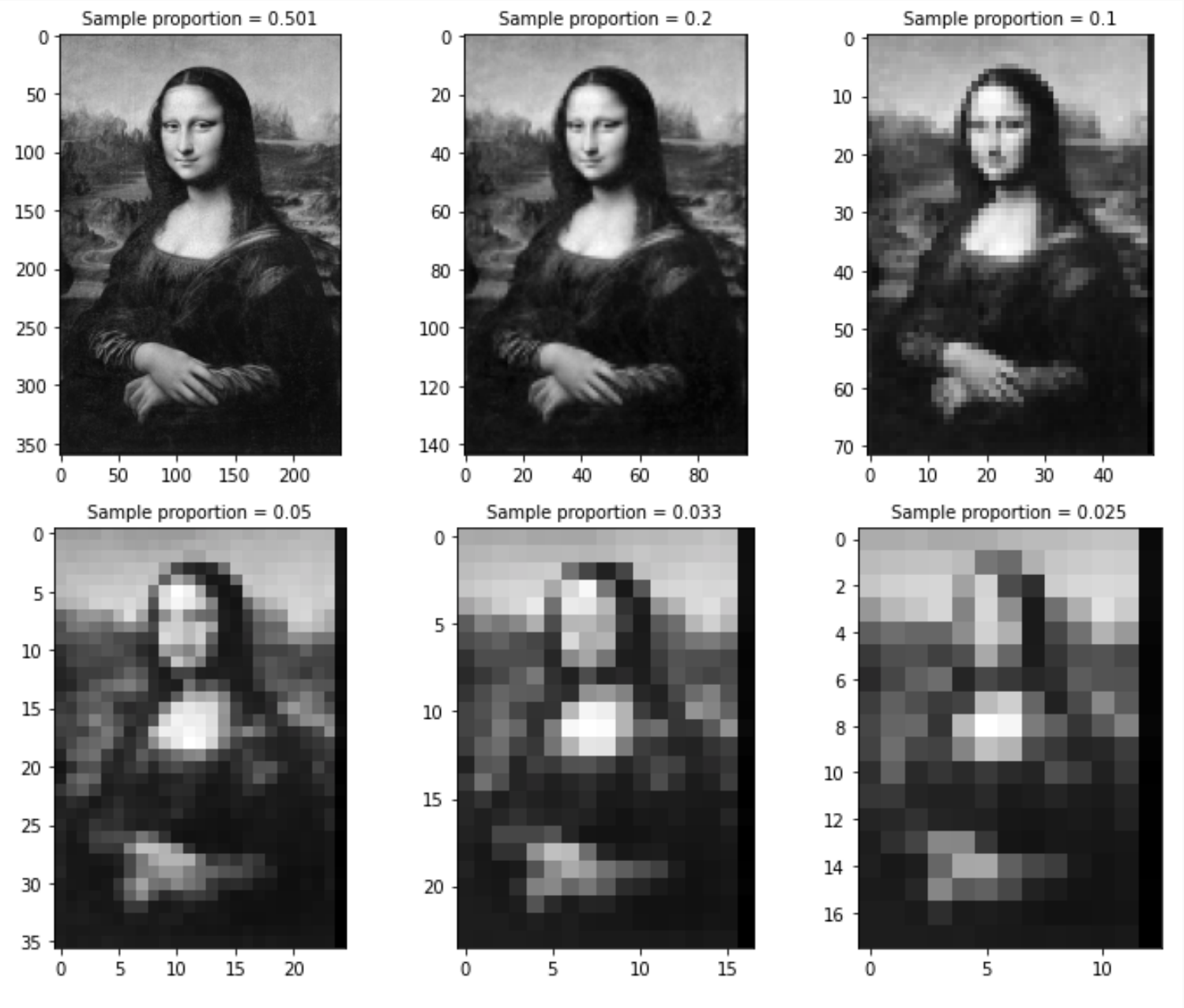
- Sampling: from samples $\rightarrow$ unbiased estimates
- Inference: statistics, margins of errors
- Measurement: the art of asking questions
- Analysis: multivariate, complex survey designs

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

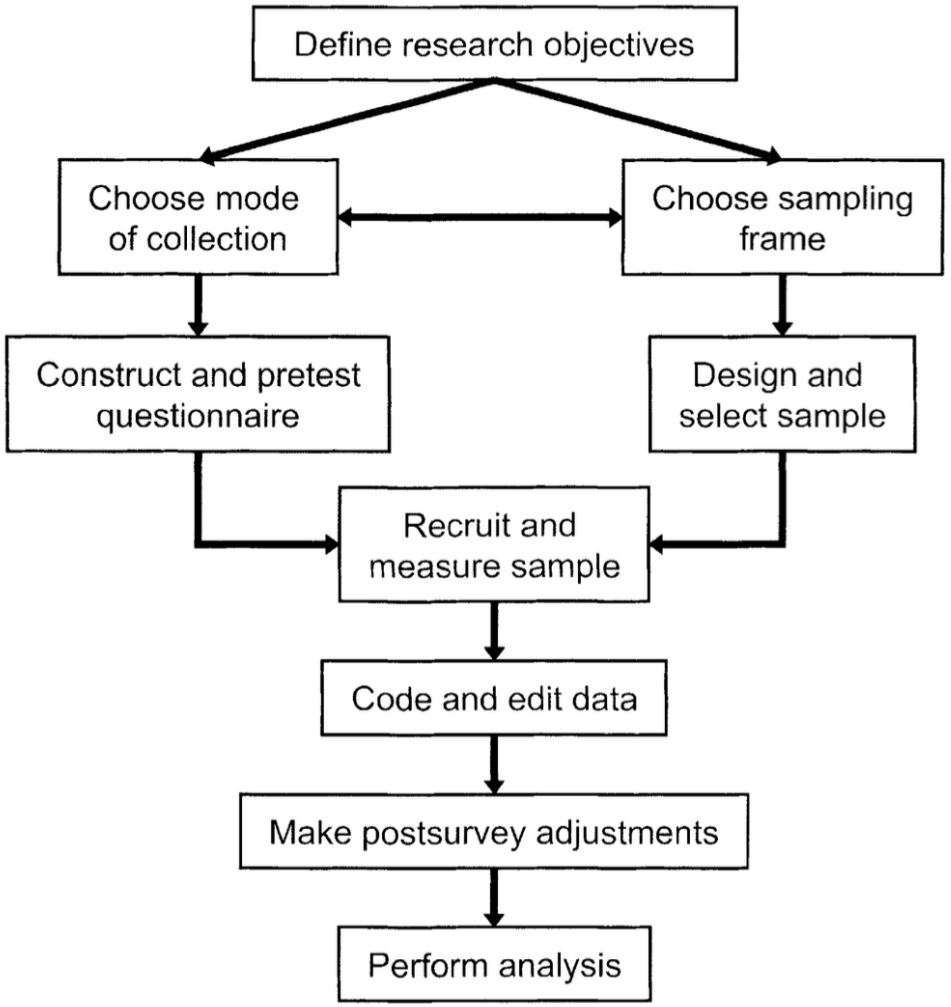
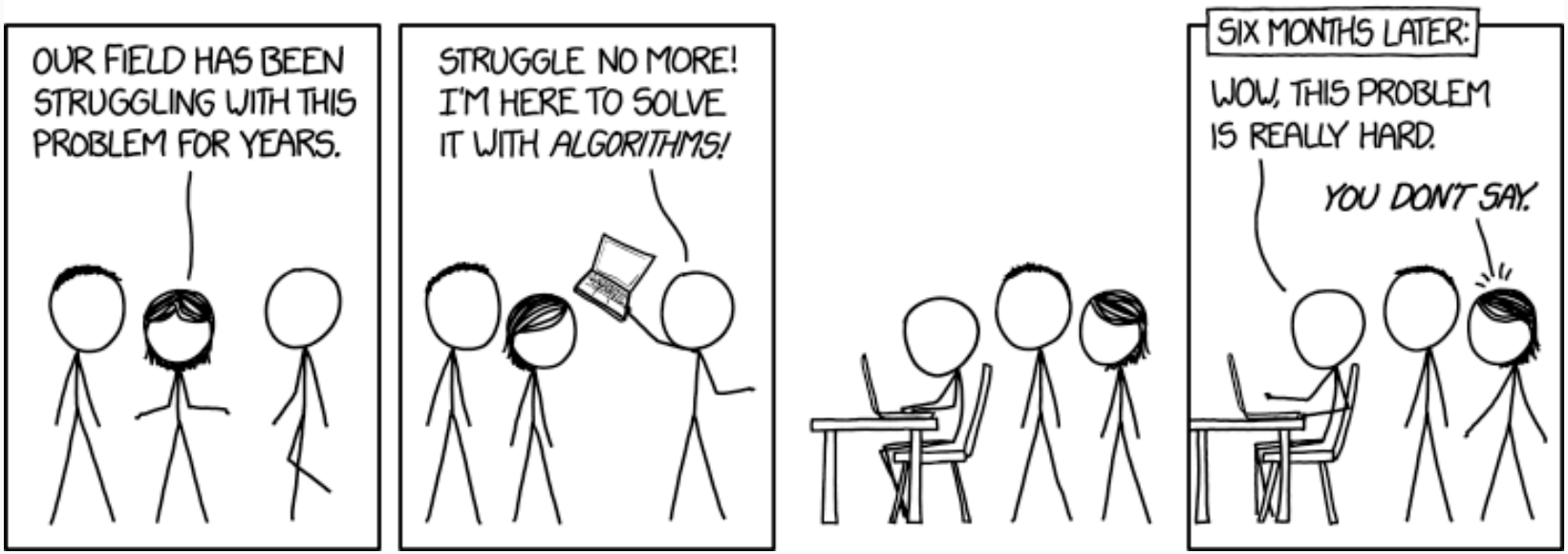


Figure 2.4 A survey from a process perspective.

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

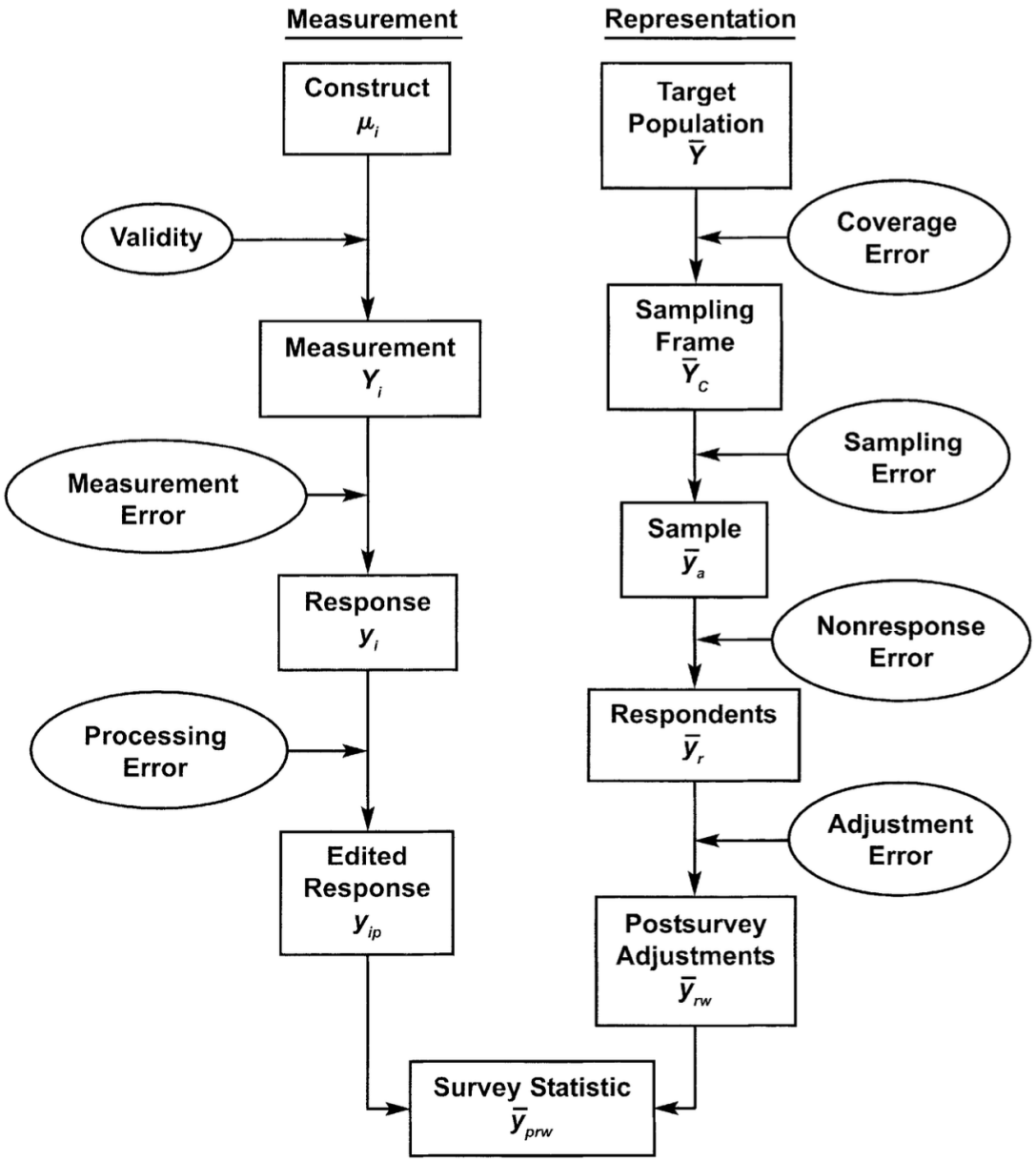
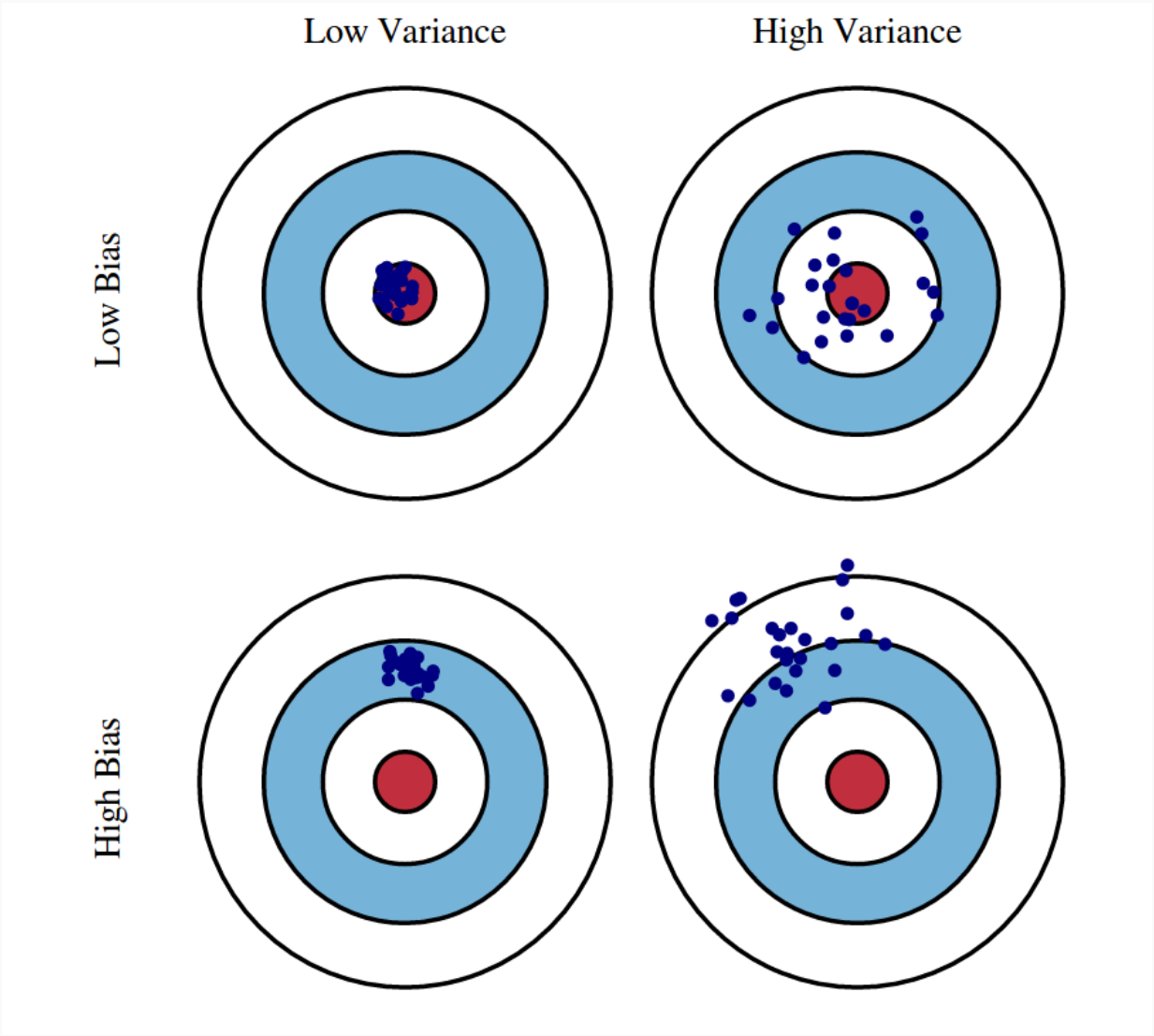
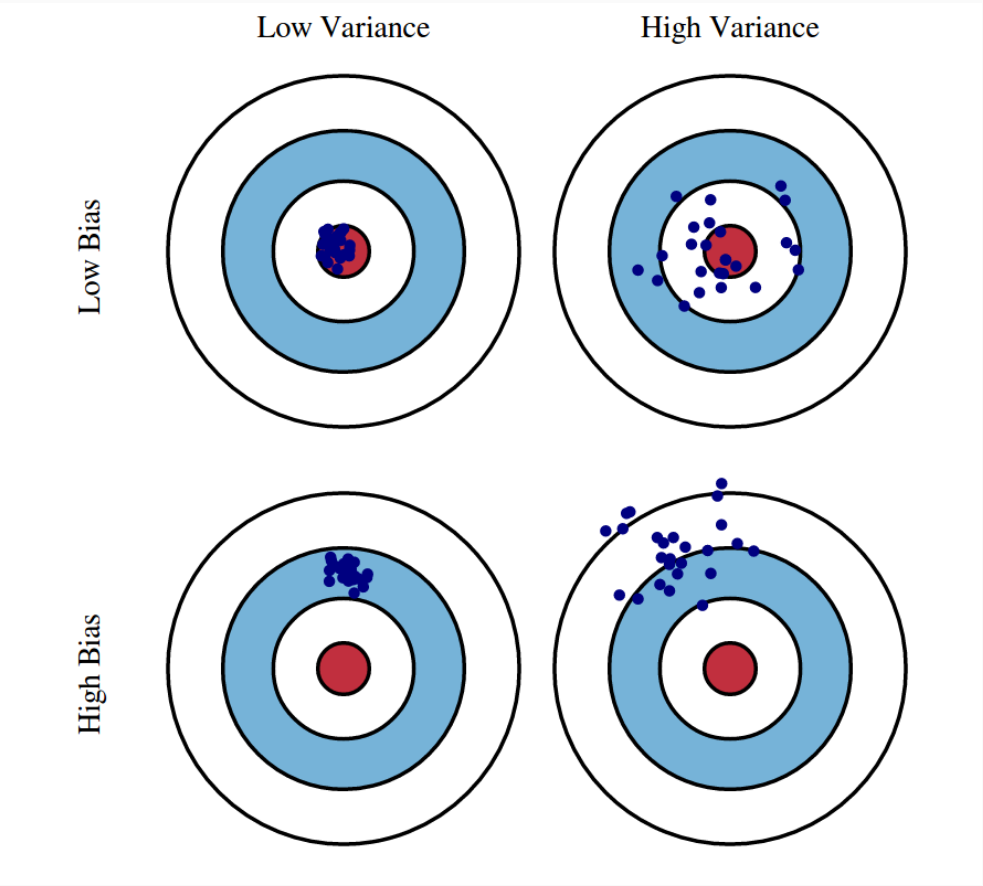


Figure 2.5 Survey life cycle from a quality perspective.

