Survey Research Methodology I

Computational Social Science UC3M
Fall 2022-2023

Sebastian Daza

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 + plain R code
- I recommend you use Visual Studio Code
  - Integrated development environment (IDE)
- R
  - `data.table`
  - More info here

Meaning of Emojis

🤔 = Expecting your participation and discussion

💻 = We will do some live coding

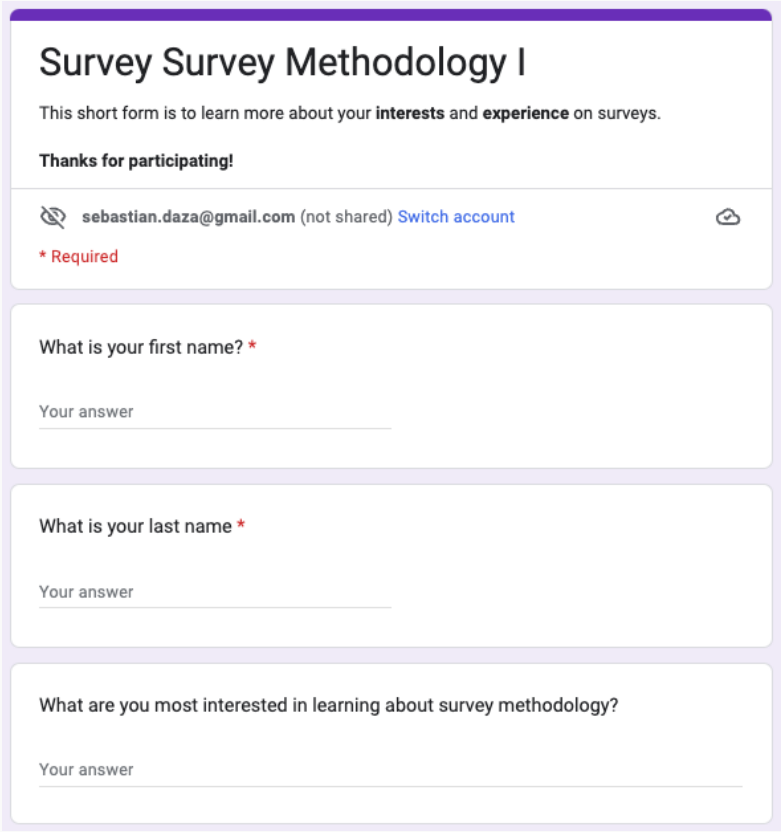
🤓 = Extra knowledge

🥷 = Ninja level

Students’ Survey

Response rate

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

Most interested in learning about survey methodology

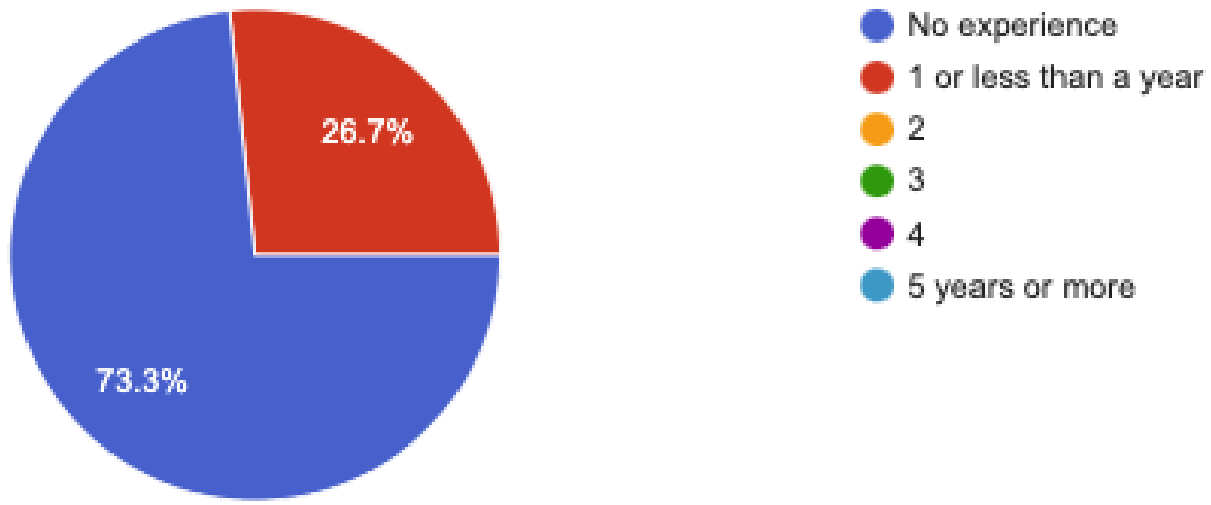
Analyzing the results of it

I want to learn to be very critical of data, both from social media and from surveys.

Experience

How many years of experience do you have in conducting or designing social surveys?