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

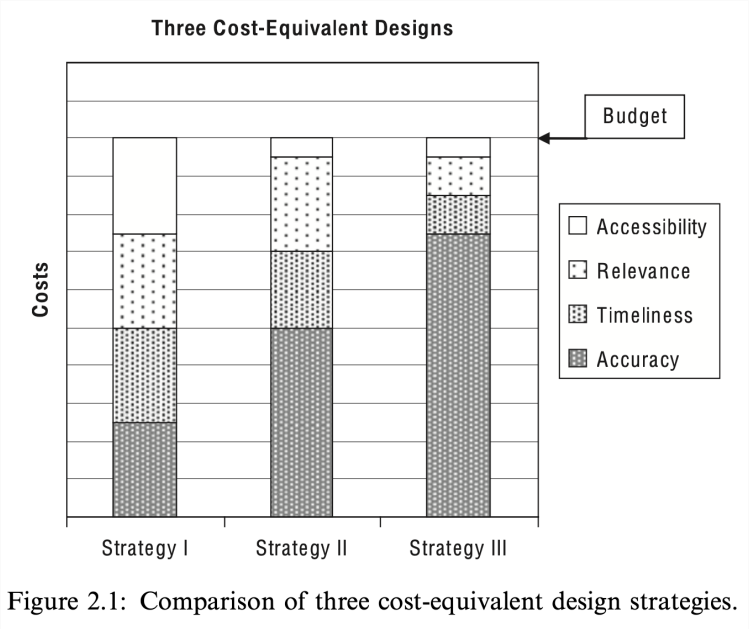


Figure 2.1: Comparison of three cost-equivalent design strategies.

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

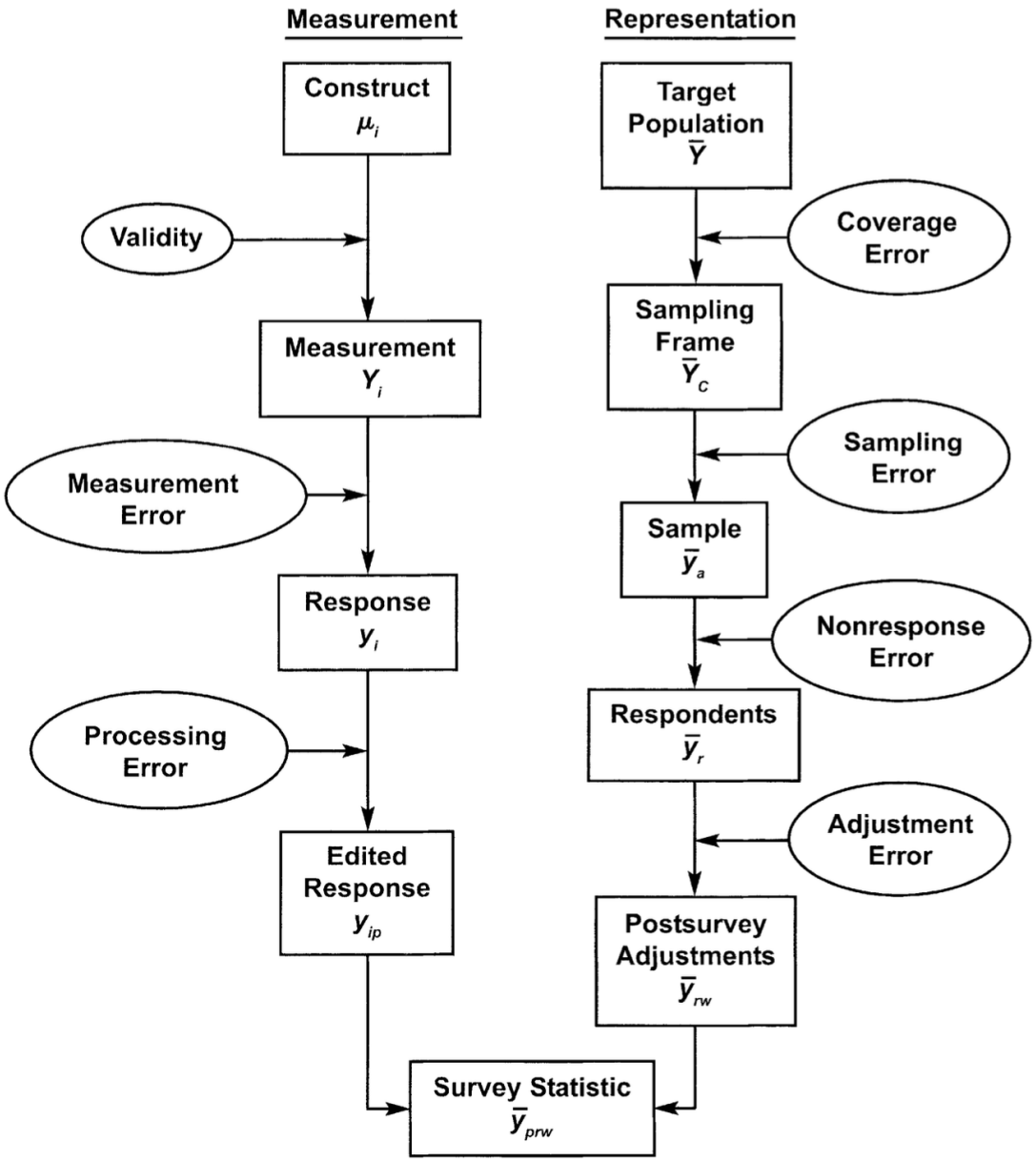
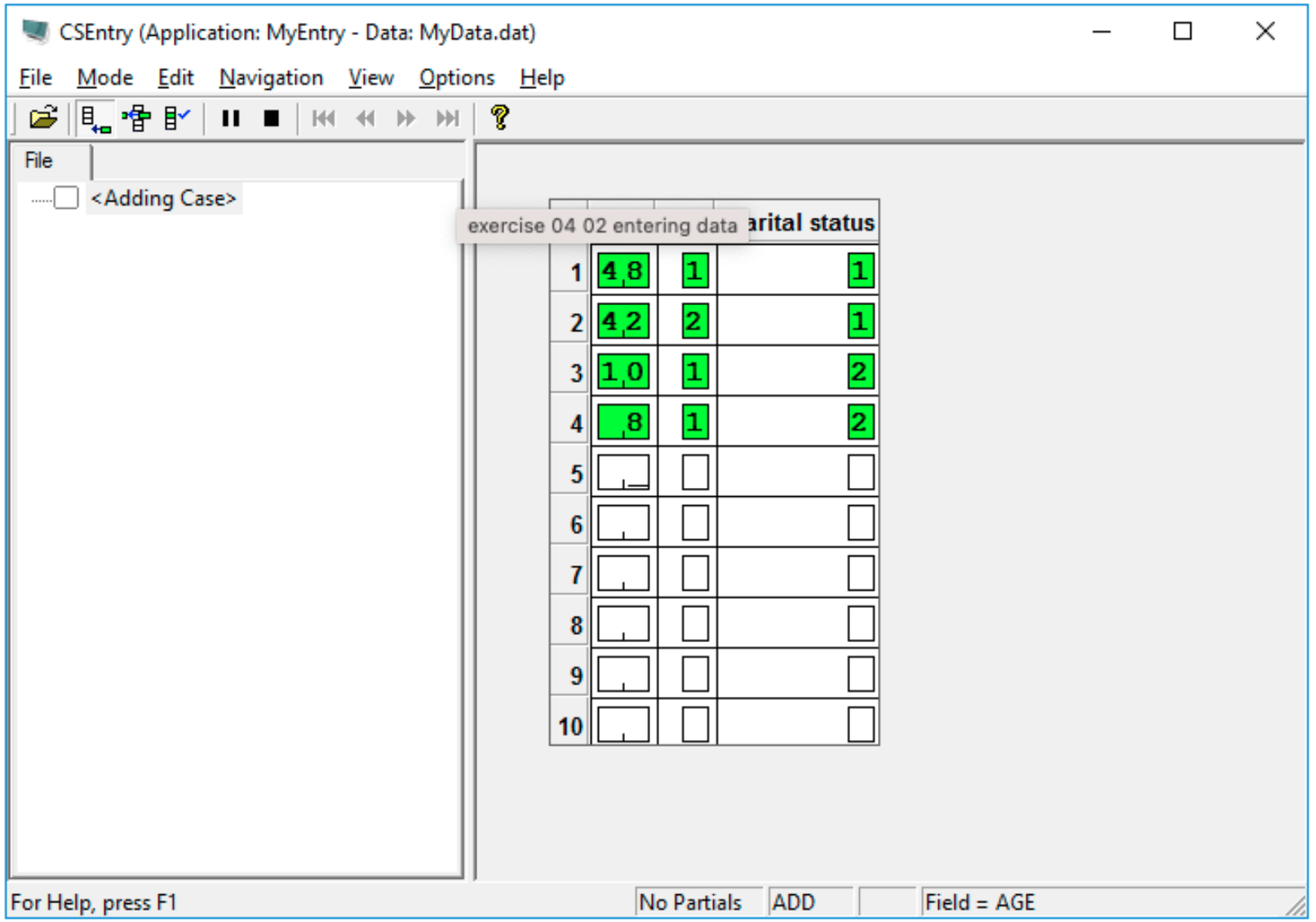


Figure 2.5 Survey life cycle from a quality perspective.

Processing error

Census and Survey Processing System (CSPro)



Coverage error 🤔

Systematic vs random distortion?

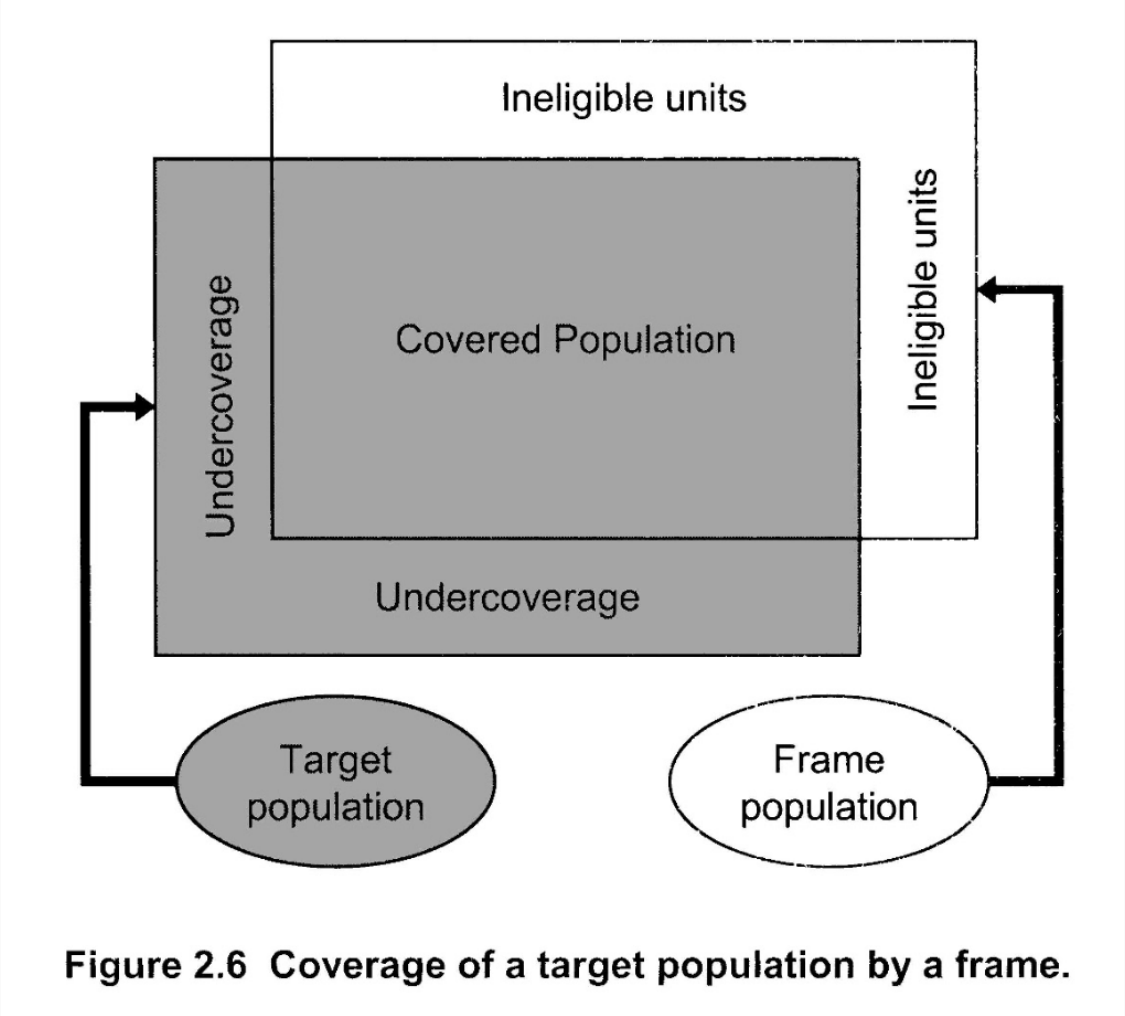


Figure 2.6 Coverage of a target population by a frame.

Sampling error 🤔

Systematic vs random distortion?

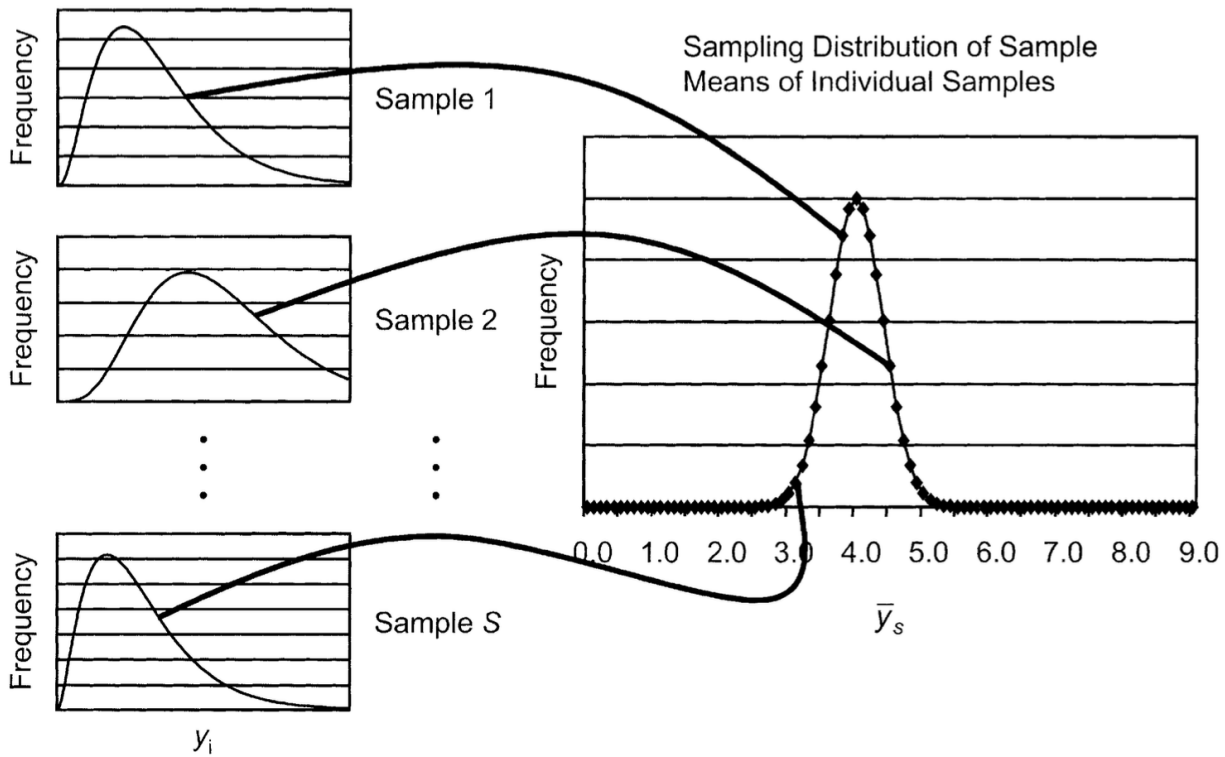


Figure 2.7 Samples and the sampling distribution of the mean.

💻 Let's move to R...

Non-response error 🤔

Systematic vs random distortion?

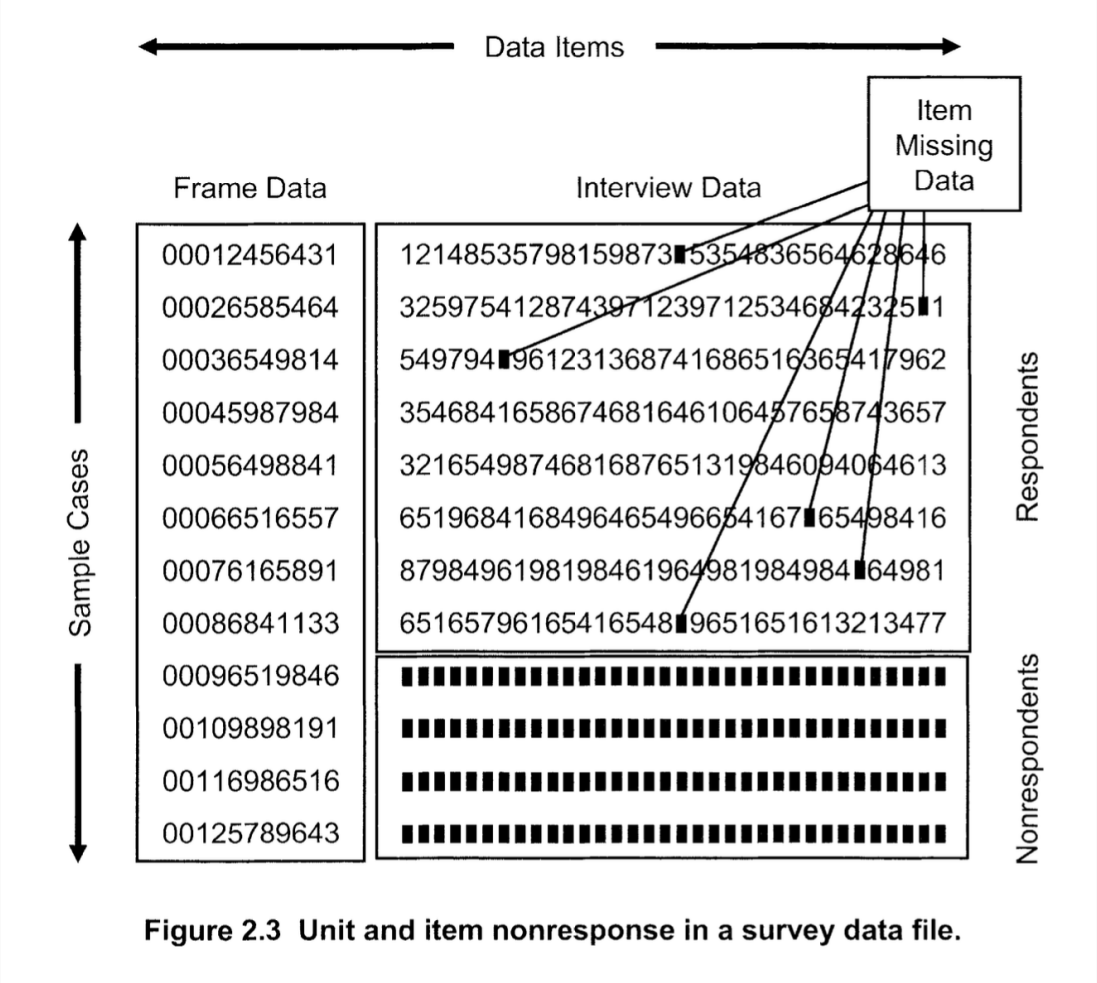
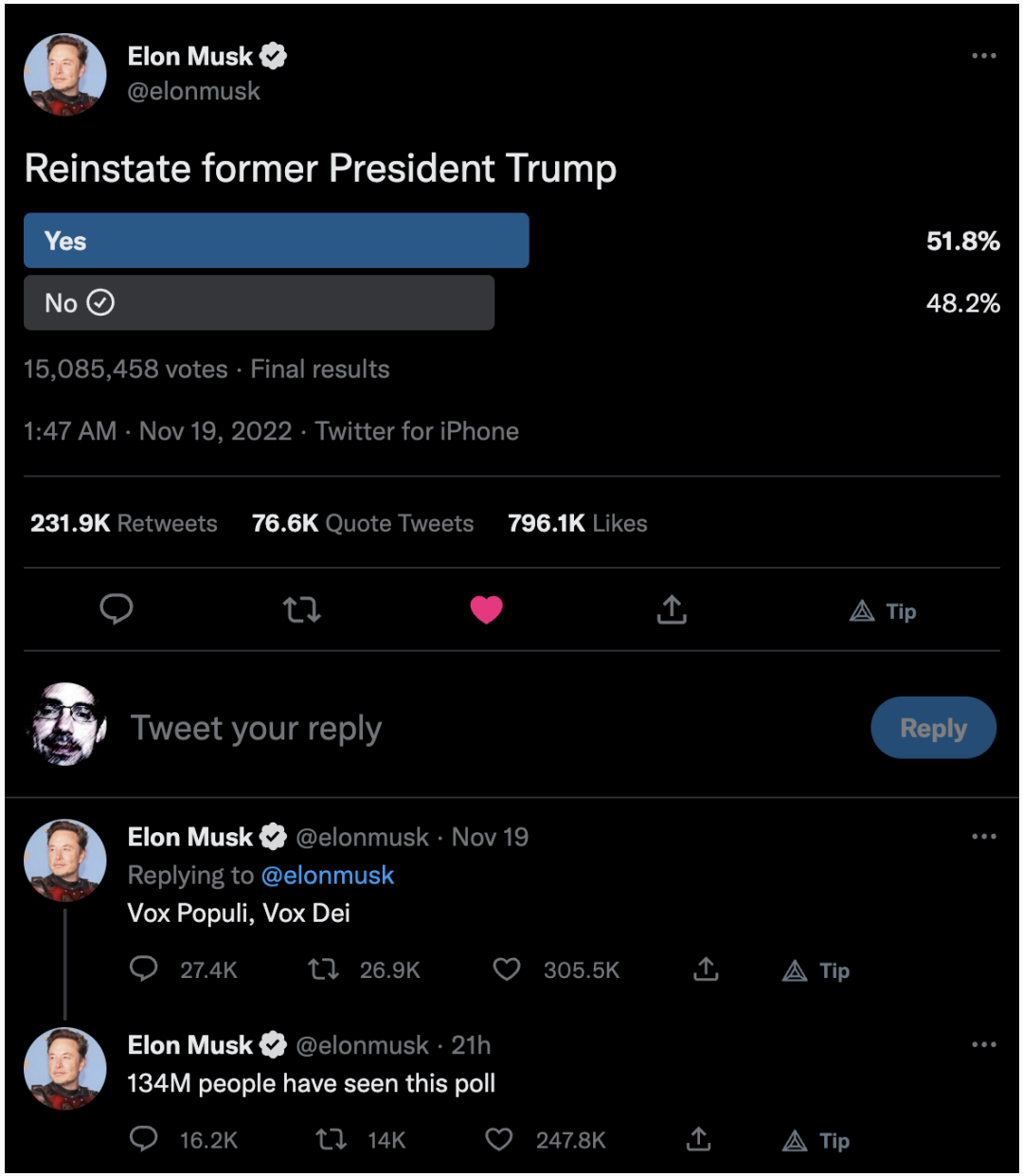


Figure 2.3 Unit and item nonresponse in a survey data file.

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

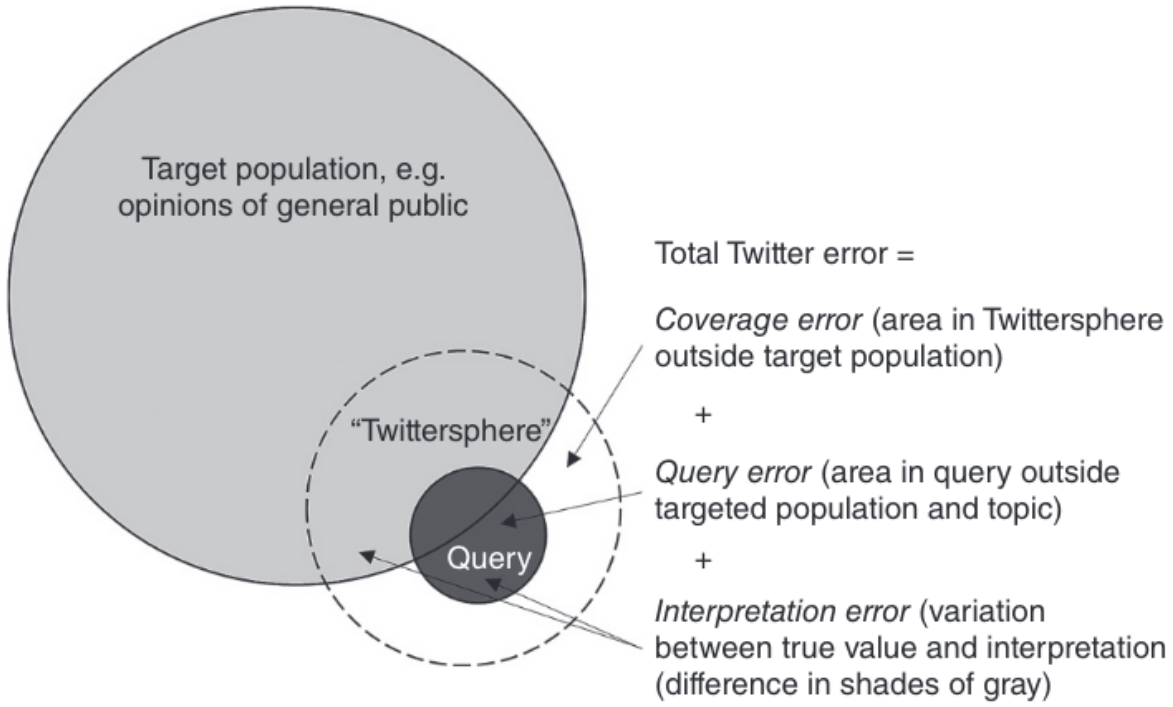


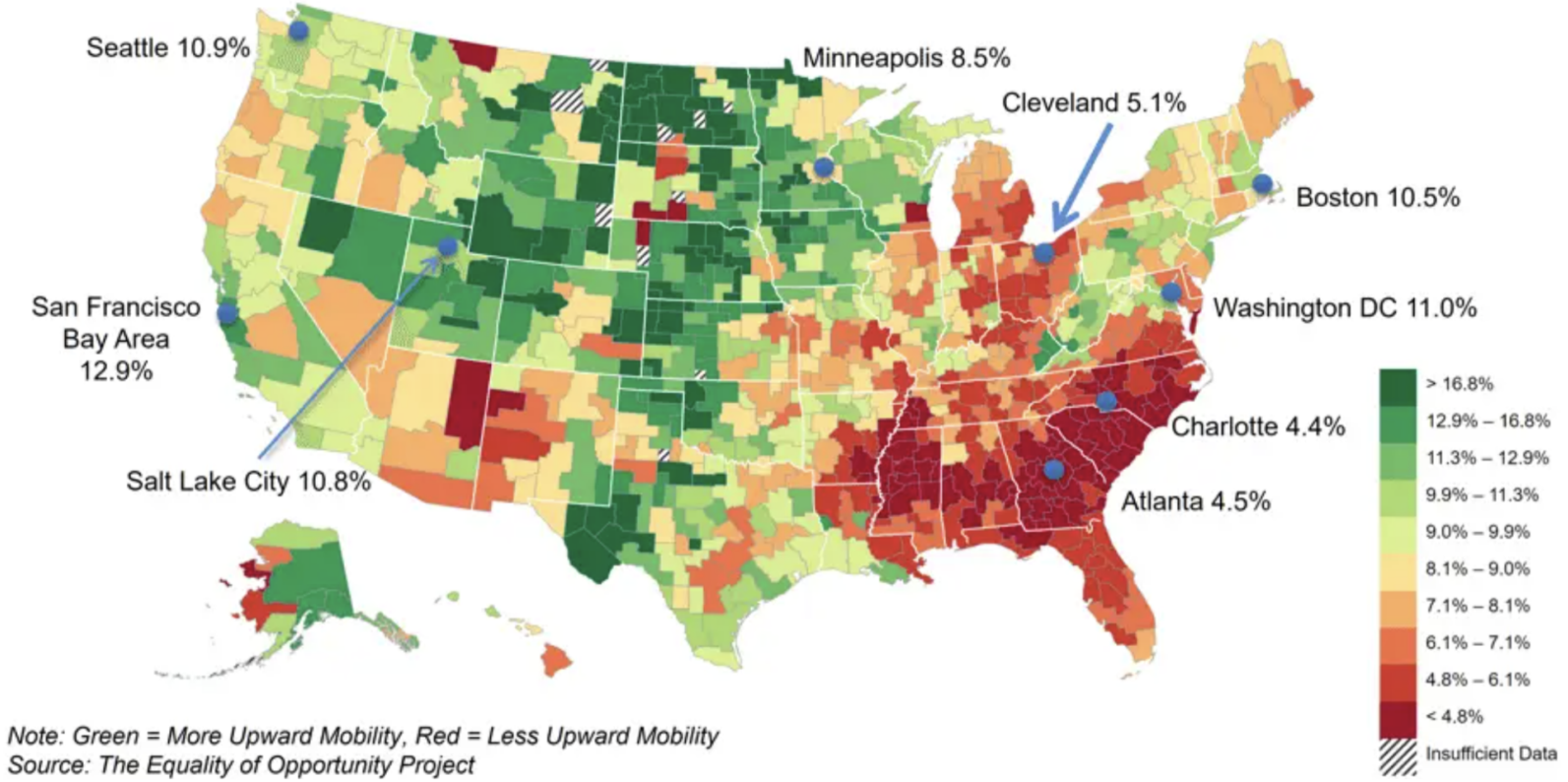
Figure 2.1 Theoretical spaces of Twitter data error.

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

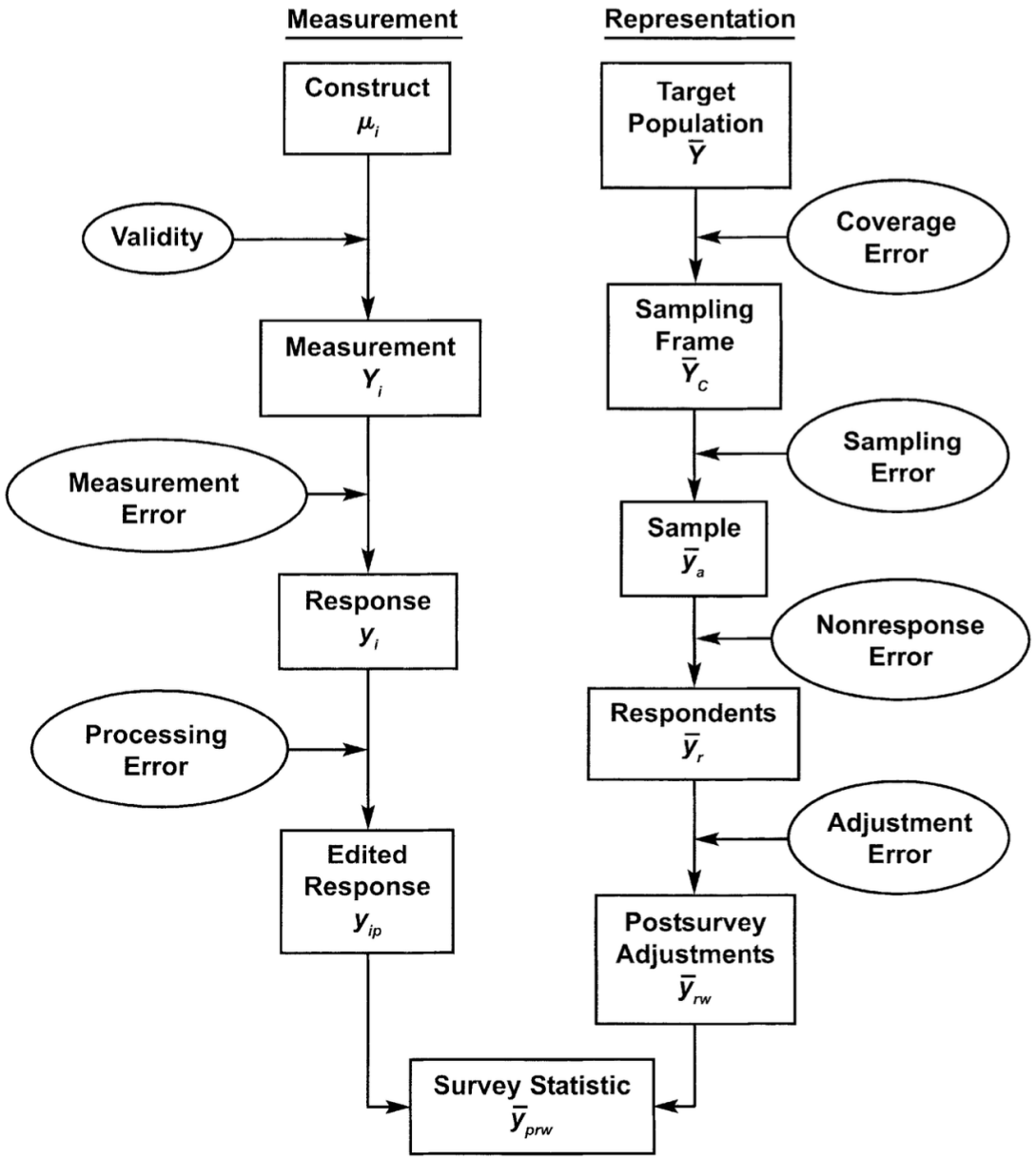


Figure 2.5 Survey life cycle from a quality perspective.