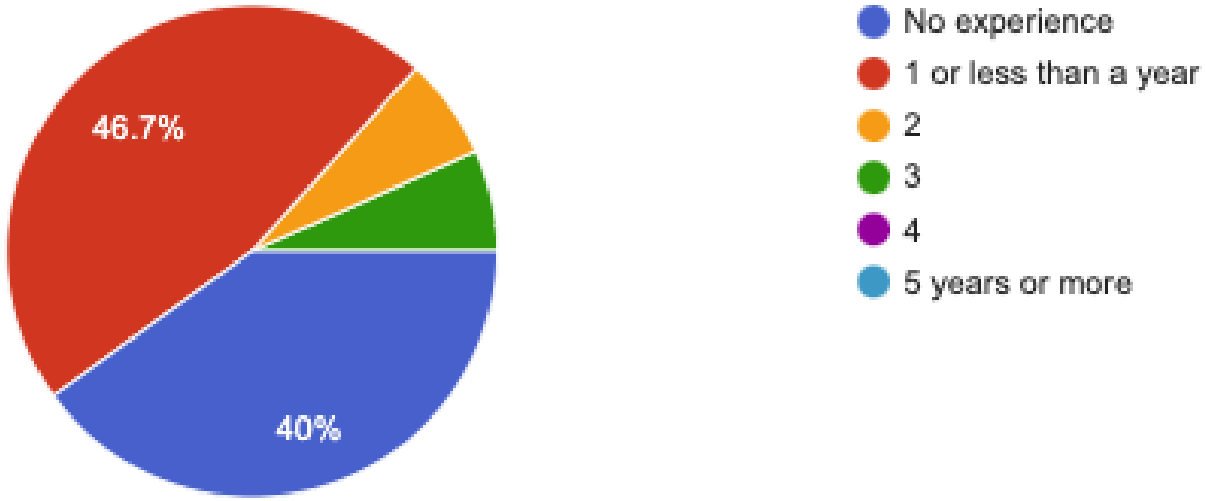
15 responses



Experience

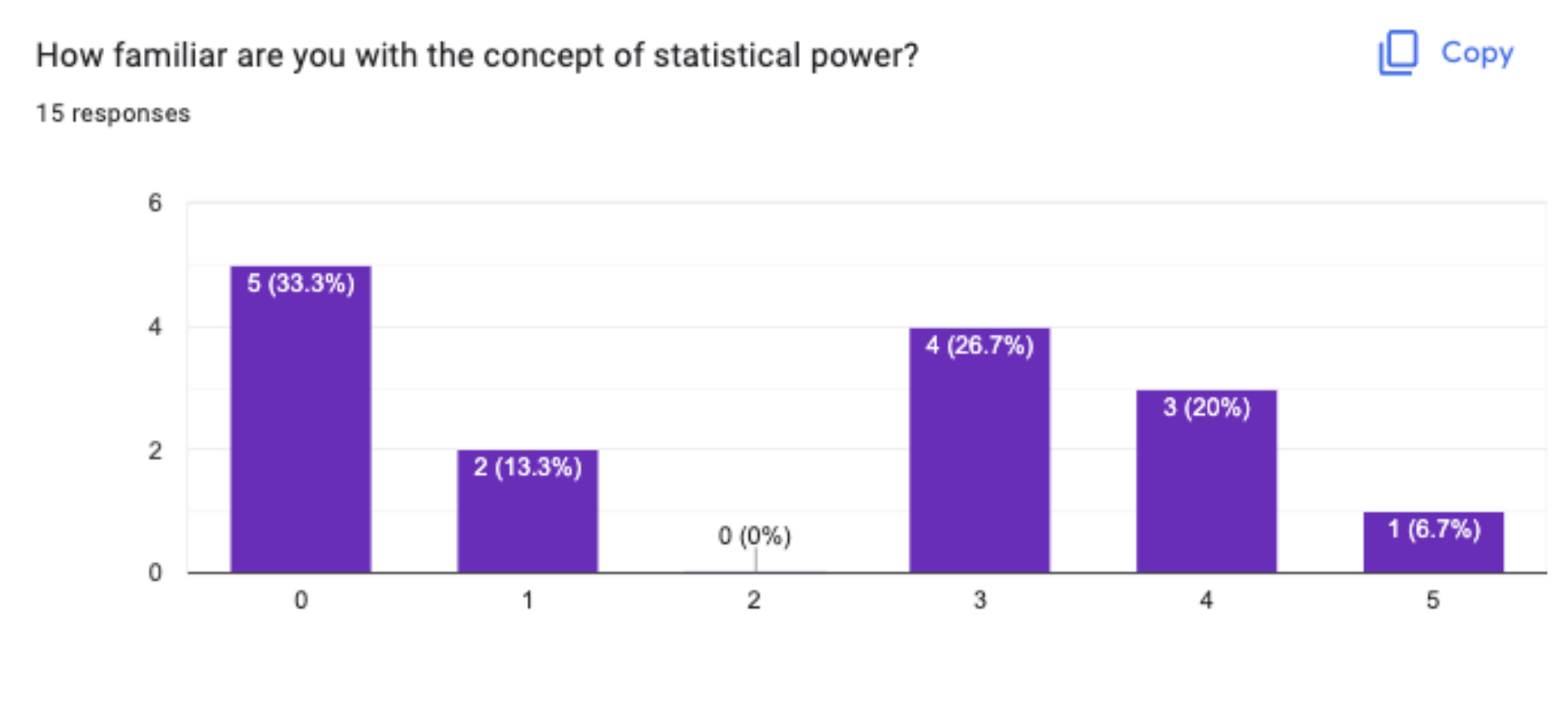
How many years of experience do you have in analyzing social surveys?

15 responses

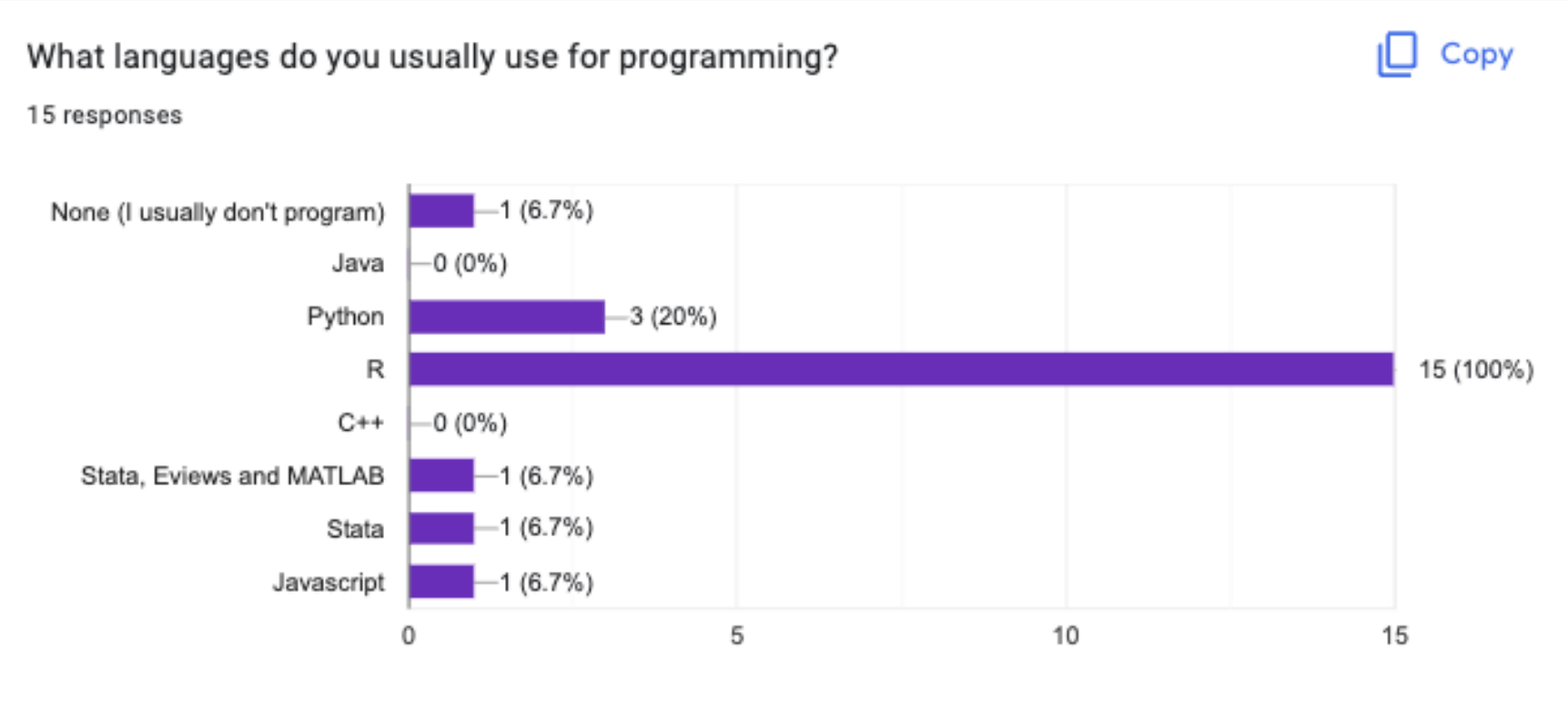


Statistical Power

Mean = 2.067



Programming

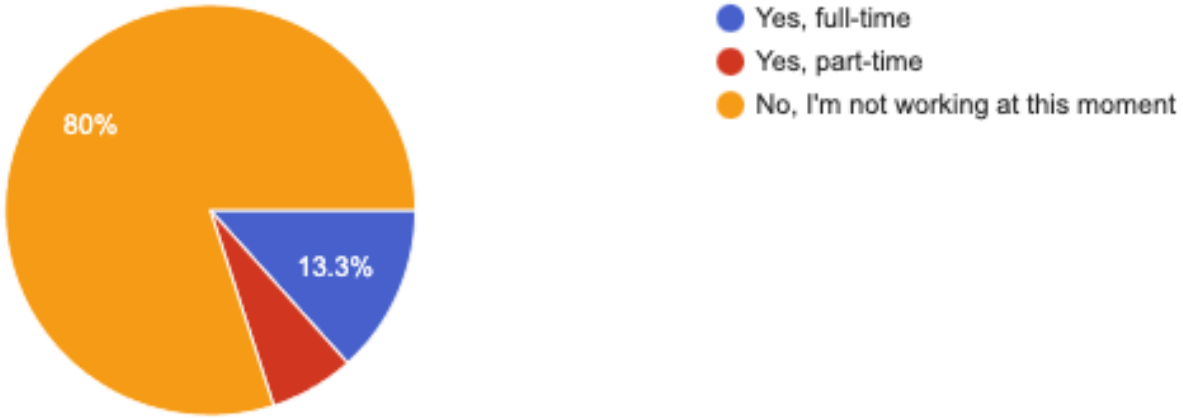


Working

Are you currently working?