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

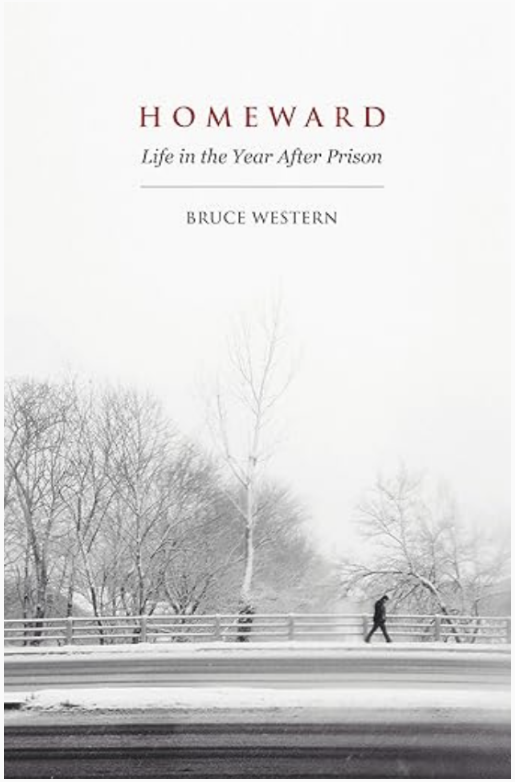
Error sources from design decisions 🤔

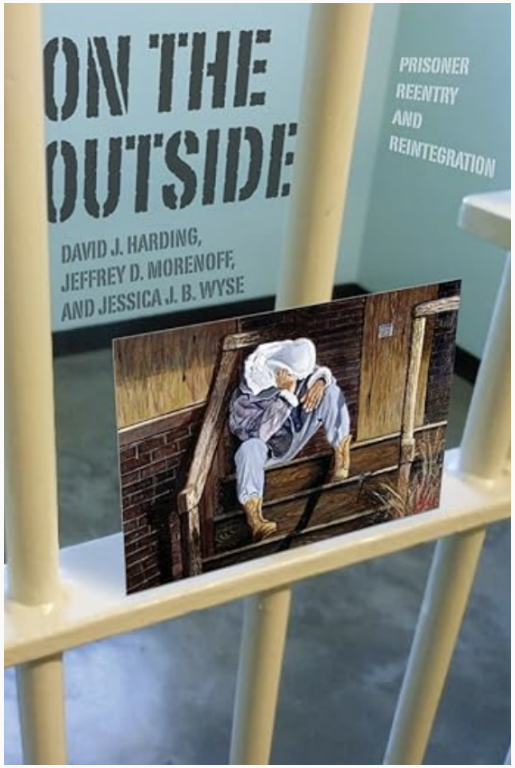
Identify which error sources might be affected

> *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

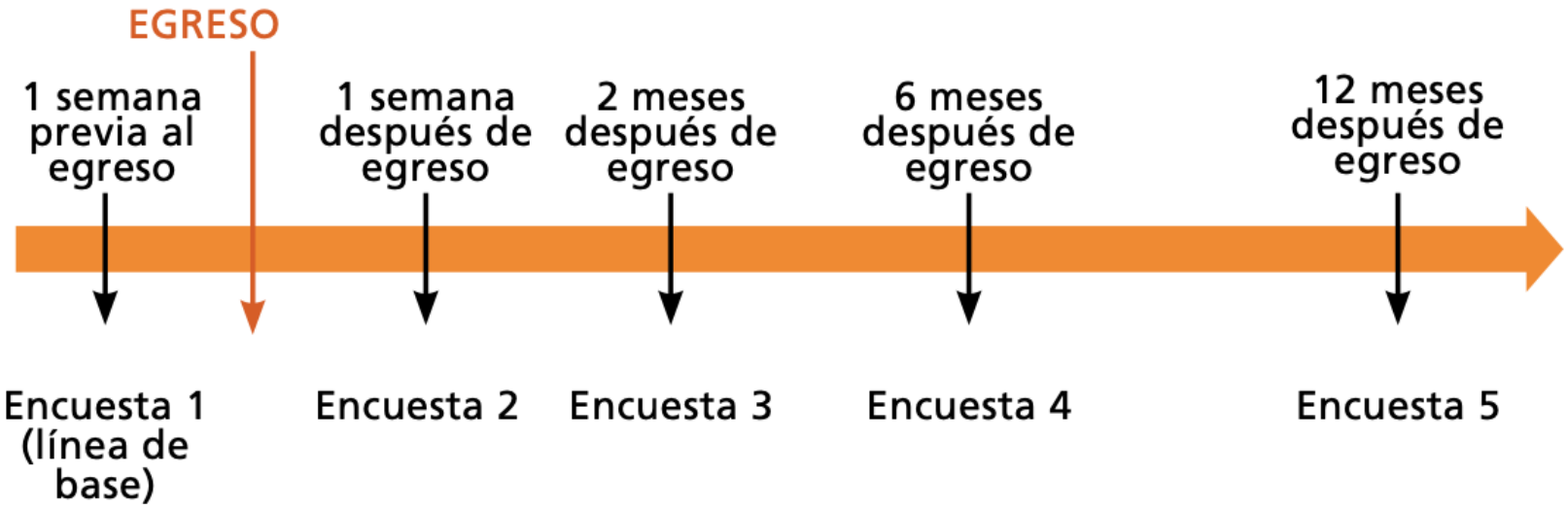
Process that individuals go through when transitioning from incarceration back into the community





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.*

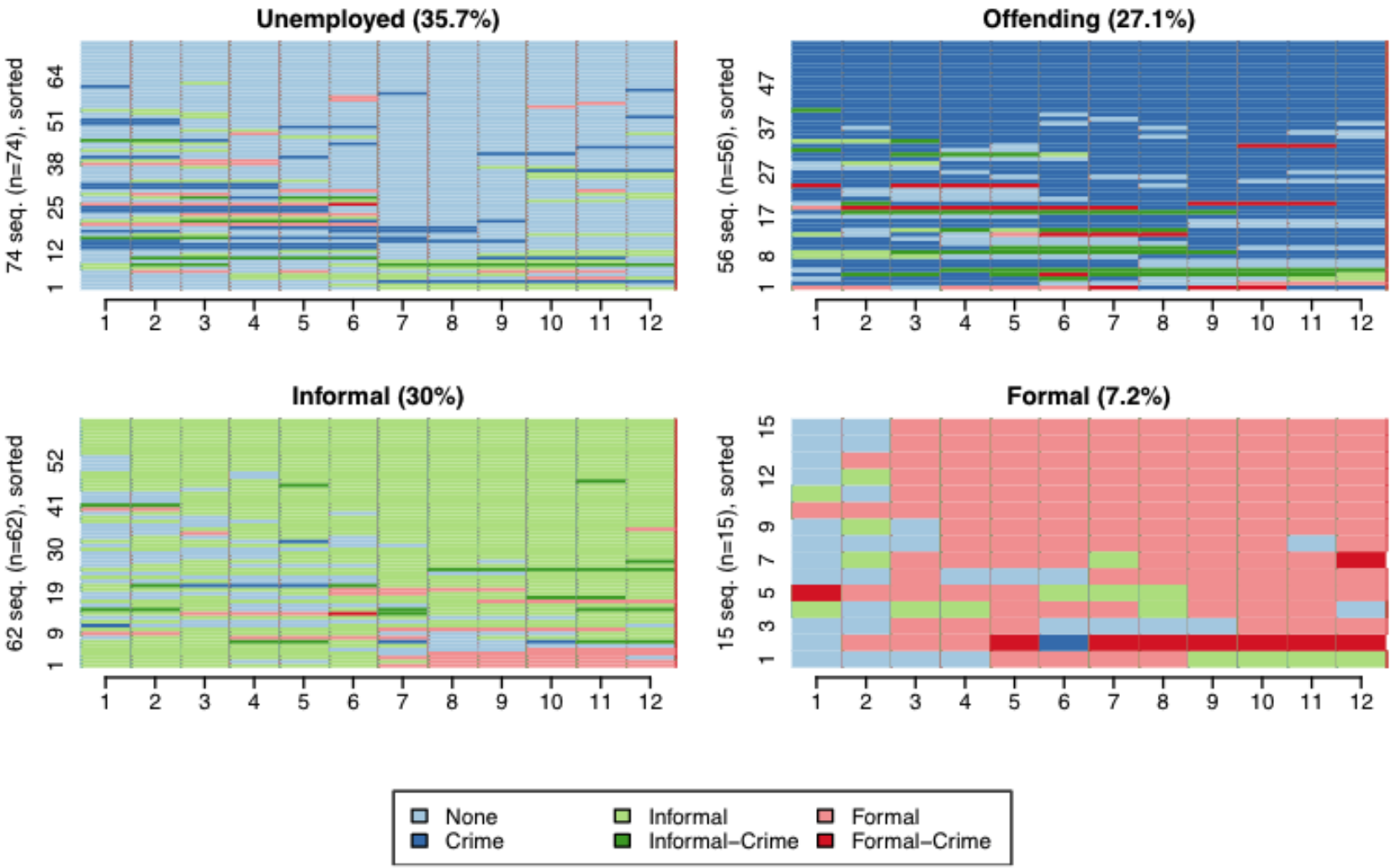
Study Design

- 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 🤓

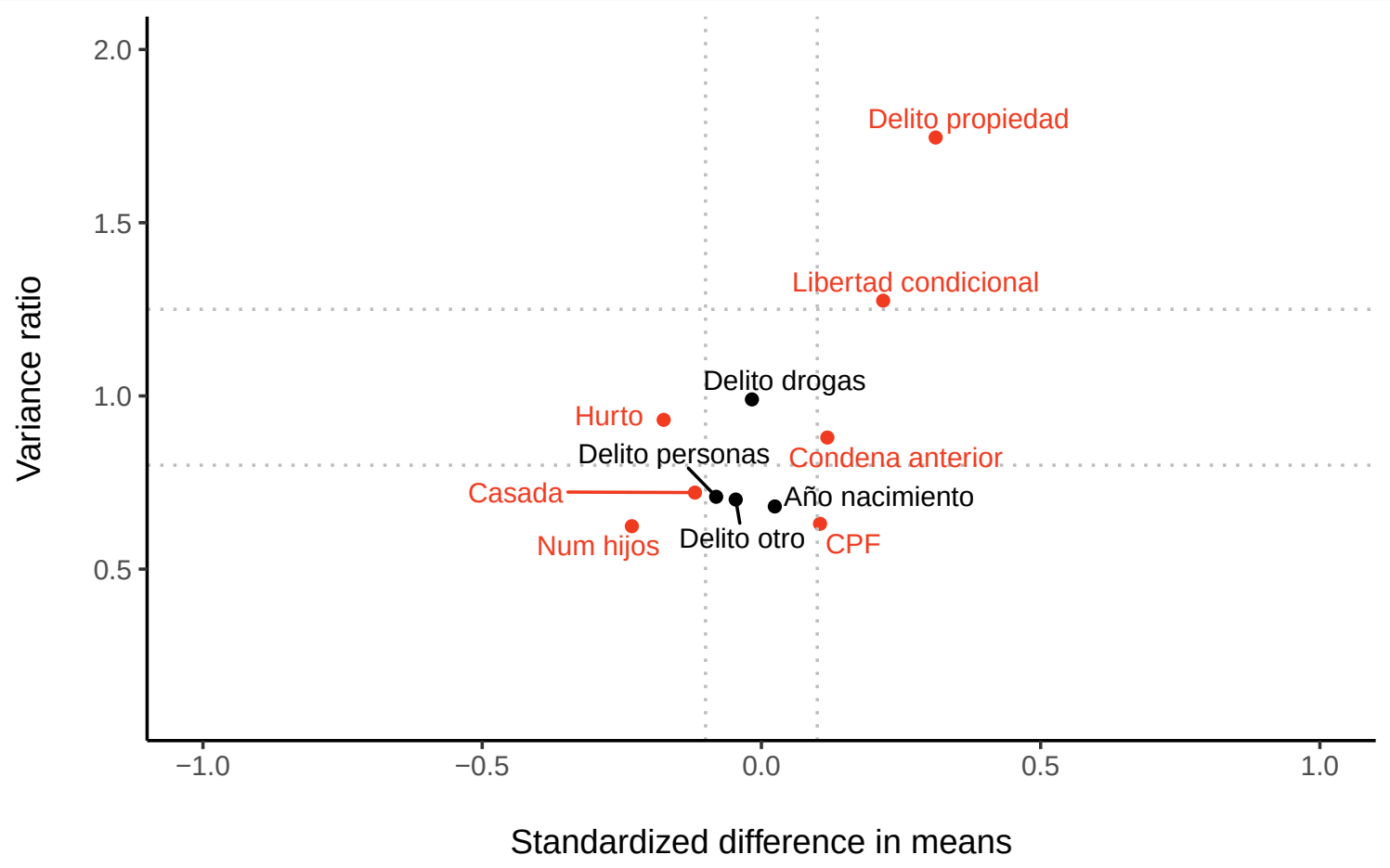
Figure S2: Sequences job-crime categories of women inmates during the first 12 months following their release by four employment-crime clusters (N = 207)



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

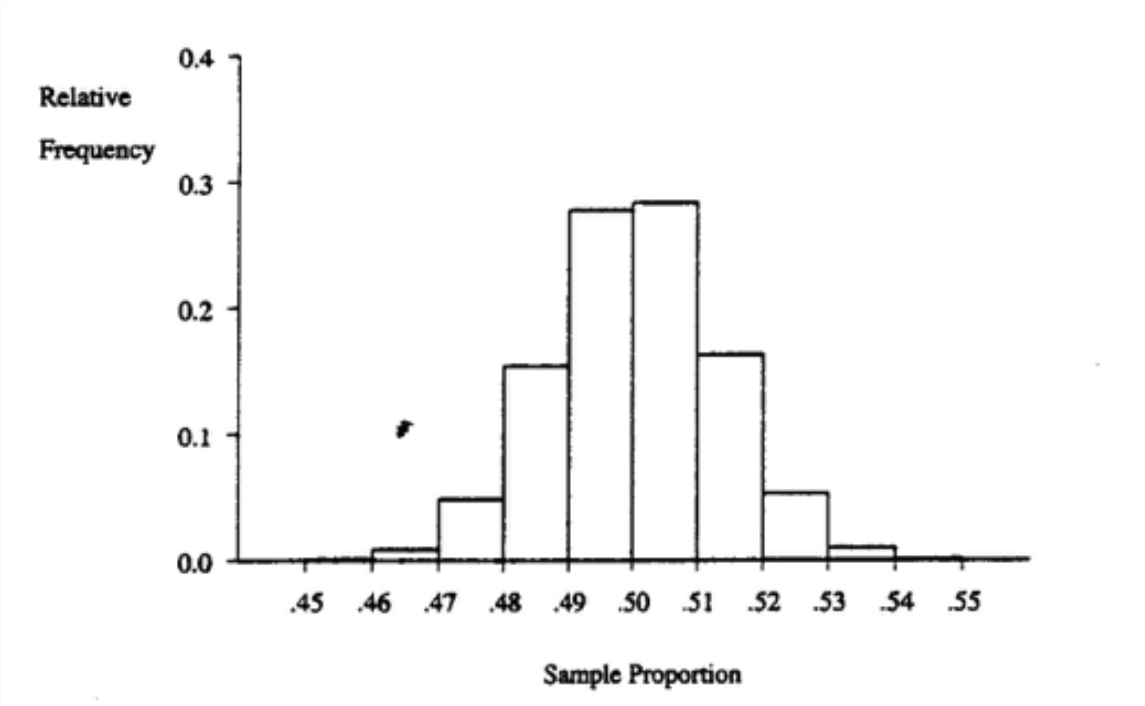
- Convenience sampling
- Quota sampling
- Judgmental or purposive sampling
- Snowball sampling

For instance, street corner interviews will be...

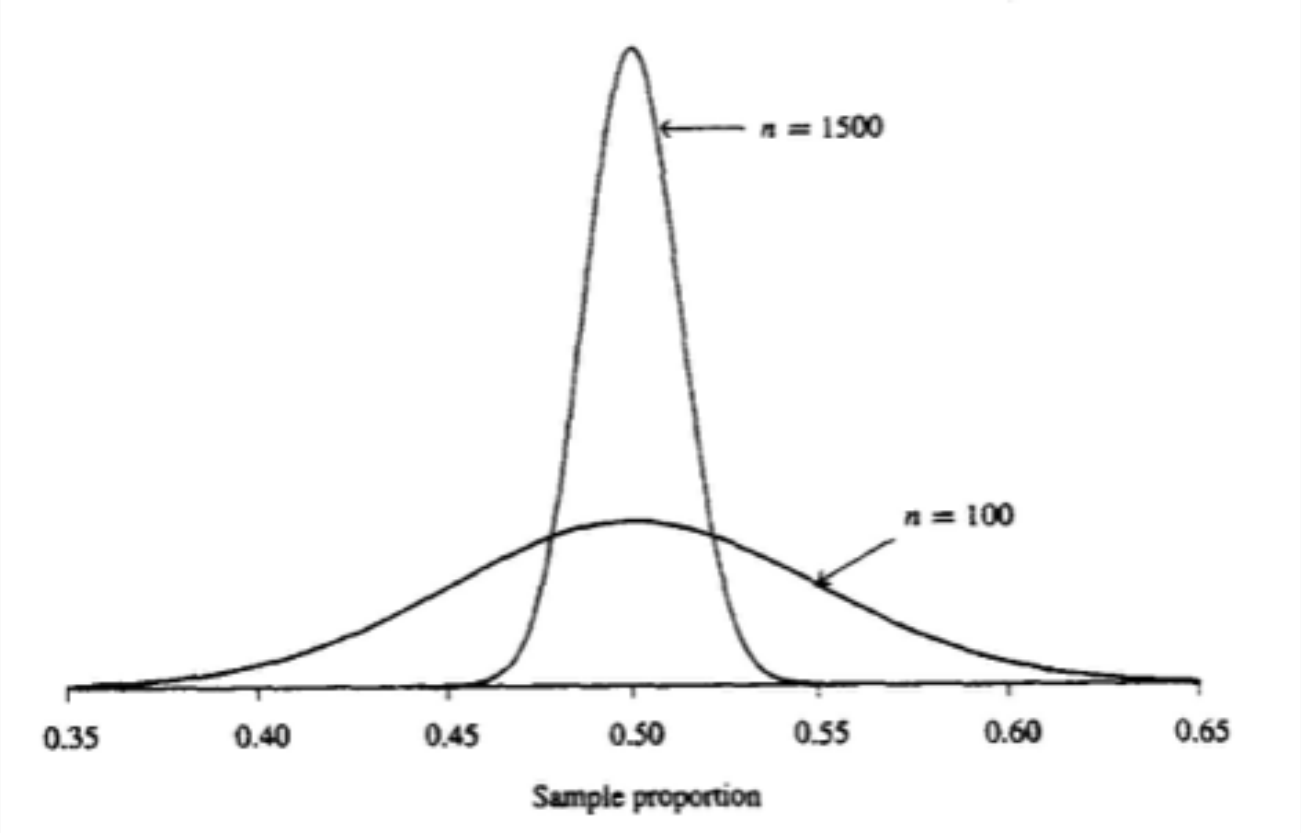
Frequentist inference refreshment

Sampling distribution

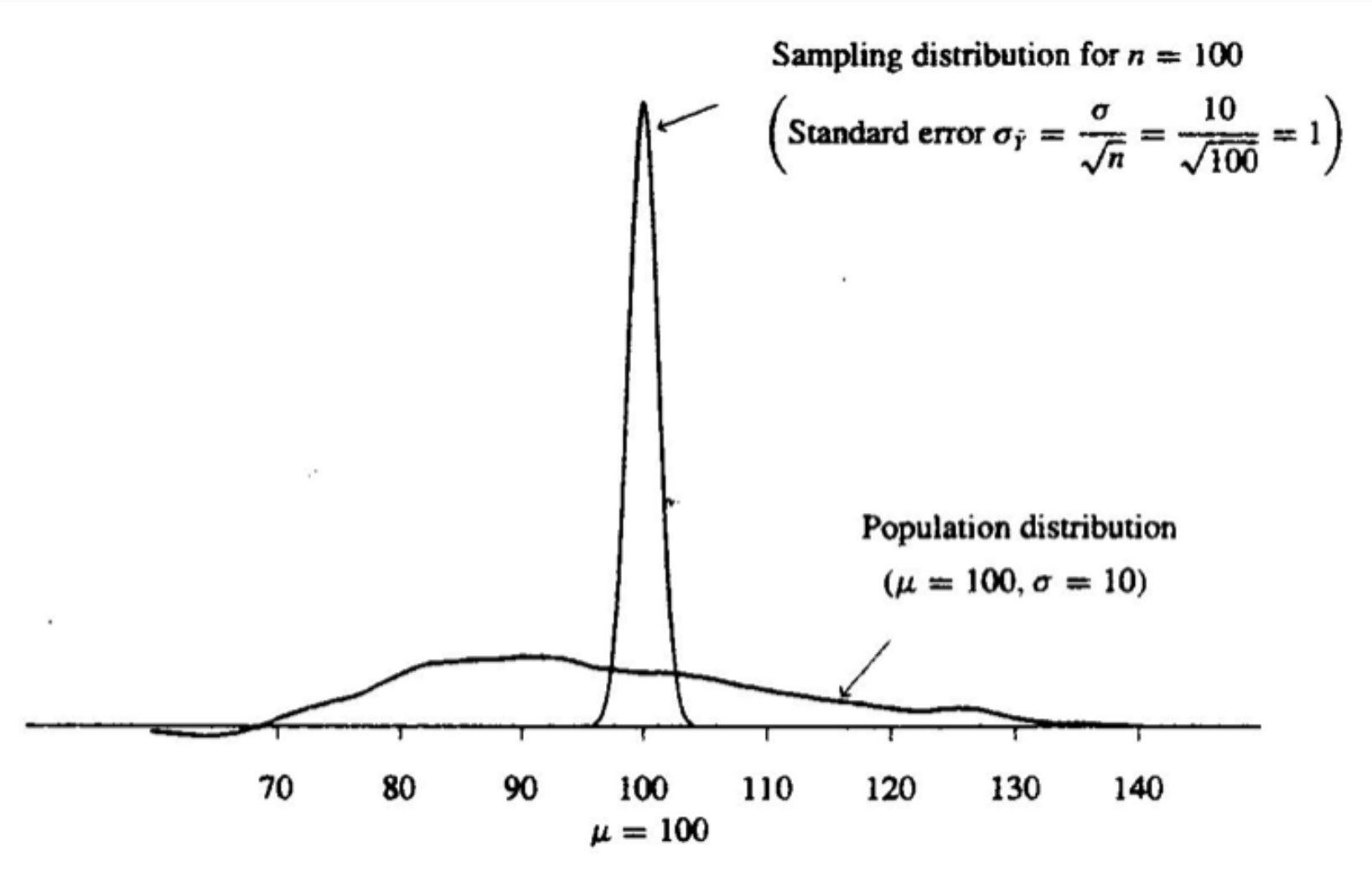
We repeat (random) sampling to get a theoretical set of possible values.
But in practice, we only get one sample.
We rest on statistical theory and the properties of repeating sampling infinitely



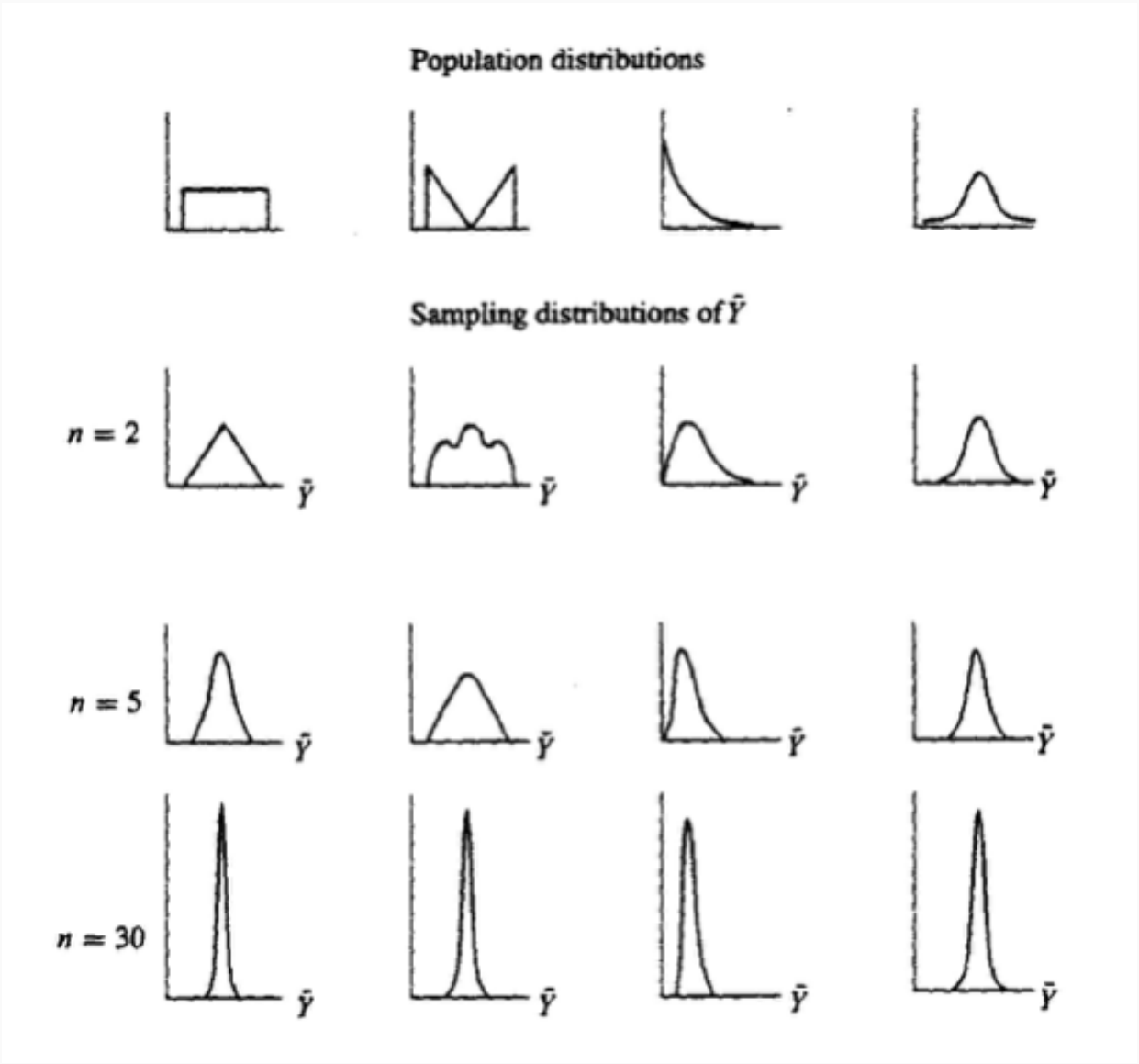
Sampling distribution



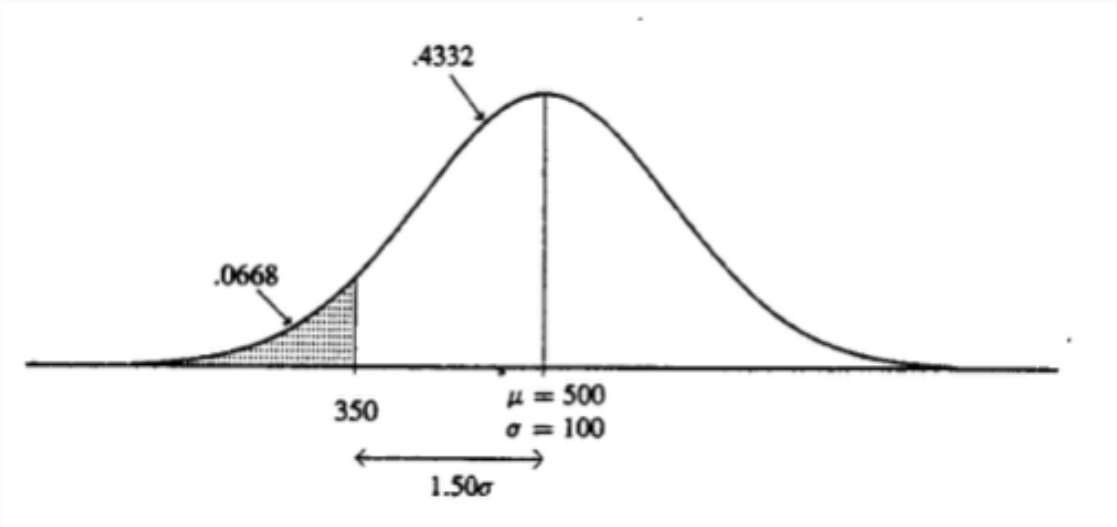
Central limit theorem



Central limit theorem

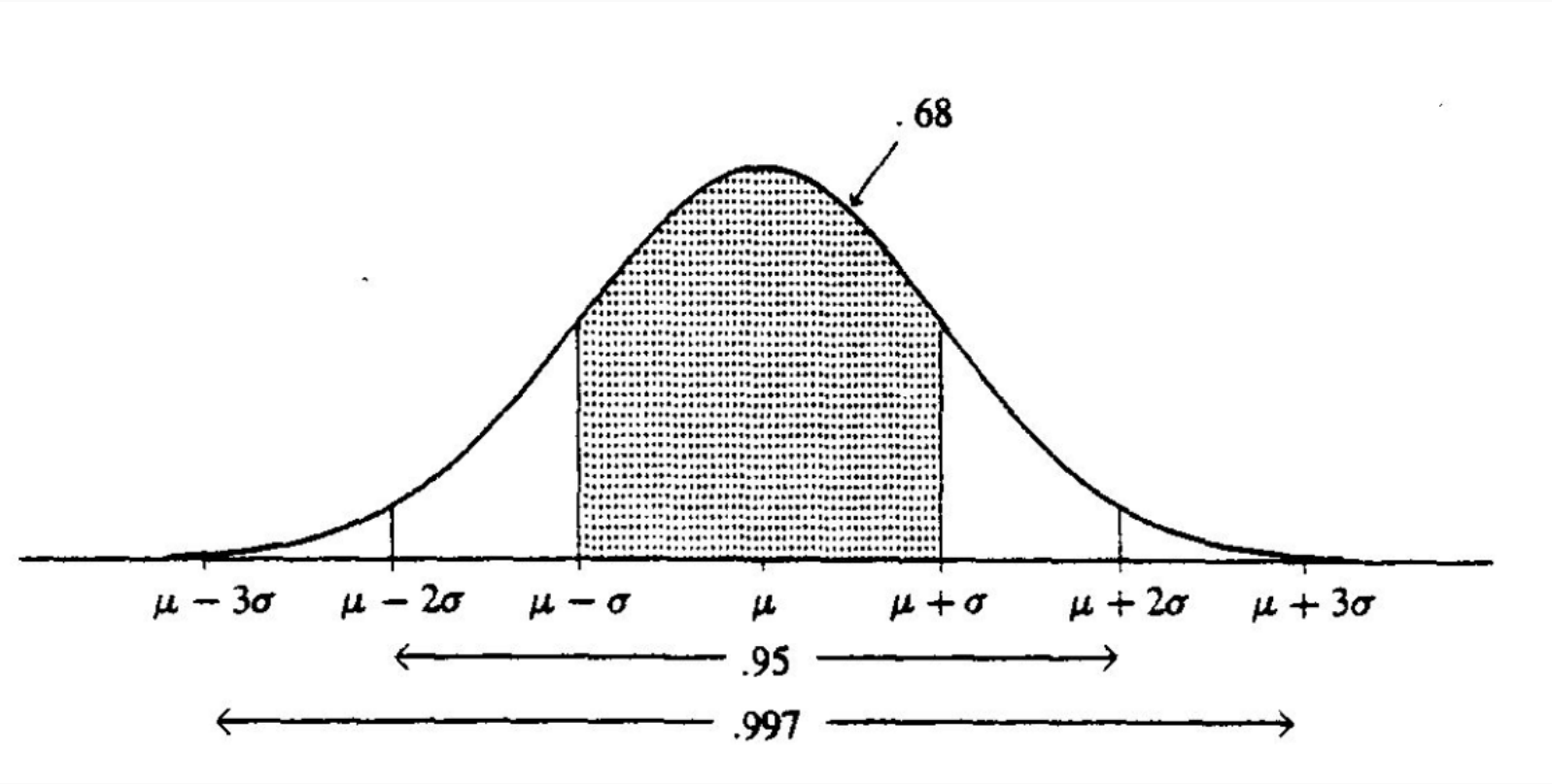


Central limit theorem

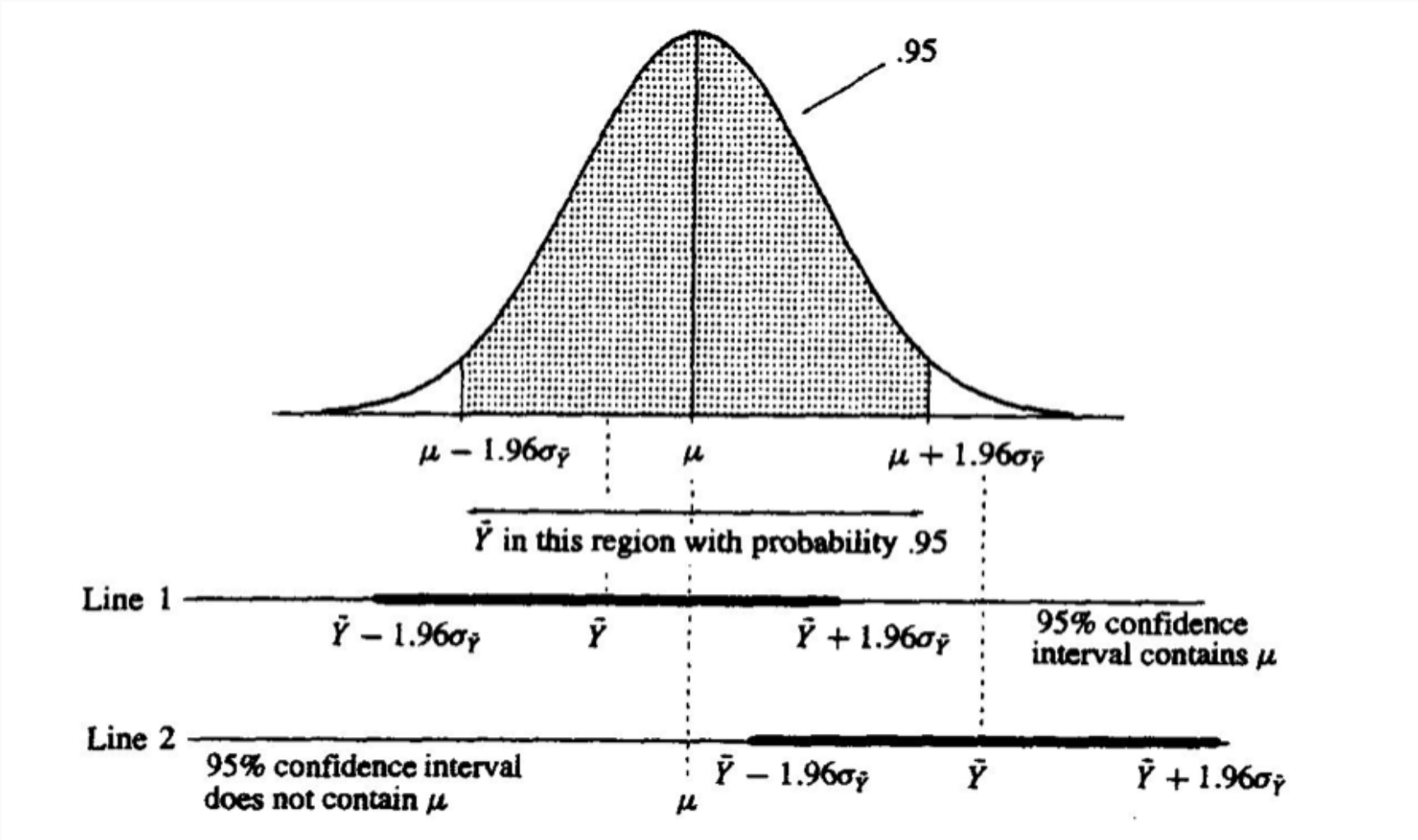



```
z = (350-500)/100
print(z)
[1] -1.5

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

Normal distribution



Confidence interval



Margin of error (MOE)

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

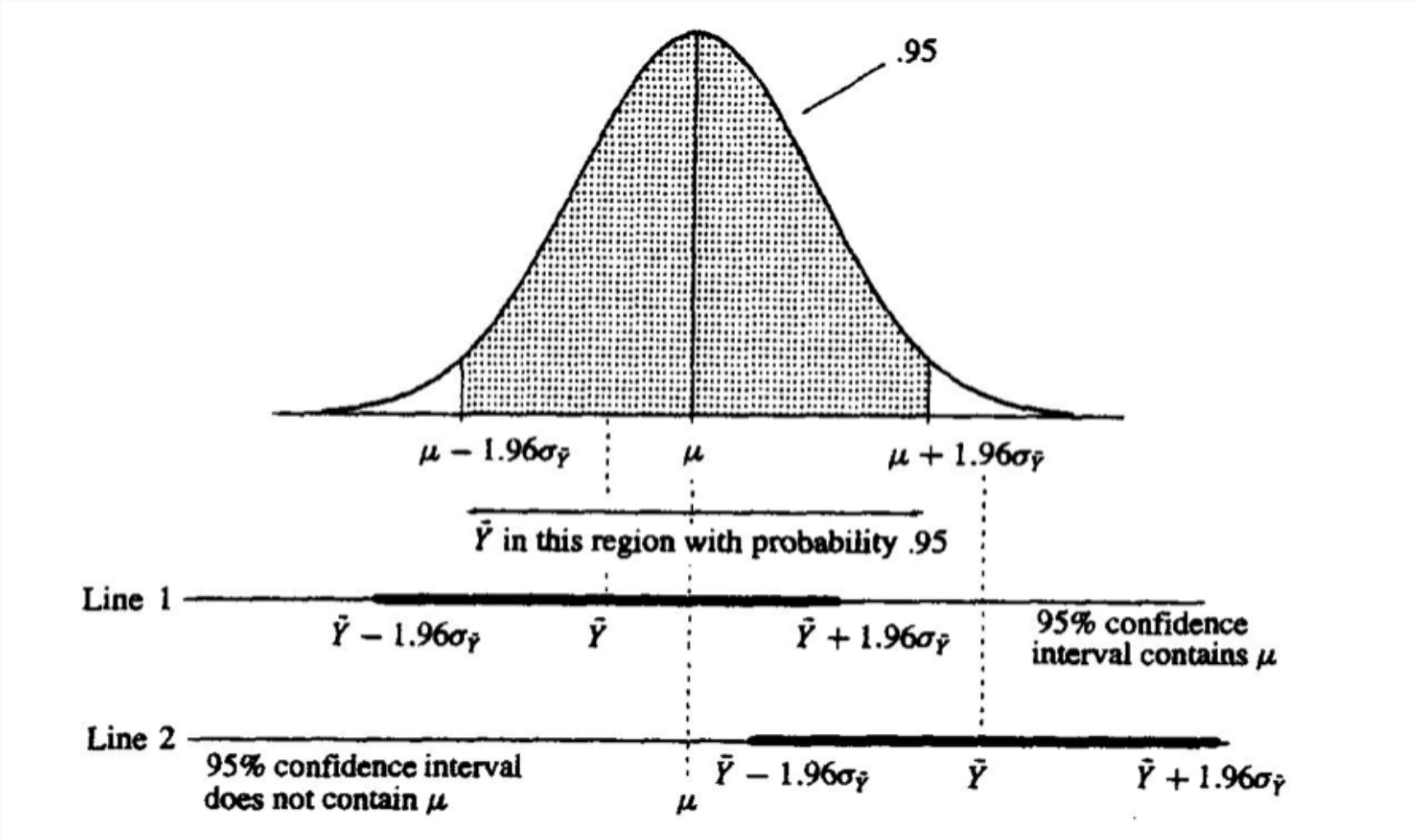
$$MOE = z * SE$$

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...

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(1 - 0.025)
[1] 1.959964 ~ 1.96
```