15 responses

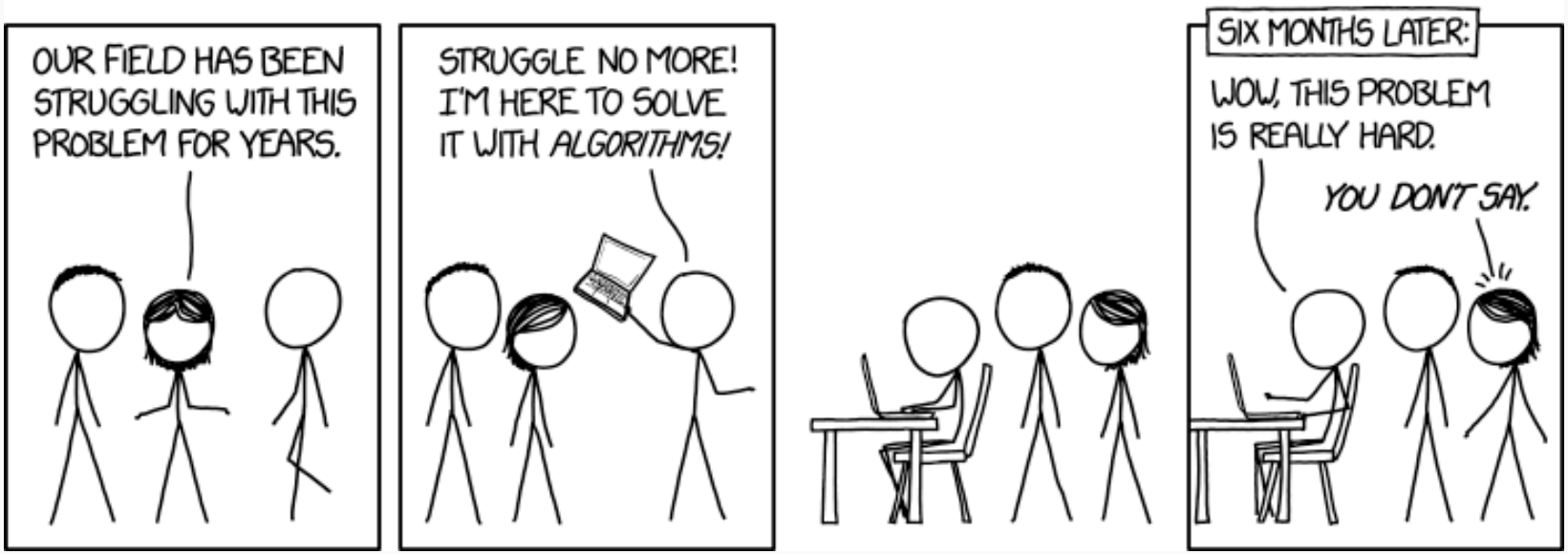
Copy



What is the position you would most like to be working?

Disclaimer

Complexity of social problems



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

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

Many decisions

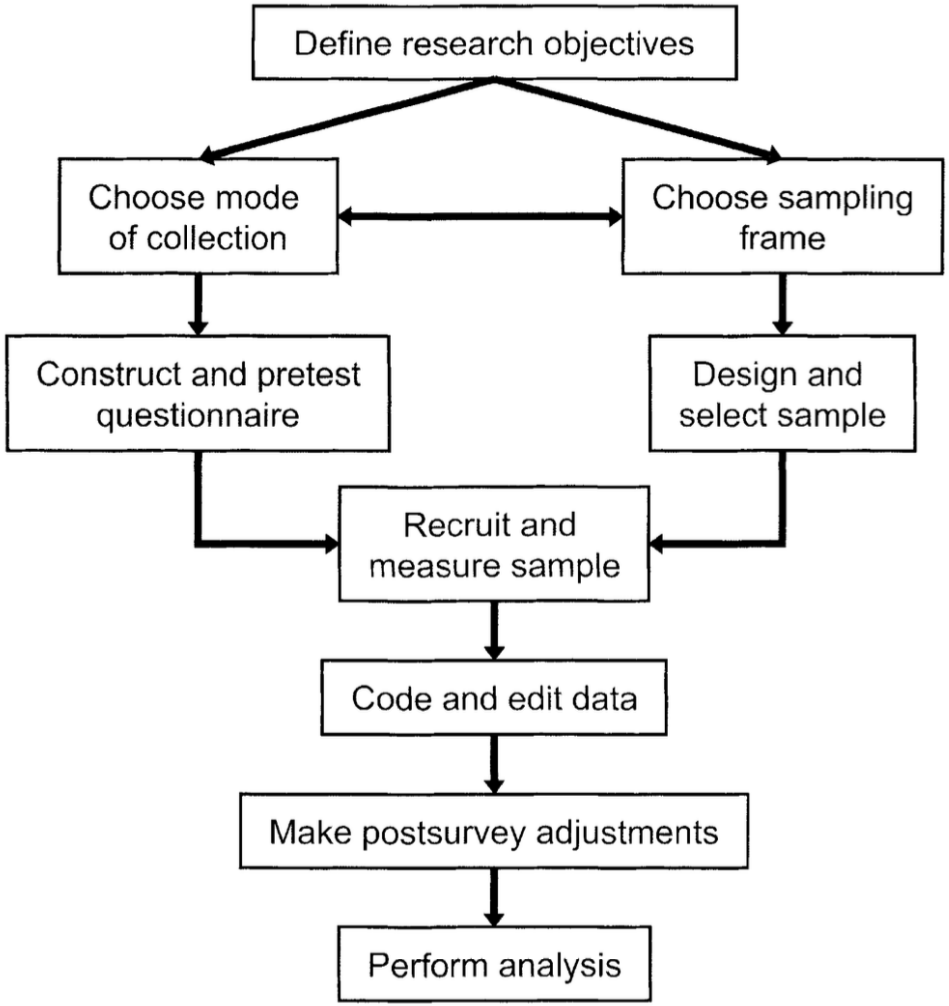


Figure 2.4 A survey from a process perspective.

Many decisions

- To make a large set of decisions about several individual features of a survey
- Those decisions have the potential to affect the quality of estimates that come from surveys
- Surveys are conducted in the uncontrolled settings of the real world and can be affected by those settings!

Key challenges

- How to best use the available resources
- How to balance the investments in each of the components of a survey to maximize the value of the data that will results
- Rather than focusing on just one or few of the elements of a survey, all the elements are considered as a whole $\sim$ total survey error approach
- Many trade-offs!

Some history

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

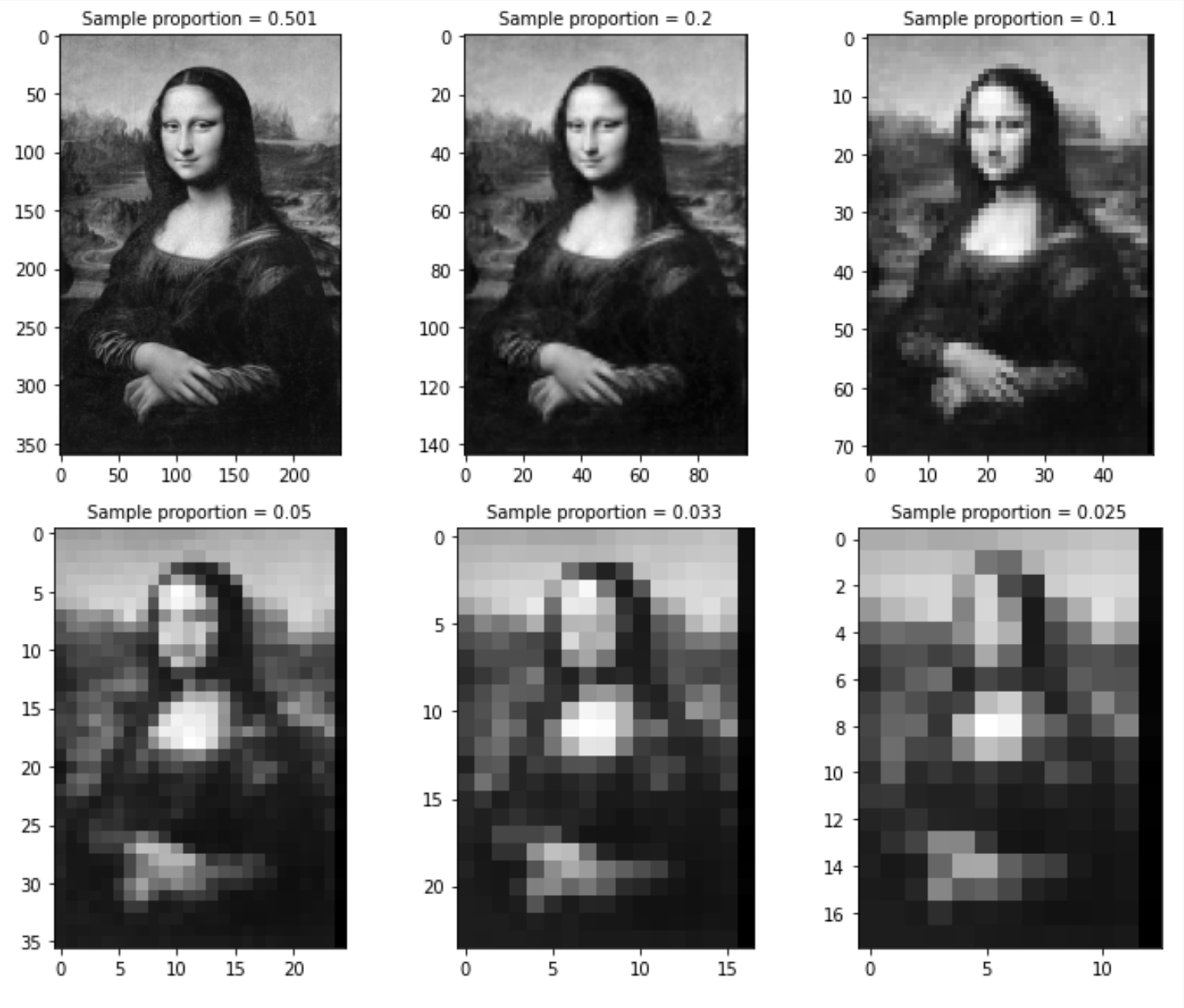
- Sampling: from samples $\rightarrow$ unbiased estimates
- Inference: statistics, margins of errors
- Measurement: the art of asking questions
- Analysis: multivariate, complex survey designs