💻 03-freq-inference.ipynb

Confidence Interval Clarification

- A CI is a numerical interval constructed around the estimate of a parameter
- Such an interval does not, however, directly indicate a property of the parameter; instead, it indicates a property of the procedure
  - e.g., repeat across a series of hypothetical data sets (i.e., the sample space), so that to get intervals that contain the true parameter value in 95% of the cases

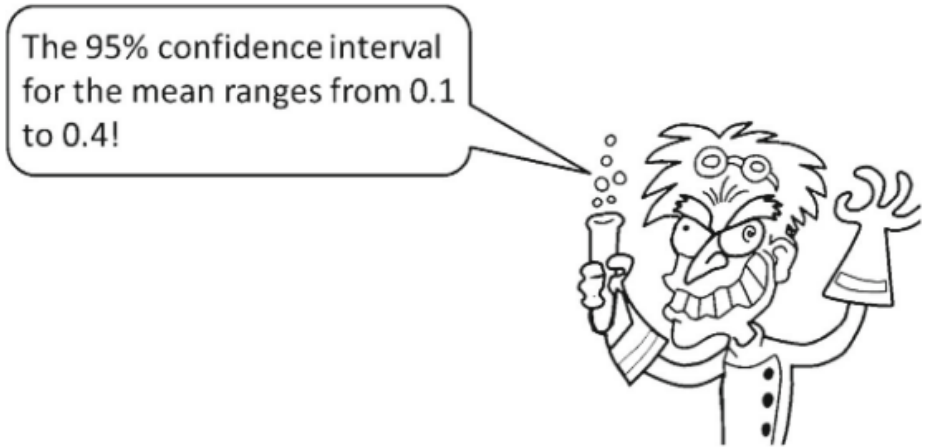
Confidence Interval Clarification

- CIs do not provide a statement about the parameter as it relates to the particular sample at hand
- They provide a statement about the performance of the procedure of drawing such intervals in repeated use
- It is incorrect to interpret a CI as the probability that the true value is within the interval

Confidence Interval Clarification

Paper: Robust misinterpretation of confidence intervals

Professor Bumbledorf conducts an experiment, analyzes the data, and reports:



Please mark each of the statements below as "true" or "false". False means that the statement does not follow logically from Bumbledorf's result. Also note that all, several, or none of the statements may be correct:

4. There is a 95% probability that the true mean lies between 0.1 and 0.4. ☐True ☐False

5. We can be 95% confident that the true mean lies between 0.1 and 0.4. ☐True ☐False

6. If we were to repeat the experiment over and over, then 95% of the time the true mean falls between 0.1 and 0.4. ☐True ☐False

Please indicate the level of your statistical experience from 1 (no stats courses taken, no practical experience) to 10 (teaching statistics at a university): _____

Confidence Interval Clarification

Paper: Robust misinterpretation of confidence intervals

Statement	First-year (442)	Master (34)	Researchers (118)
There is a 95 % probability that the true mean lies between 0.1 and 0.4	58	50	59
We can be 95 % confident that the true mean lies between 0.1 and 0.4	49	50	55
If we were to repeat the experiment over and over, then 95 % of the time, the true mean falls between 0.1 and 0.4	66	79	58

Confidence Interval Clarification

Paper: Robust misinterpretation of confidence intervals

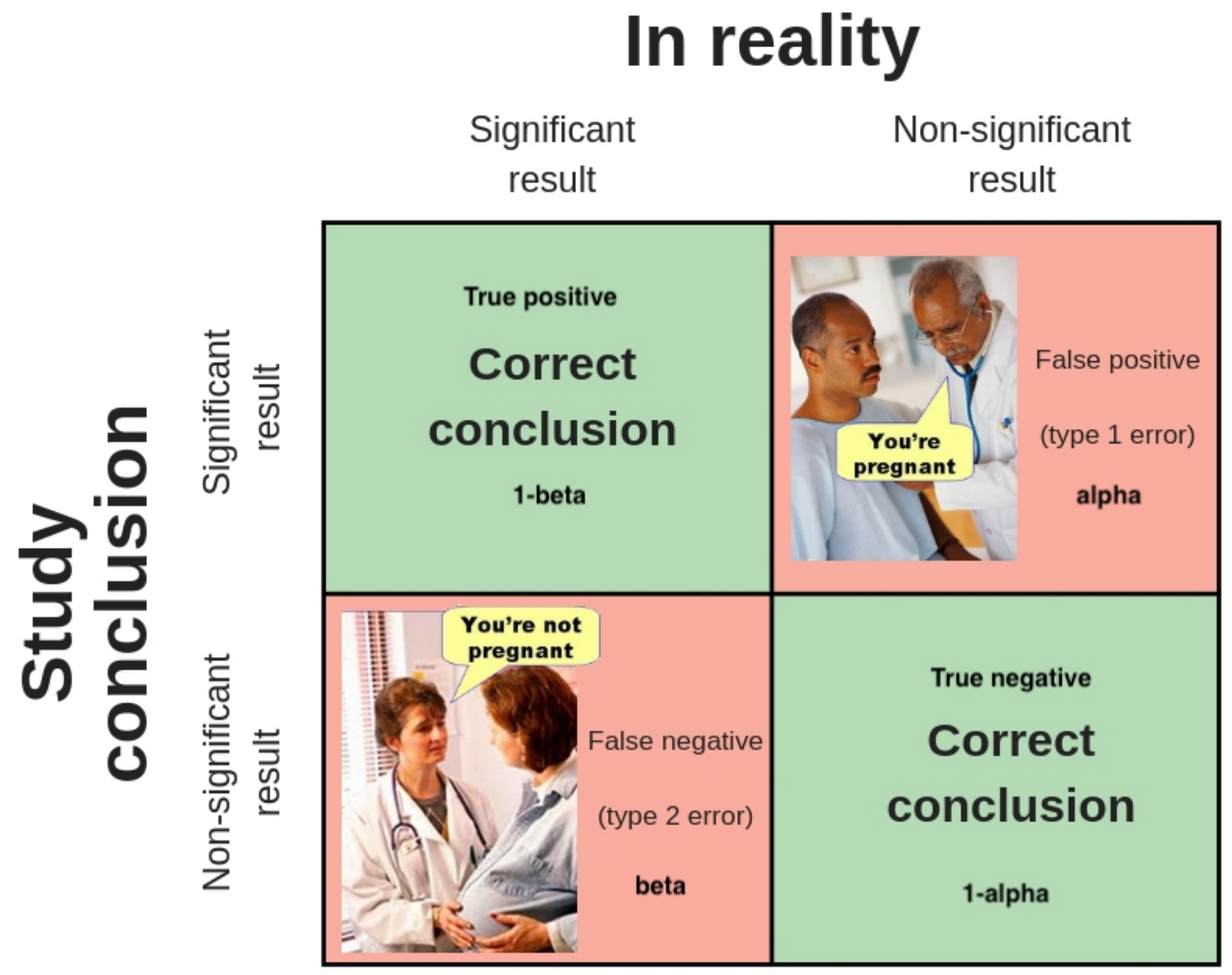
If we were to repeat the experiment (sample) over and over, then 95% of the time, the confidence intervals contain the true mean

p-values

- Hypothesis
  - The **null hypothesis** is the statement of no effect or no relationship between variables, while the **alternative hypothesis** represents the presence of an effect or relationship
  - Ordinal claims to state that there is a difference between conditions, not about the size of the effect
- p-values
  - **Probability** of observing the sample data, or **more extreme data**, assuming that the null hypothesis is true
  - Statement about the **probability of the data**, not about the probability of a hypothesis or theory!
  - **Simply tell** you that an **observation is surprising** given a null model

💻 `03-freq-inference.ipynb`

Matrix of reality



Probabilities

- α = probability of a **Type I error**, known as a *false positive*
- $1 - \alpha$ = probability of a *true negative* i.e., correctly not rejecting the null hypothesis
- β = probability of a **Type II error**, known as a *false negative*
- $1 - \beta$ = probability of a *true positive* -> correctly rejecting the null hypothesis. **Power of the test**.

p-values misconceptions

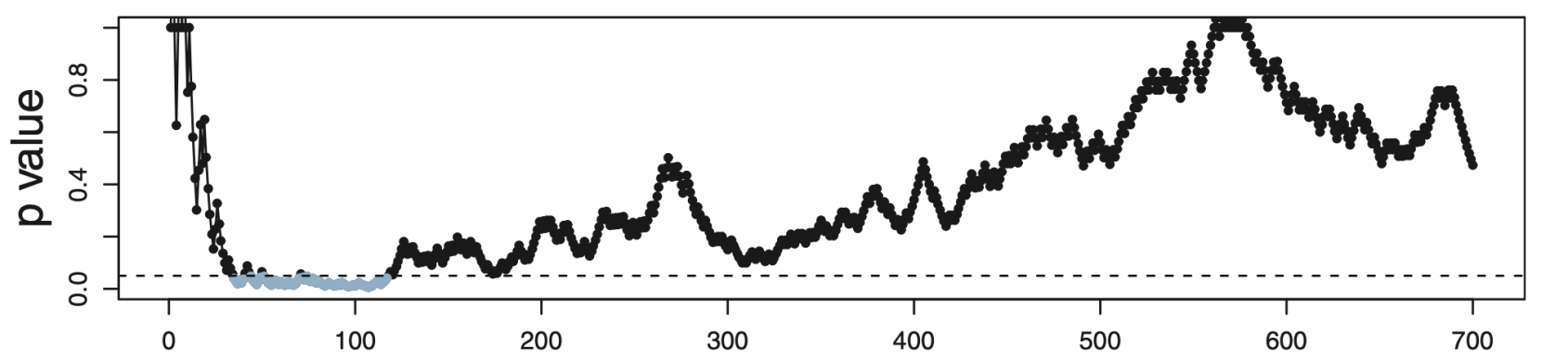
1. A non-significant p-value means that the **null hypothesis is true**
2. A significant p-value means that the **null hypothesis is false**
3. If you have observed a significant finding, the probability that you have made a **Type 1 error** (false positive) is 5%
4. One minus the p-value is the probability that the effect will replicate when repeated
5. A significant p-value means that a **practically important effect** has been discovered

Beyond power calculations

- **Type S error rate**: the probability that the replicated estimate has the incorrect sign if it is statistically significantly different from zero
- **Exaggeration ratio (expected Type M error)**: the expectation of the absolute value of the estimate divided by the effect size, if statistically significantly different from zero

Beyond power calculations

Flip a coin, where the null hypothesis is true (p=0.5)
Let's look at the evolution of p-values assuming $\alpha = 0.05$



Beyond power calculations

Neighborhood Effects in Temporal Perspective: The Impact of Long-Term Exposure to Concentrated Disadvantage on High School Graduation

Table 5. Effects of Duration-Weighted Exposure to Neighborhood Disadvantage on High School Graduation (log odds ratios)

	Blacks (n = 834)			Nonblacks (n = 1,259)		
Model	Coef	SE		Coef	SE	
Unadjusted	−.703	(.170)	***	−.581	(.109)	***
Regression-adjusted	−.416	(.196)	*	−.212	(.125)	
Stabilized IPT-weighted	−.525	(.190)	**	−.274	(.128)	*

Note: Analyses based on children not lost to follow-up before age 20. Coefficients and standard errors are combined estimates from five multiple imputation datasets.
*$p < .05$; **$p < .01$; ***$p < .001$ (two-sided tests of no effect).

Multiple comparisons

- When conducting multiple statistical tests or comparisons within a study or analysis, there is a risk of obtaining false positive results
- Family-wise Error Rate (FWER): likelihood of committing at least one Type I error (false positive) within the entire family (or batch) of comparisons

$$FWER = 1 - (1 - \alpha)^n$$

Where n is the number of comparisons and α the Type 1 Error rate, in general, 0.05.

https://sdaza.com/blog/2023/statistical-power/

p-values 🤔

Let's assume that the random number generator in R works, and we use:

```
rnorm(n = 50, mean = 0, sd = 1)
```

We generate 50 observations, and the mean of these observations is 0.5, which in a one-sample t-test against an effect of 0 produces a p-value of 0.03, which is less than the alpha level (which we have set at 0.05).

What is the probability that we have observed a significant difference (p < alpha) just by chance?

3%, 5%, 95% or 100%

Sample size computations

Sample size (proportion)

Which sample size will you need to estimate a proportion (any drug use at UCM3) with a MOE of 0.04 with a 95% confidence?

$$0.04 = 1.96\sqrt{\frac{0.5(1-0.5)}{n}}$$

$$n = \frac{1.96^2 p(1-p)}{0.04^2}$$

Which proportion should we use?

Sample size (proportion)

Which sample size will you need to estimate a proportion (any drug use at UCM3) with a MOE of 0.04 with a 95% confidence?

```
p = 0.5
sigma2 = p * (1-0.5)
print(sigma2)
0.25
n = (1.96^2 * sigma2) / 0.04^2

print(n)
600.25
```

$$n = p(1-p)\left(\frac{z}{MOE}\right)^2$$

Sample size (average)

Which sample size will you need to estimate the average number of years of education at a company with a MOE of 0.5 years and 95% confidence?

$$n = \sigma^2 \left(\frac{z}{MOE} \right)^2$$

How we get σ? Let's get a crude estimation...

Sample size (average)

- Let's say the range of values is from 0 to 15 years (of education)
- **Chebyshev's theorem**: the proportion of values in any distribution within k standard deviations will be close to $1 - (1/k^2)$ where $k > 1$.
- 3 standard deviations (SDs) will be $1 - (1/3^2) = 0.88$
- Let's say the range 0 to 15 years consists of 6 SDs
- Our estimate could be $15/6 = 2.5$

Sample size (average)

$$n = \sigma^2 \left(\frac{z}{MOE}\right)^2$$

$$n = 2.5^2 \left(\frac{1.96}{0.5}\right)^2$$

$$n = 96.04$$

Finite population correction (fpc)

- All the formulas above assume an infinite population
- When we work with finite populations, we can apply a correction factor (fpc)
  - A sample of 10 units from a population of 20 units has more information than a sample of 10 units from a population of 20,000 units
  - Decrease sampling variance

$$fpc = 1 - \frac{n}{N}$$

- When n (sample) remains relatively small concerning the population size N, fpc will be very close to 1
- If $1 - \frac{n}{N} \geq 0.95$ or if $n \leq \frac{N}{20}$, fpc corrections are not necessary

Sample size formulas with FPC: Proportion

$$n = \frac{Np(1-p)}{(N-1)\left(\frac{MOE}{z}\right)^2 + p(1-p)}$$

```
N = 2000
z = 1.96
moe = 0.04
p = 0.5
print( ( N*p*(1-p) ) / ( (N-1)*(moe/z)^2 + p*(1-p) ) )
461.86
```

Sample size formulas with FPC: Proportion

$$n = \frac{Np(1-p)}{(N-1)\left(\frac{MOE}{z}\right)^2 + p(1-p)}$$

```
N = 1000000
z = 1.96
moe = 0.04
p = 0.5
print( ( N*p*(1-p) ) / ( (N-1)*(moe/z)^2 + p*(1-p) ) )
599.89
```

Sample size formulas with FPC: Average

$$n = \frac{N\sigma^2}{(N-1)\left(\frac{MOE}{z}\right)^2 + \sigma^2}$$

```
N = 100
z = 1.96
moe = 0.5
sigma = 2.5
print( ( N*sigma^2) / ( (N-1)*(moe/z)^2 + sigma^2 ) )
49.24
```

SRS as benchmark

- All the formulas above assume simple random sampling
- However, sampling design might affect sampling variance and standard error
- We can compare variances between the sampling design used, and SRS
- An efficient sampling design concerning SRS will have value 1, less efficient will be higher than 1

$$d^2 = \frac{v(\bar{y})}{v_{srs}(\bar{y})}$$

d^2 = **design effect (deff)**

💻 `04-sample-size.ipynb`

Probability of selection

Probability of selection

- Sampling ratio: fraction of population represented by a sample $f = \frac{n}{N}$
- Probability a unit is selected from a population (when the probability of selection is constant)

Expansion factor (design weights)

- The inverse or reciprocal of the probability of selection corresponds to the **expansion factor** $W = \frac{N}{n}$
- The **expansion factor** is the number of elements represented by an element in the sample
- The sum of the **expansion factor** will be equal to the total population

What will be the expansion factor or the inverse of the probability of selecting an element in simple random sampling?