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 (∼ 5000 respondents), correctly predicted a Franklin Delano Roosevelt 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 Harry S Truman against popular New York governor Thomas E. 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)
- 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

All surveys are not created equal

Total survey error

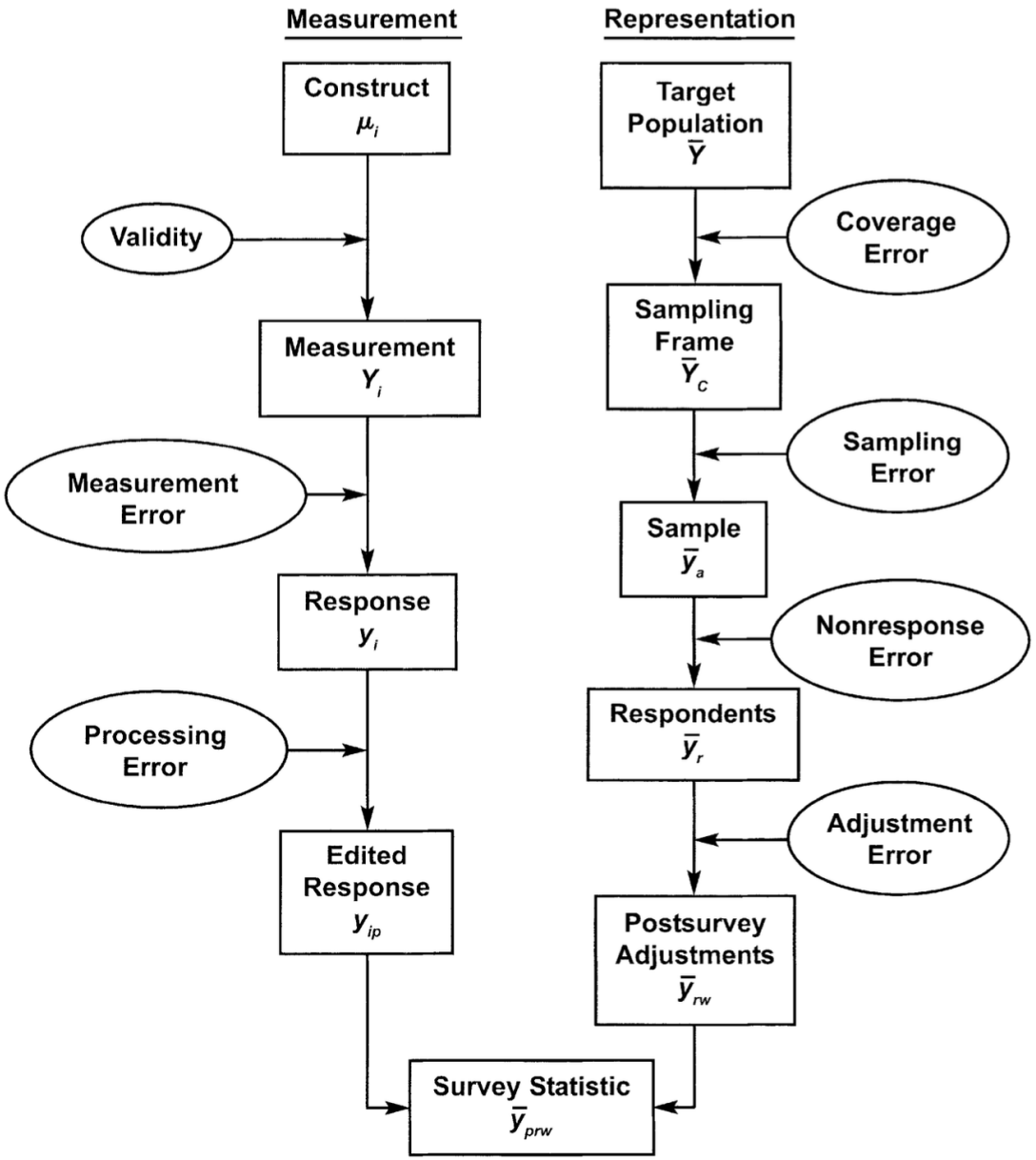
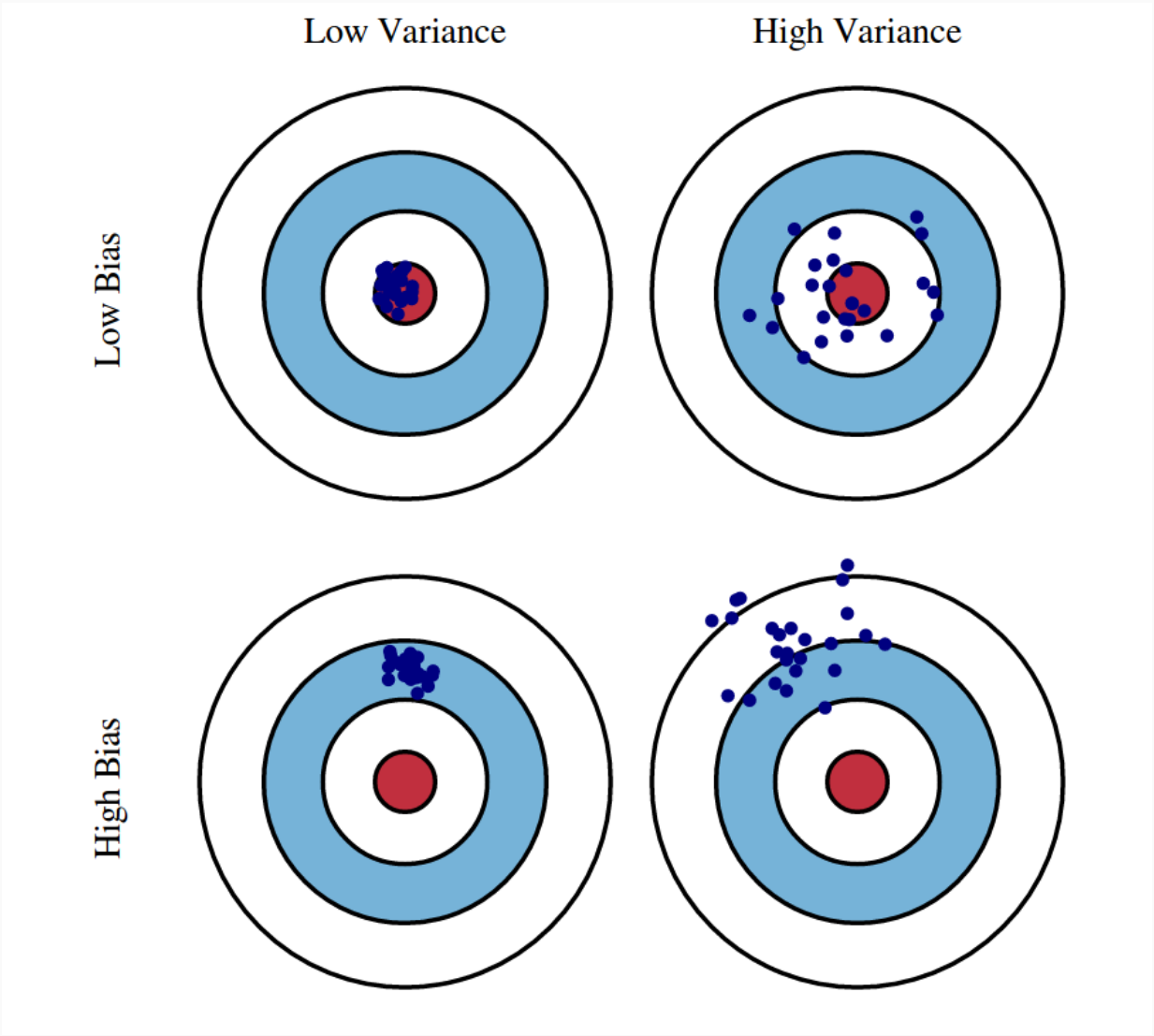
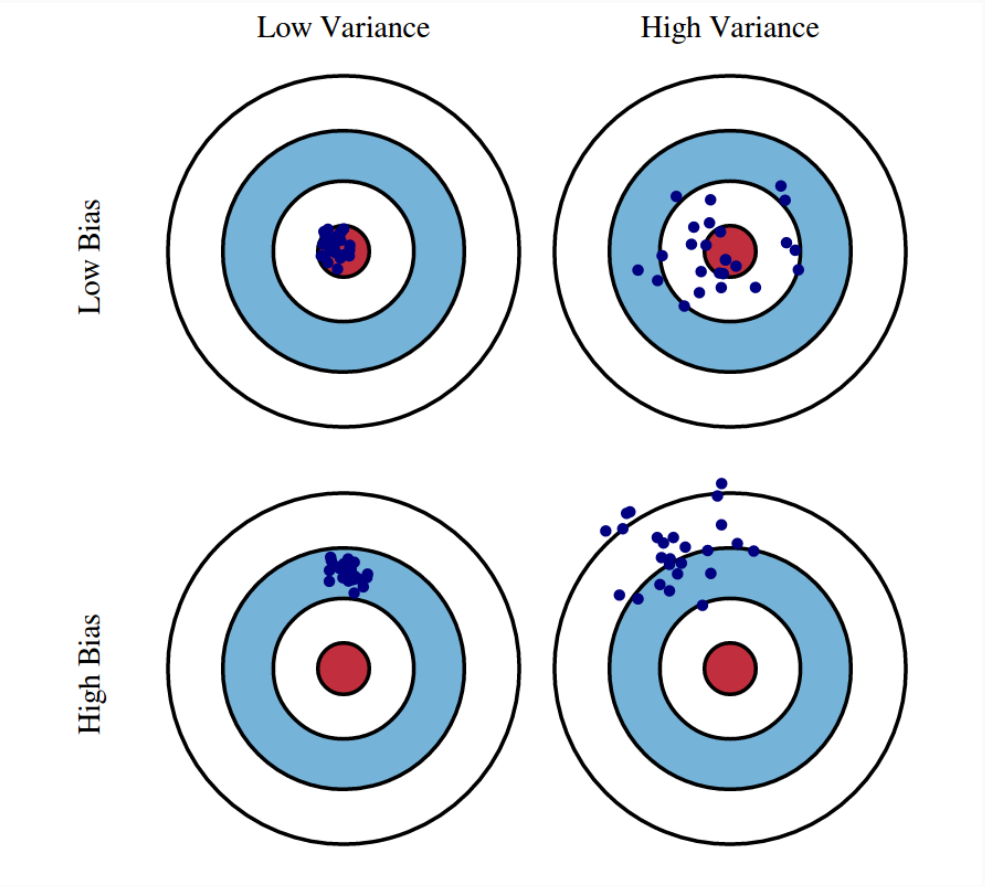


Figure 2.5 Survey life cycle from a quality perspective.

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

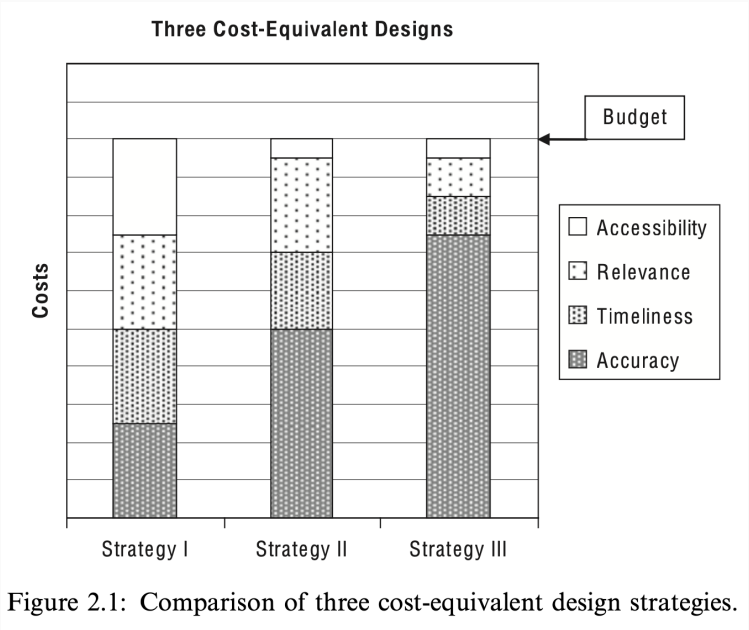


Figure 2.1: Comparison of three cost-equivalent design strategies.