SRS Expansion factor (design weights)

Population $N = 13$, sample $n = 3$

Unit	Selected	$f = \frac{n}{N}$	$W = \frac{1}{f}$	$W_p = W * \frac{n}{N}$
1	0	0.231		
2	0	0.231		
3	0	0.231		
4	1	0.231	4.33	1
5	0	0.231		
6	0	0.231		
7	0	0.231		
8	1	0.231	4.33	1
9	0	0.231		
10	0	0.231		
11	1	0.231	4.33	1
12	0	0.231		
13	0	0.231		
Total	**3**	**3**	**13**	**3**

Using auxiliary data

- The sampling frame usually has information that can be used to design a sample
- The goal, **increase the efficiency of the sample**
- **What is the relation between the variable of study and the auxiliary information?**

Using auxiliary data

- Simple random sampling (SRS):
  - No auxiliary info is used. It serves as a benchmark
- Stratified sample (STR):
  - The population is divided into **non-overlapping** subpopulations called strata
  - Sampling is done **independently** in each stratum (in all strata).
  - If a large part of the variation in the study variable is captured by the variation between strata, then STR will be more efficient than SRS

Stratified sampling (STR)

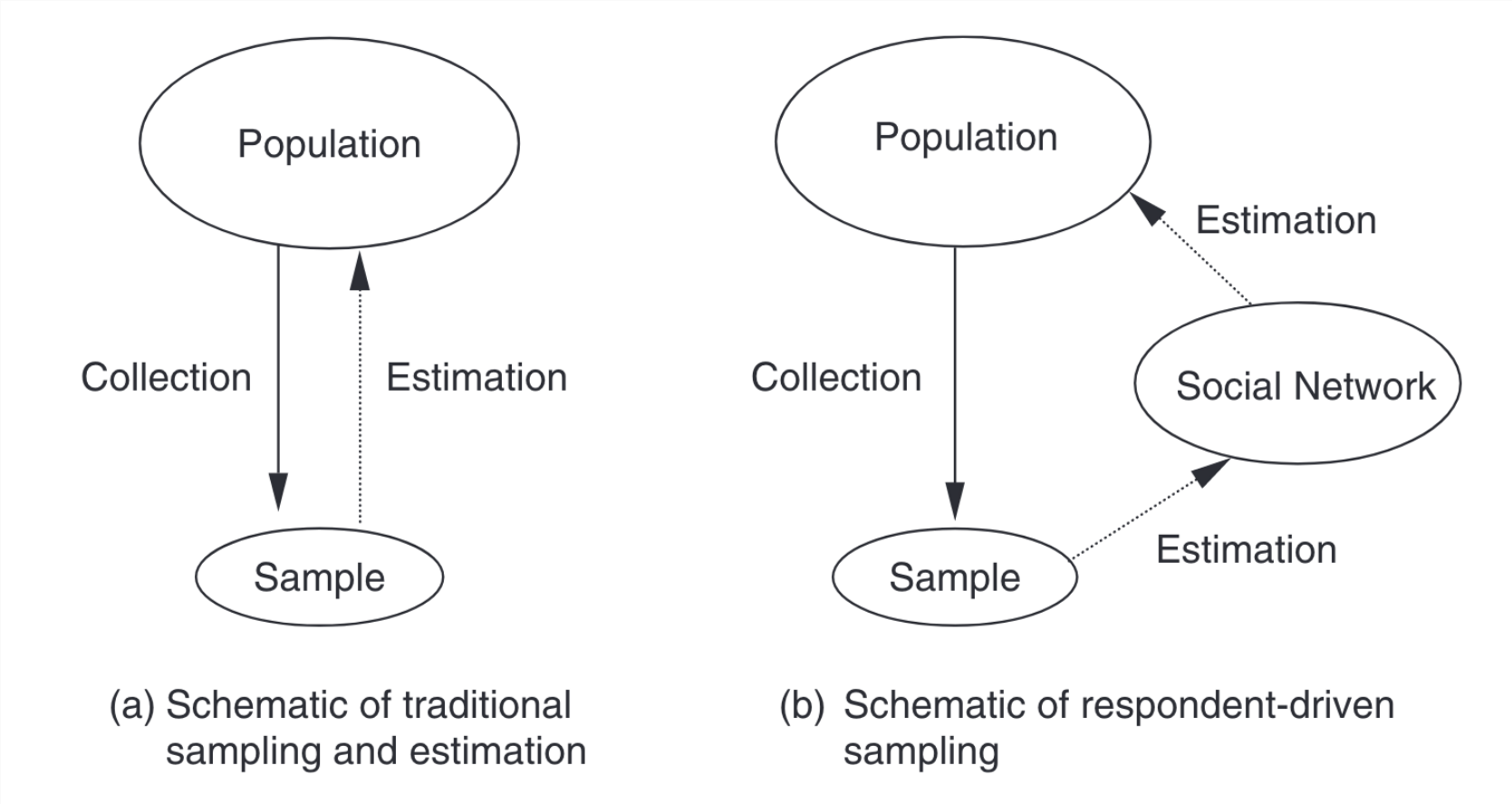
Advantages

- Estimation variance can be minimized through the definition of strata as homogeneous as possible (compared to simple random sampling)
- The cost per unit can be reduced by stratifying the elements into convenient groups (definition of compact geographic areas)
- Parameter estimates can be obtained for subgroups of the population

Respondent-driven sampling (RDS)

A method to survey populations that are difficult to reach because they are small, hidden, or mobile or because members of the target population are not interested in participating in the survey, for example, because they are engaged in socially undesirable behaviors

Respondent-driven sampling (RDS)



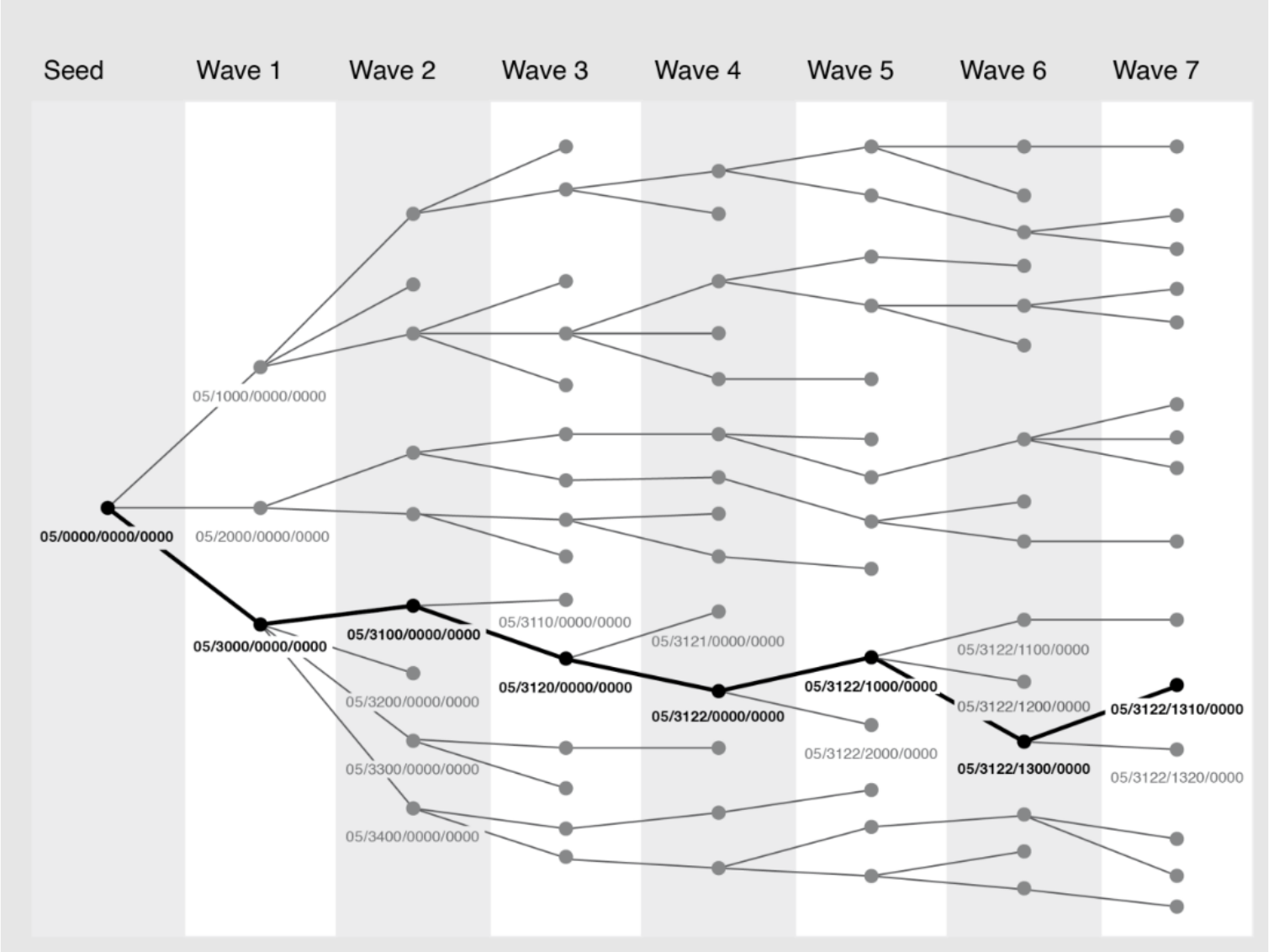
(a) Schematic of traditional sampling and estimation

(b) Schematic of respondent-driven sampling

Respondent-driven sampling (RDS)

- RDS combines **snowball sampling** and **network analysis** to achieve high statistical validity in samples that were collected using non-random procedures
- Researchers begin by selecting an initial convenience sample (seeds). They then encourage respondents to recruit their peers to participate in the survey as well.

RDS process


Seed
Wave 1
Wave 2
Wave 3
Wave 4
Wave 5
Wave 6
Wave 7
05/0000/0000/0000
05/1000/0000/0000
05/2000/0000/0000
05/3000/0000/0000
05/3100/0000/0000
05/3110/0000/0000
05/3120/0000/0000
05/3121/0000/0000
05/3122/0000/0000
05/3122/1000/0000
05/3122/1100/0000
05/3122/1200/0000
05/3122/1300/0000
05/3122/1310/0000
05/3122/1320/0000
05/3122/2000/0000
05/3200/0000/0000
05/3300/0000/0000
05/3400/0000/0000

Respondent-driven sampling

- By keeping track of the respondents' social networks and recruiting patterns, researchers can applying mathematical models of the recruitment process using Markov chains and network theory and weight the sample to compensate for the initial non-randomness of the seeds
- When correctly applied, RDS can provide population estimates of the target population that are only modestly biased

RDS Assumptions

- Network of a hidden population forms one connected component
    - A subgraph in which each pair of nodes is connected via a path
- All respondents receive and use one coupon, and when respondents recruit others, they recruit randomly from all edges that involve them
- The seeds are selected based on their degree, which represents the number of connections a node has to other nodes.

RDS - Comparisons

TABLE 1
Three Different Estimates of the Characteristics of Jazz Musician Populations

Jazz musicians in New York			
Characteristic	Union Sample ($n = 415$)	Chain-Referral Sample ($n = 251$)	RDS Estimate ($n = 251$)
Union member	100	39.6	25.4
Female	16.1	26.8	23.7
Solo only	19.9	8.8	11.6
Received airplay	79.6	81.6	74.9
Jazz musicians in San Francisco			
Characteristic	Union Sample ($n = 237$)	Chain-Referral Sample ($n = 221$)	RDS Estimate ($n = 221$)
Union member	100	11.2	4.0
Female	21.9	14.5	11.1
Solo only	15.0	9.2	21.0
Received airplay	81.6	48.7	31.4

Note: The data reflect different conclusions researchers would make depending on the source of data used. The estimates from the union sample and chain-referral sample are sample means. The respondent-driven sampling (RDS) estimates use the data from the chain-referral sample and the estimation techniques presented in this paper.

Prob vs No-Prob

- **Non-probability samples**
  - Not well articulated rationale
  - Strong modeling assumptions (not tested)
  - Documentation

Even in the age of declining response rates, accuracy in probability sample surveys is generally higher than in non-probability samples

Multilevel regression with post-stratification (MRP)

MRP is analogous in many ways to cell weighting without the troubles associated with zero or small-n cells.

$$\Pr\left(candidate_i^k\right) = \text{logit}^{-1}\left(\begin{array}{l}\alpha_0 + \beta_1\left(2012\ party\ share\right)_j + \beta_2\left(black\ share\right)_j + \\ \beta_3\left(Hispanic\ share\right)_j + \beta_4\left(white\ evang.share\right)_j + \alpha_{1[i]}^{gender} + \\ \alpha_{2[i]}^{age5} + \alpha_{3[i]}^{race4} + \alpha_{4[i]}^{edu5} + \alpha_{5[i]}^{state} + \alpha_{6[i]}^{region} + \alpha_{7[i]}^{age5,edu5} + \\ \alpha_{8[i]}^{gender,edu5} + \alpha_{9[i]}^{race5,age5} + \alpha_{10[i]}^{race5,edu5} + \alpha_{11[i]}^{race5,gender} + \\ \alpha_{12[i]}^{race,region} + \alpha_{13[i]}^{wave}\end{array}\right)$$

$$\alpha_{l[i]}^S \sim N(0,(\sigma^S)^2)$$

Back to Probabilistic Sampling

Stratified sampling (STR)

Steps

- **First step**: specify and define the strata (as homogeneous as possible)
- **Second step**: define the sample size in each stratum (allocation)
- **Third step**: select a random and independent sample in each stratum (all strata)

Different allocations

We can afford a sample of ($n_t = 2000$) individuals across ($r = 13$) regions

Region	Population 18+ (N_r)	$\frac{N_r}{N_t}$	$\frac{N_r}{N_t} \times n_t$	$\frac{n_t}{r}$
1	291,854	0.028	56	154
2	335,859	0.032	64	154
3	168,264	0.016	32	154
4	407,105	0.039	78	154
5	1,086,122	0.104	208	154
6	531,142	0.051	102	154
7	618,865	0.059	119	154
8	1,274,496	0.122	244	154
9	583,294	0.056	112	154
10	729,897	0.070	140	154
11	60,736	0.006	12	154
12	108,129	0.010	21	154
13	4,248,842	0.407	814	154
Total	**10,444,605** (N_t)	**1**	**2000**	**2000**

The efficiency of allocation will depend on:

- Units in each stratum
- Variance within the stratum
- Cost

Proportion formula with fpc:

$$n = \frac{\sum_{i=1}^{L} \frac{N_i^2 p_i(1-p_i)}{w_i}}{N^2 \left(\frac{MOE}{z}\right)^2 + \sum_{i=1}^{L} N_i p_i (1-p_i)}$$

STR: w_i

w_i = fraction of units assigned to a stratum

c_i = cost per unit in stratum i

$$w_i = \frac{N_i \sqrt{\frac{p_i(1-p_i)}{c_i}}}{\sum_{i=1}^{L} N_i \sqrt{\frac{p_i(1-p_i)}{c_i}}}$$

STR

Define a sample from an organization of **4500** workers, where there are three types of workers $A = 400$, $B = 2300$, and $C = 1800$, to estimate the proportion of workers who agree with a union demand. Assume max variance, 95% confidence, and a MOE of **0.03**. The cost per survey is the same in all strata.

STR: Allocation

First, define the assignment per stratum. Let's assume proportional allocation with $c = 1$

$$w_i = \frac{N_i \sqrt{\frac{p_i(1-p_i)}{c_i}}}{\sum_{i=1}^{L} N_i \sqrt{\frac{p_i(1-p_i)}{c_i}}}$$

- $A = 400\sqrt{\frac{0.25}{1}} = 200$
- $B = 2300\sqrt{\frac{0.25}{1}} = 1150$
- $C = 1800\sqrt{\frac{0.25}{1}} = 900$
- $\sum_{i=1}^{L} N_i \sqrt{\frac{p_i(1-p_i)}{c_i}} = 200 + 1150 + 900 = 2250$
- $w_A = 0.09$; $w_B = 0.51$; $w_C = 0.4$

STR: Sample size

$$n = \frac{\sum_{i=1}^{L} \frac{N_i^2 p_i(1-p_i)}{w_i}}{N^2 \left(\frac{MOE}{z}\right)^2 + \sum_{i=1}^{L} N_i p_i(1-p_i)}$$

- $\sum_{i=1}^{L} \frac{N_i^2 p_i(1-p_i)}{w_i} = \frac{400^2 * 0.25}{0.09} + \frac{2300^2 * 0.25}{0.51} + \frac{1800^2 * 0.25}{0.40} = 5062582$
- $\sum_{i=1}^{L} N_i p_i(1-p_i) = 400 * 0.25 + 2300 * 0.25 + 1800 * 0.25 = 1125$
- $n = \frac{5062582}{4500^2 \left(\frac{0.03}{1.96}\right)^2 + 1125} = 863$

STR: Sample size with simple allocation

If we assign the same w_i to each stratum:

$$n = \frac{\sum_{i=1}^{L} \frac{N_i^2 p_i(1-p_i)}{w_i}}{N^2 \left(\frac{MOE}{z}\right)^2 + \sum_{i=1}^{L} N_i p_i(1-p_i)}$$

- $\sum_{i=1}^{L} \frac{N_i^2 p_i(1-p_i)}{w_i} = \frac{400^2 * 0.25}{0.33} + \frac{2300^2 * 0.25}{0.33} + \frac{1800^2 * 0.25}{0.33} = 6517500$
- $\sum_{i=1}^{L} N_i p_i(1-p_i) = 400 * 0.25 + 2300 * 0.25 + 1800 * 0.25 = 1125$
- $n = \frac{6517500}{4500^2 \left(\frac{0.03}{1.96}\right)^2 + 1125} = 1110$

STR: So when to use it?

When:

- The allocation is proportional
- Costs are constant across strata
- Variance is the same across strata

We get the same results using the stratified sample size formula as the SRS formula

Still, it is a good idea to stratify because the conditions above are almost always not met

💻 `05-stratification.ipynb`

Using auxiliary information

- Systematic sampling (SYS)
  - Auxiliary information is used to order the list of elements in a population
  - If the variable of interest changes systematically with the order of the list, **SYS** will be more efficient than **SRS**

Systematic sampling (SYS)

```
population = 340; sample = 39; diff = c()
k = population/sample
print(k)
8.717949

unit = 3 # random start from 1 to 9
for (i in 2:38) {
v = round(unit[length(unit)] + k)
diff = c(diff, v - unit[length(unit)])
unit = c(unit, v)
}
head(unit, 10)
3 12 21 30 39 48 57 66 75 84 ...
mean(diff)
9
```

Systematic sampling (SYS)

- A population is random if its elements are randomly ordered. SRS will be completely equivalent to SYS
- A population is ordered if the elements within the population are ordered in magnitude according to some scheme related to the variable of interest
  - The variance will be smaller than in SRS if such an order exists
- A population is periodic if the elements of the population have a cyclic variation
  - Selection of a sample of sales from a company every Wednesday.
  - In this case, it is convenient to keep changing the starting point of the systematic jump

Cluster sampling

Random sample in which each sampling unit is a collection or cluster of elements

Why?

- No good sampling frame that lists the elements of the population (unavailable or too expensive)
- A sampling frame that lists clusters is readily available
- The cost of obtaining observations (in general) increases with the distance separating the elements

Cluster sampling (CL)

Random sample in which each sampling unit is a collection or cluster of elements

Definition of clusters

- Elements within a cluster are often physically close together and so that, tend to resemble each other
- Thus, the amount of information about a population parameter may not increase substantially when taking new measurements from a cluster
- Clusters should (ideally) be as heterogeneous (or different) as possible internally and similar between them

CL vs STR

There might be some similarity between cluster and stratified sampling as the population is divided into non-overlapping groups of elements, but...

- If groups are considered strata, then a random sample is selected within each group (all of them)
- If groups are considered clusters, then a random sample of groups is drawn (we don't use all of them)

Cluster sampling designs

Popular designs

- Single-stage
  sample of clusters $\rightarrow$ all units within clusters
- Two-stage (sub-sampling)
  sample of clusters $\rightarrow$ sample of units within clusters
- Multiple-stage
  sample clusters L1 $\rightarrow$ sample clusters L2 $\rightarrow$ sample of units

The selection of clusters could also be:

- SRS
- PPS (probability-proportional-to-size)

Within-cluster homogeneity

How homogeneous is a cluster?