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

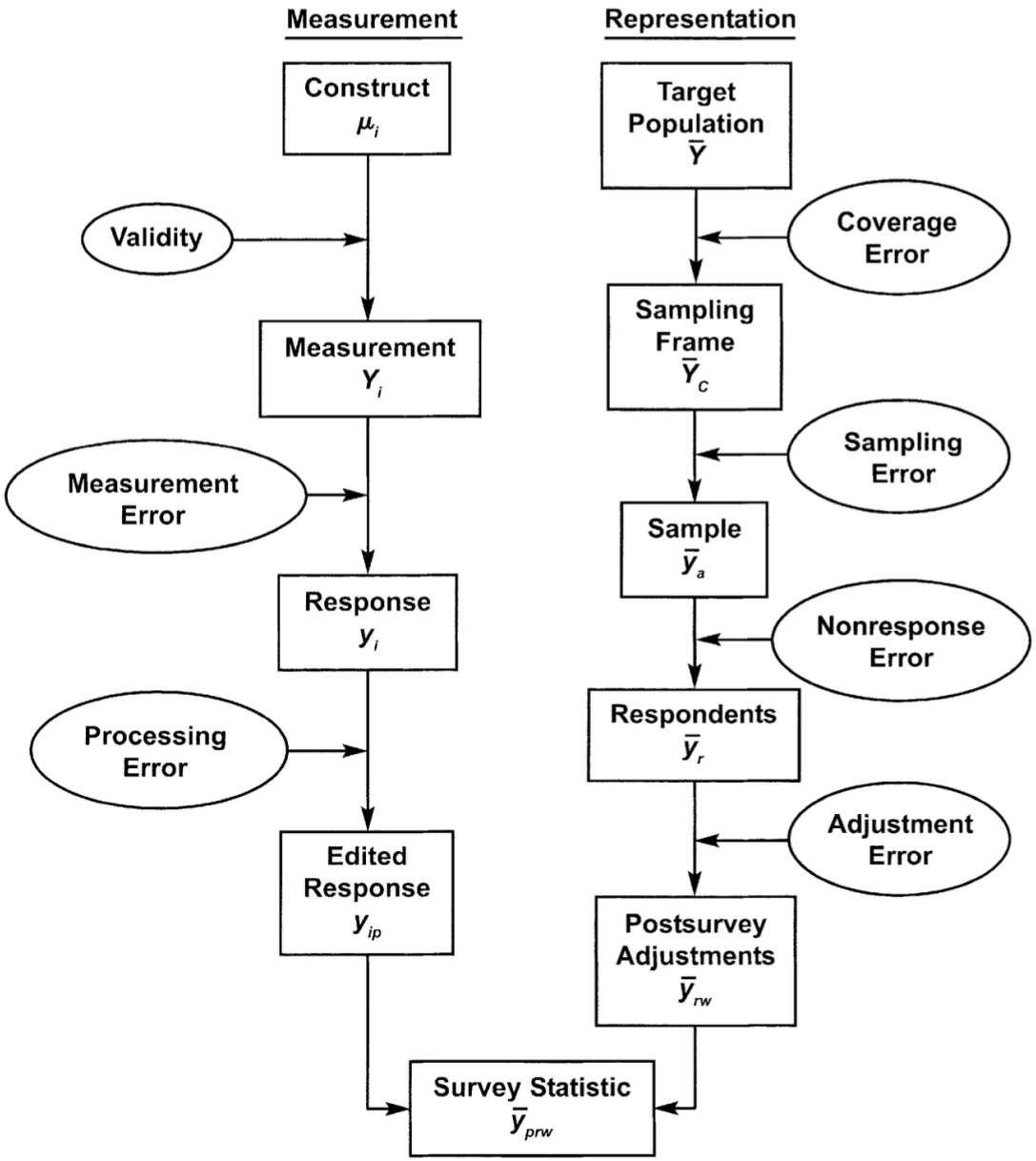
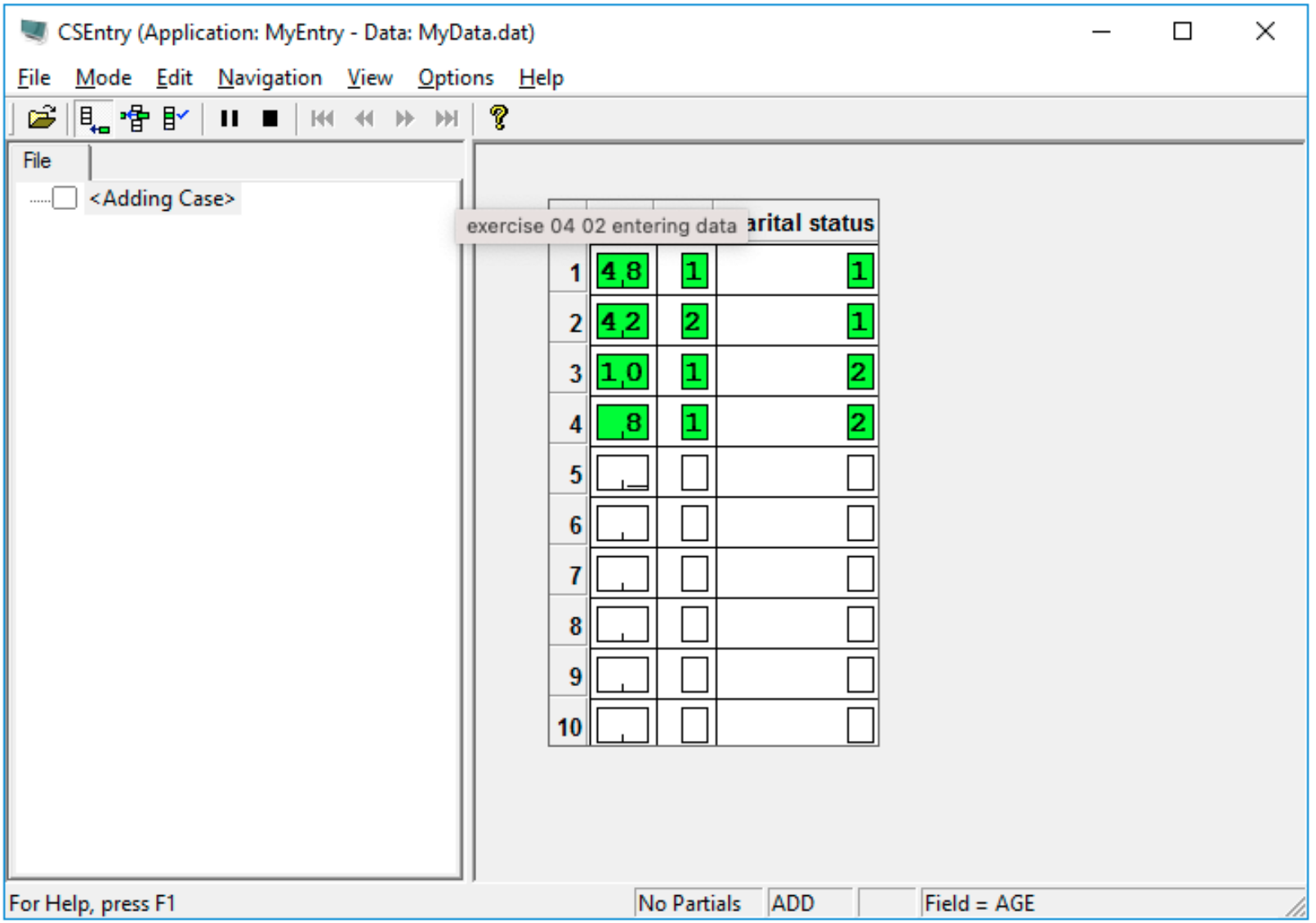


Figure 2.5 Survey life cycle from a quality perspective.

Processing error

Census and Survey Processing System (CSPro)



Coverage error 🤔

Systematic vs random distortion?