- ρ or rate homogeneity
- Intra-class correlation (ICC)

It's related to deff (approximation):

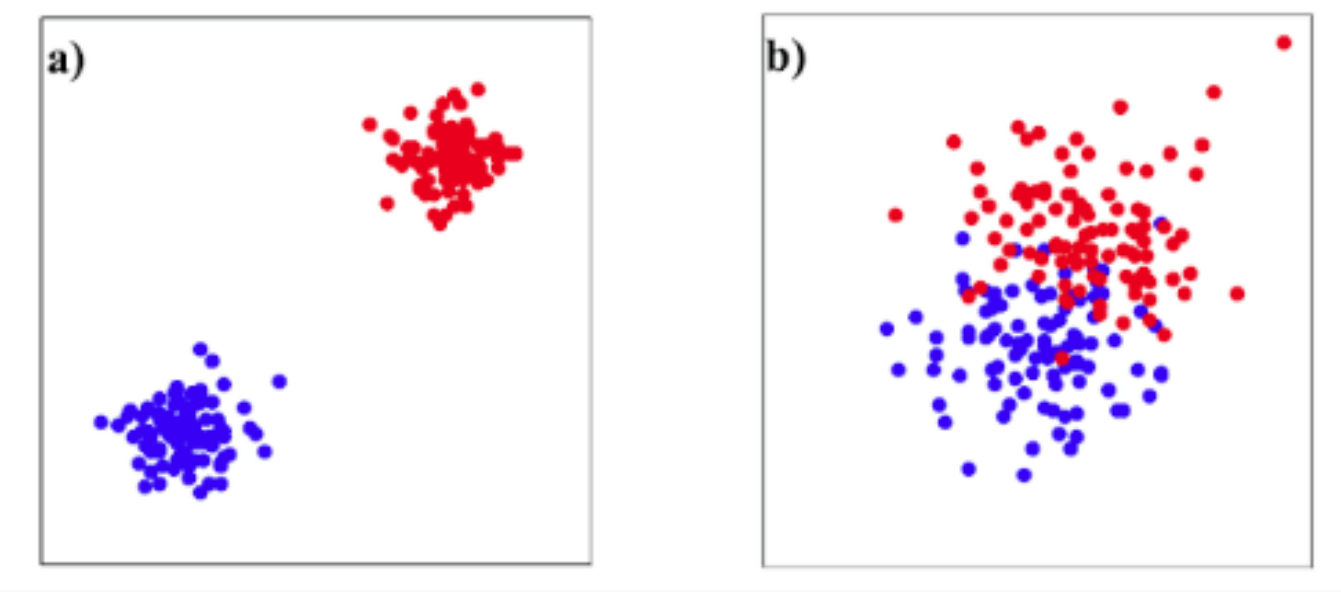
$$Deff = 1 + (n_c - 1)\rho$$

- n_c = cluster sub-sample (e.g., the average sample of clusters)

Important: all these estimates are variable-specific!

Within-cluster homogeneity

$$Deff = 1 + (n_c - 1)\rho$$



Probability-proportional-to-size sampling (PPS)

Helpful when units (clusters) have different sizes (number of units)

- When the size is correlated with the variable of interest, sample units can be selected proportionally to their size so that those of larger size are more likely to be selected than those of smaller size
- Elements (clusters) will have different probabilities of selection (based on size)
  - Estimates should be weighted
  - Otherwise, values will be biased toward larger clusters
  - There are some exceptions!

PPS: Procedure

Table 4.3. Block Housing Unit Counts and Cumulative Counts for a Population of Nine Blocks

Block	Housing units on Block α	Cumulative	Selection Numbers for the Block
1	20	20	001–020
2	100	120	021–120 ← 039
3	50	170	121–170 ← 144
4	15	185	171–185
5	18	203	186–203
6	45	248	204–248
7	20	268	249–268 ← 249
8	35	303	269–303
9	12	315	304–315

```
svalues = sample(1:315, 3)
print(svalues)
39 144 249
```

PPS: Procedure

```
block = 1:9
size = c(20, 100, 50, 15, 18, 45, 20, 35, 12)

blocks = data.table(block, size)

blocks[, end := cumsum(size)]
blocks[, start := shift(end, fill=1)]
selected = sapply(svalues, function(x) {
    blocks[block==which(start< x and end >=x),
    block]})
```

PPS: probability of selection

$$f_{block} = \frac{n_{sample} * N_{cluster_i}}{N_{total}}$$

- n_{sample} = clusters being sampled
- $N_{cluster_i}$ = size cluster i
- N_{total} = sum of all cluster's sizes

In our example, block 7 probability of selection: $\frac{3 \times 20}{315} = 0.19$

The weight for Block 7 will be $\frac{1}{0.19} = 5.2$

PPS: probability of selection

If we select a fixed number of housing units (7), what would happen with the selection probability of house units?

$$f_{house} = \frac{7}{N_{cluster_i}}$$

$$f_{total} = f_{block} * f_{house}$$

- Block 2: $\frac{3 \times 100}{315} \times \frac{7}{100} = 0.066$
- Block 3: $\frac{3 \times 50}{315} \times \frac{7}{50} = 0.066$
- Block 7: $\frac{3 \times 20}{315} \times \frac{7}{20} = 0.066$

The weight or inverse of the probability of selection will be **15.15** for each housing units (same probability of selection for each unit):

$$3 \times 7 \times 15 = 315$$

PPS: Alternative procedures

```
blocks[, pblock := size / sum(size)]
blocks[sample(.N, 3, prob=pblock)]
blocks[, block_prop := 3 * pblock]
```

💻 06-clustering.ipynb

Joint versus marginal distributions

	A1	A2	A3	Total
B1	10	23	4	37
B2	23	56	10	89
B3	32	23	34	89
Total	65	102	48	215

Let's simulate some sample data

These data don't follow the joint distribution of the previous table:

```
va = paste0('A', 1:3)
vb = paste0('B', 1:3)

a = sample(va, 500, replace=TRUE)
b = sample(vb, 500, replace=TRUE)
dat = data.table(a, b)
dat[, total_sim := .N]
nrow(dat)
500
```

Create dataframe with the joint distribution

```
joint = c(10, 23, 4, 23, 56, 10, 32, 23, 34)
la = rep(c('A1', 'A2', 'A3'), 3)
lb = c(rep('B1', 3), rep('B2', 3), rep('B3', 3))
djoint = data.table(a=la,b=lb, joint=joint)
djoint[, total_joint := sum(joint)]

# distribution
   a       prop
1: A1 0.3023256
2: A2 0.4744186
3: A3 0.2232558

   b       prop
1: B1 0.1720930
2: B2 0.4139535
3: B3 0.4139535
```

Weights to emulate joint distribution in our fake sample?

If j are the joint numbers from the previous table, and s are the fake sample we created:

$$w = \frac{\frac{n_{j_{ab}}}{N_j}}{\frac{n_{s_{ab}}}{N_s}}$$

$$w = \frac{n_{j_{ab}}}{N_j} \times \frac{N_s}{n_{s_{ab}}}$$

$$w_{a_1 b_1} = \frac{10}{215} \times \frac{500}{42}$$

$$w_{a_1 b_1} = 0.5537$$

Create data with joint distributions

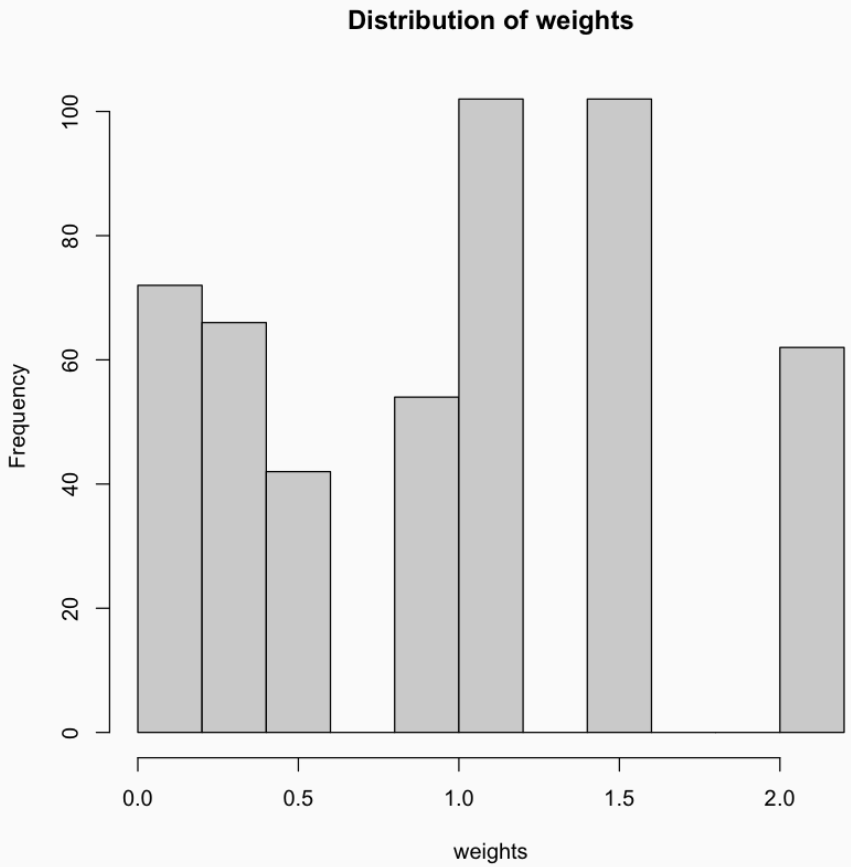
```
dat = merge(dat, djoint, by=c('a', 'b'))
dat[, sim_group := .N, .(a, b)]
dat[, w := joint/total_joint * total_sim/sim_group]

head(dat, 4)
   a  b         w
1: A1 B1 0.5537099
2: A1 B1 0.5537099
3: A1 B1 0.5537099
4: A1 B1 0.5537099
```

Checking distributions

```
wpct(dat$sa)
A1    A2    A3
0.282 0.336 0.382
wpct(dat$a, dat$w)
    A1        A2        A3
0.3023256 0.4744186 0.2232558

wpct(dat$b)
B1    B2    B3
0.332 0.356 0.312
wpct(dat$b, dat$w)
    B1        B2        B3
0.1720930 0.4139535 0.4139535
```

Distribution of weights

What is inflation in the variance of sample estimates attributed to weighting?



DEFF due to weighting

$$DEFF_w \approx \frac{\sum_{i=1}^{n} w_i^2}{\left(\sum_{i=1}^{n} w_i\right)^2} \times n$$

```
(sum(dat$w^2) / (sum(dat$w)) ^ 2) * nrow(dat)
```

```
1.384
```

Raking

What if we only have information on marginal distributions?

	A1	A2	A3	Total
B1	10	23	4	37
B2	23	56	10	89
B3	32	23	34	89
Total	65	102	48	215

Raking ratio estimation, or iterative proportional fitting, is the statistical process of adjusting data sample weights to match desired marginal totals

Raking using R

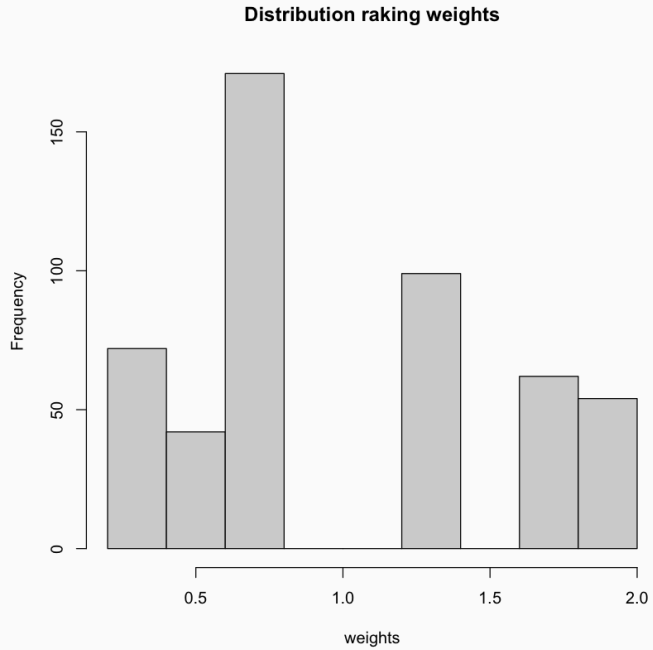
```
remotes::install_github("sdaza/autumn-adjustments")
library(autumn)

# define targets
target = list(
a = c(A1=0.302, A2=0.474, A3=0.223),
b = c(B1=0.172, B2=0.413, B3=0.413)
)
target = normalize(target)
```

Raking using R

```
result = harvest(dat, target)
diagnose_weights(data=result, target=target,
weights=result$weights)
```

variable	level	prop_original	prop_weighted	target
a	A1	0.282	0.30230	0.30230
a	A2	0.336	0.47447	0.47447
a	A3	0.382	0.22322	0.22322
b	B1	0.332	0.17234	0.17234
b	B2	0.356	0.41382	0.41382
b	B3	0.312	0.41382	0.41382

Distribution of weights with raking

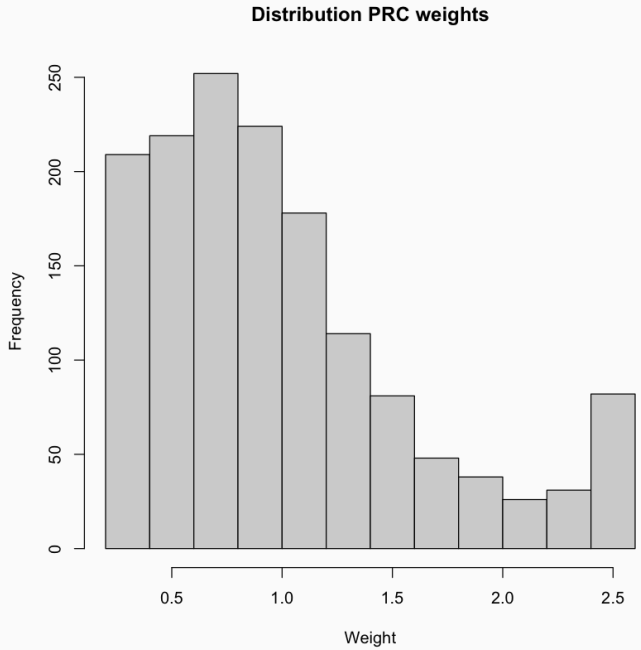



```
design_effect(result$weights)
1.24

effective_sample_size(result$weights)
402.95
```

Pew Research Center (PRC) Dataset

- 2021 Core Trends Survey
- Telephone survey (300 landlines, 1202 cellphones)
- Random-digit-dial (RDD)
- 18 years of age or older
- Weights: iterative adjustments (raking) by gender, age, education, race, Hispanic origin/nativity and region using ACS estimates

Distribution of PRC weights



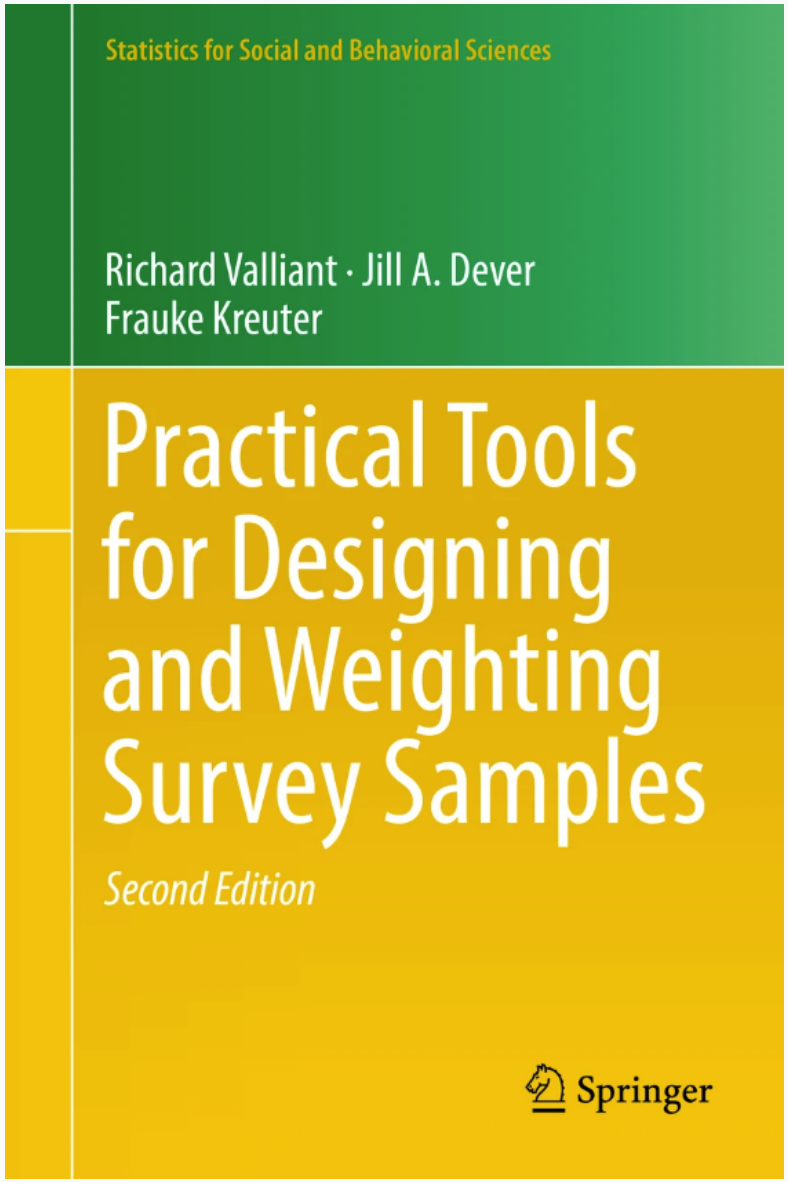

```
design_effect(dat$weights)
1.34

effective_sample_size(dat$weights)
1118.19
```

TikTok users?

```
library(survey)

d = svydesign(ids=~0, strata=~state+sample,
data=dat, weights=~weight)

(s = svymean(~tiktok, d))
         mean     SE
tiktok 0.2096 0.0128

confint(s)
         2.5 %    97.5 %
tiktok 0.1844557 0.2347533
```

💻 08-raking-pew-research.ipynb

Check out this book!



USTS

The U.S. Trans Survey (USTS) is the largest survey of trans people, by trans people, in the United States. The USTS documents the lives and experiences of trans and non-binary people ages 16+ in the U.S. and U.S. territories.

https://www.ustranssurvey.org/ Info on the design 2015 here

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

sebastian.daza@gmail.com