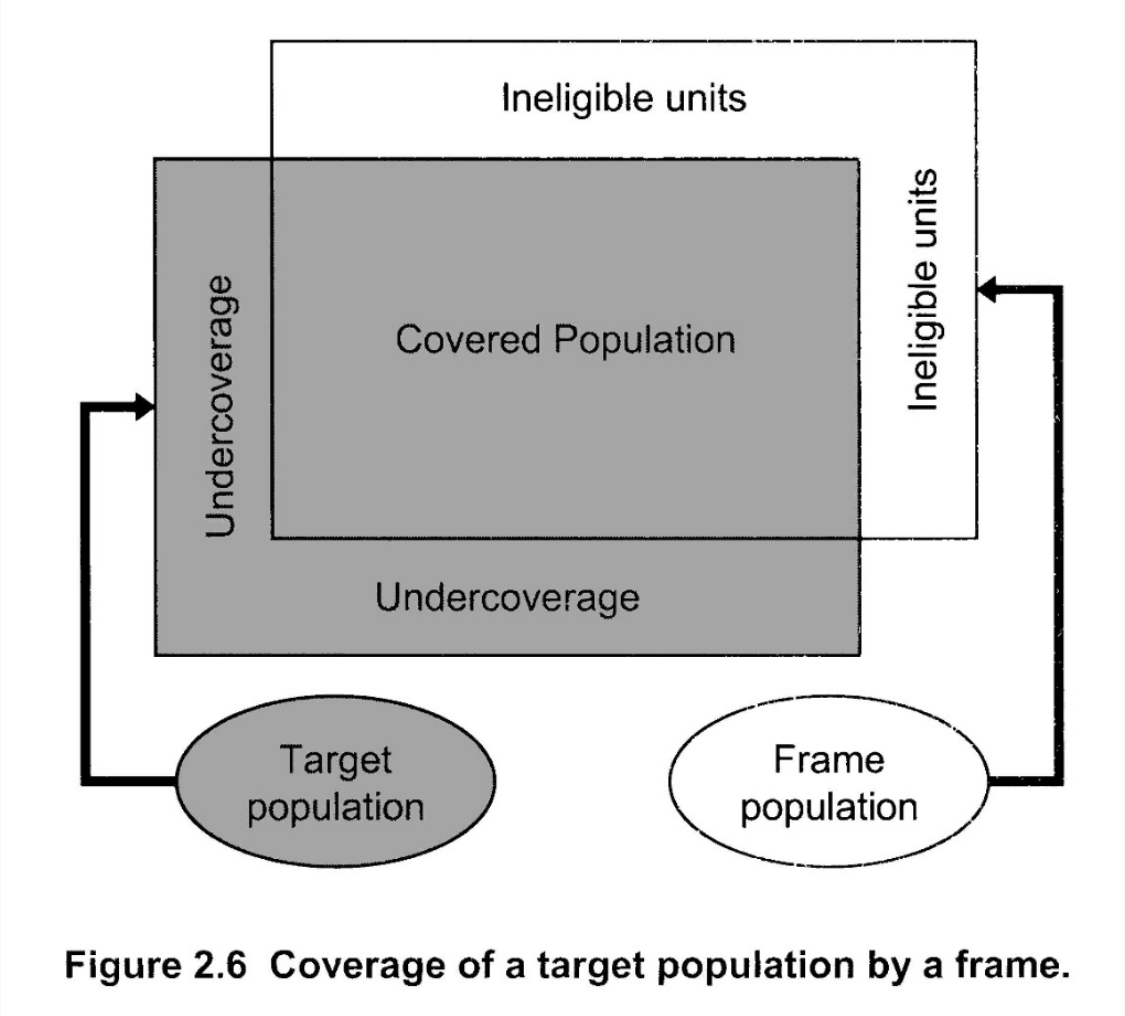


Figure 2.6 Coverage of a target population by a frame.

Sampling error 🤔

Systematic vs random distortion?

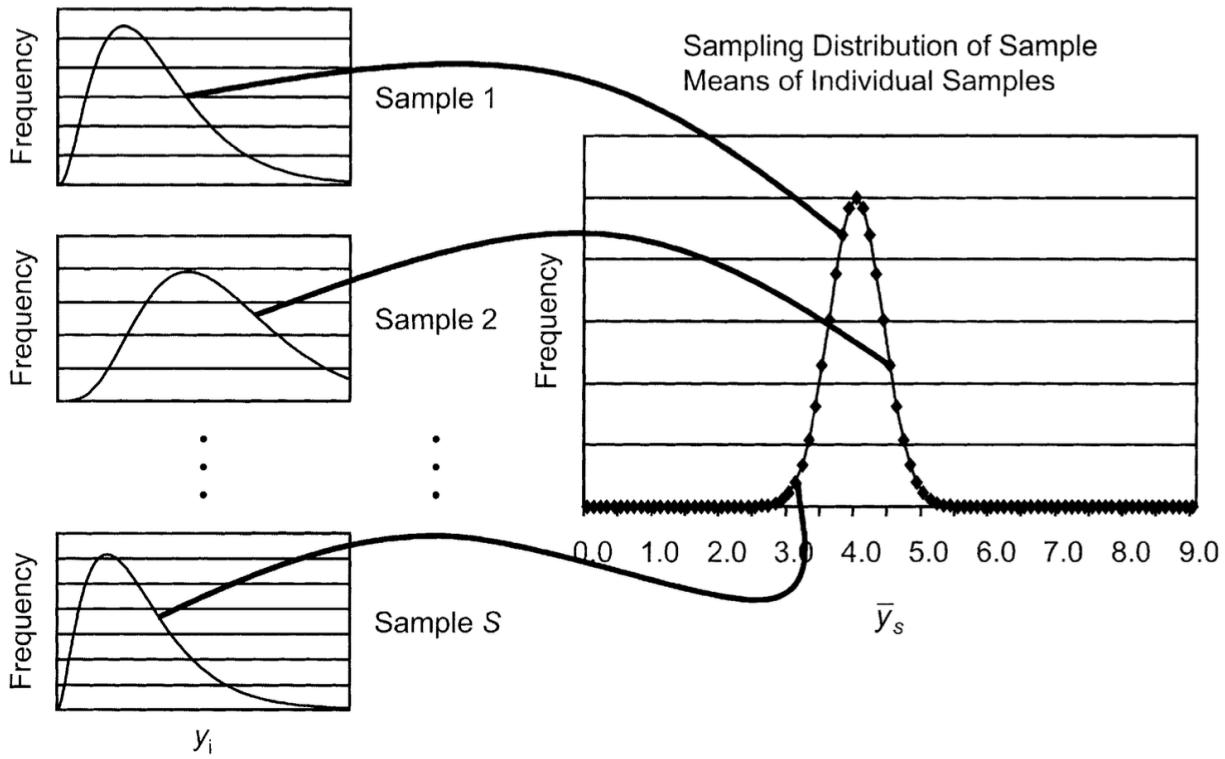


Figure 2.7 Samples and the sampling distribution of the mean.

💻 Let's move to R...

Non-response error 🤔

Systematic vs random distortion?

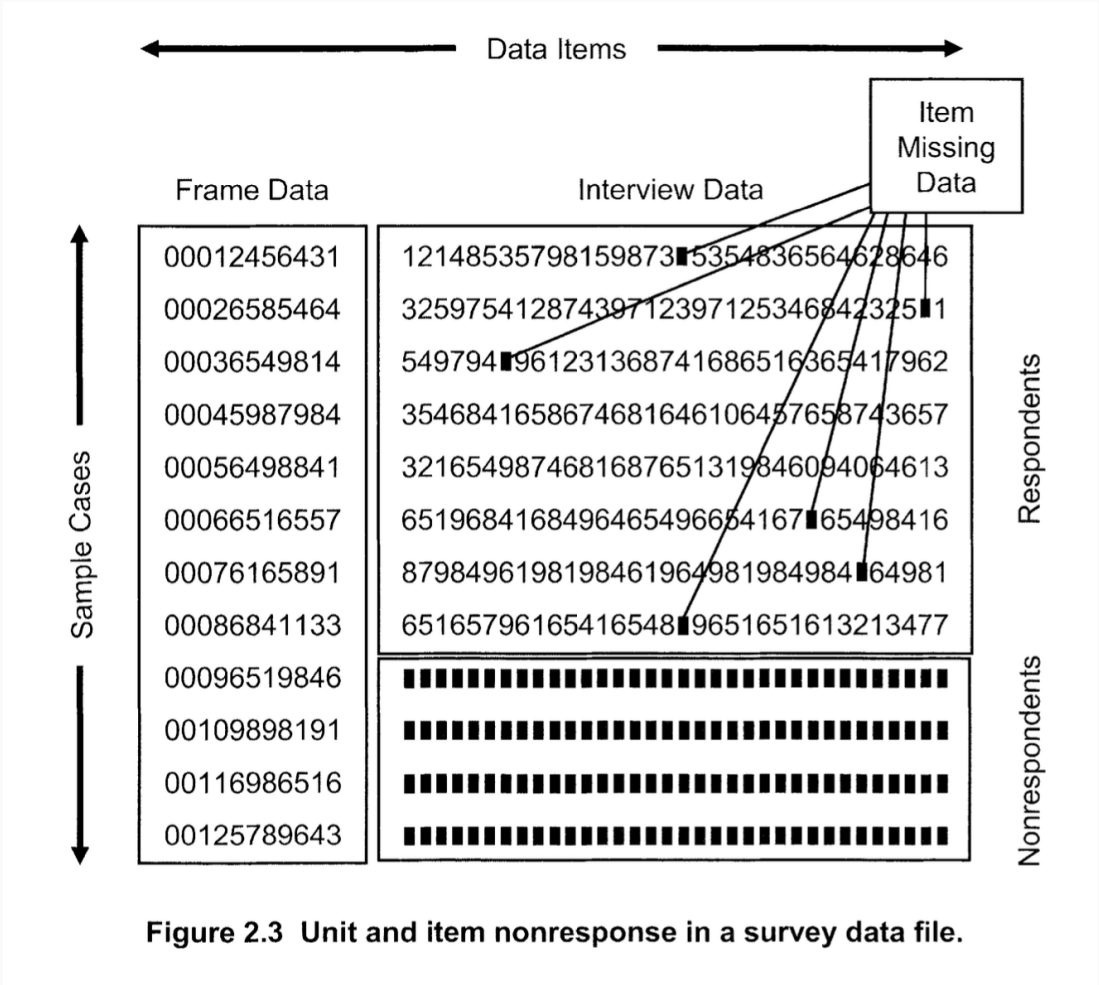


Figure 2.3 Unit and item nonresponse in a survey data file.

Steps and errors

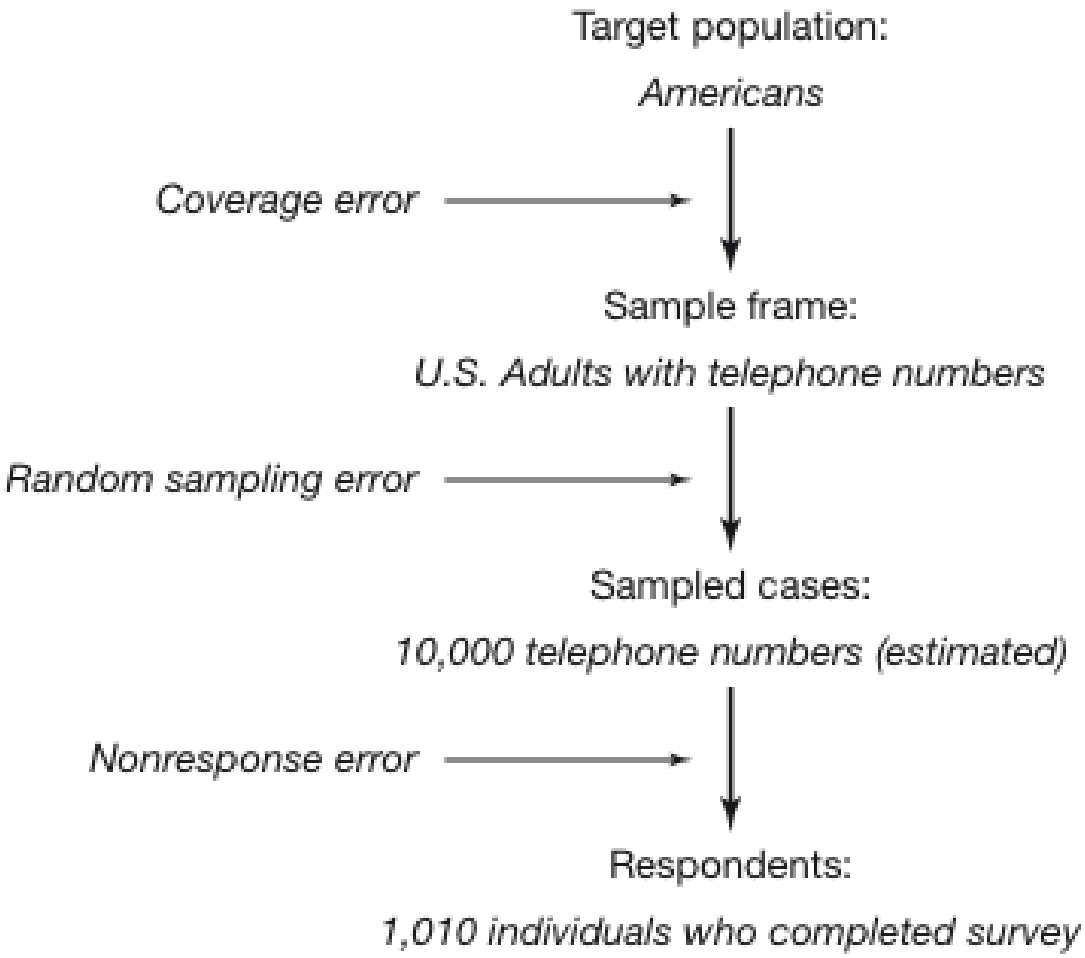


Figure 1.2 Steps in Survey Process

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-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

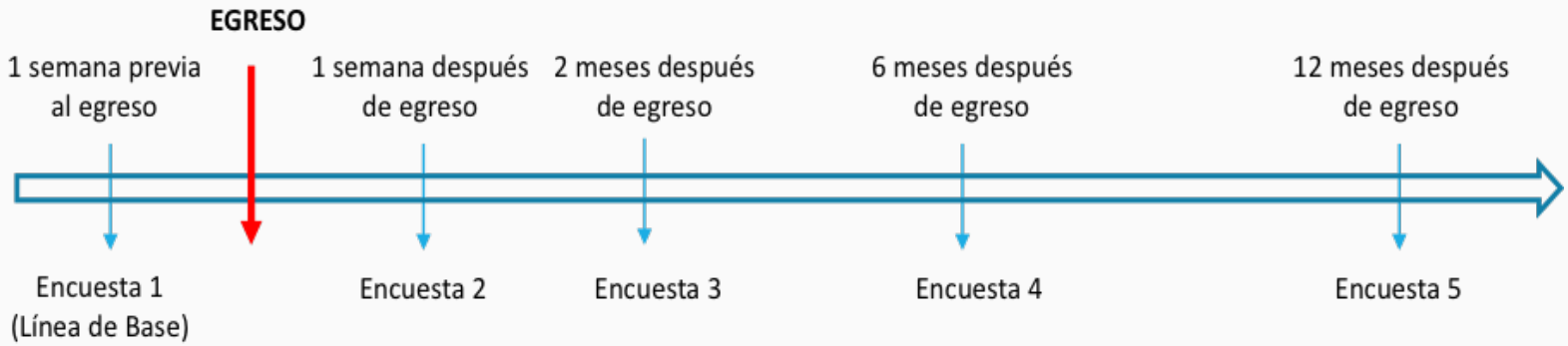
Chilean Women Reentry Study

Reentry Study

- Longitudinal study that followed a cohort of 225 women released from prison in Santiago, Chile, between September 2016 and March 2017
- The target population was Chilean women who served custodial sentences of at least 30 days and were released on parole or after completing their entire sentence. About 80% of those released under these conditions participated in the study

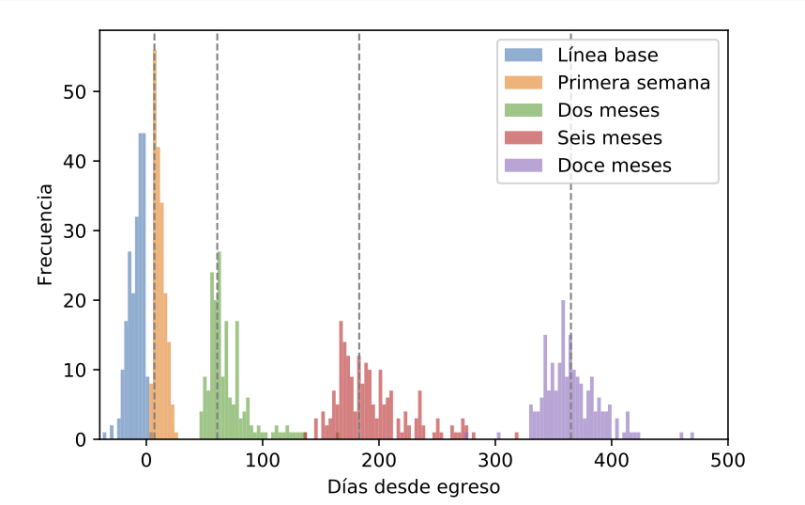
Reentry Study

- The study included five interviews over a year. The first one was conducted about two weeks before prison release
- At each interview, a face-to-face questionnaire was applied, consisting mainly of closed-ended survey questions covering employment, housing, relationships, and offending.



Reentry Study

- The questionnaires also include the use of life event calendars to capture within-waves information in some key life domains
- The application lasted between 1 and 2 hours and was conducted by trained interviewers



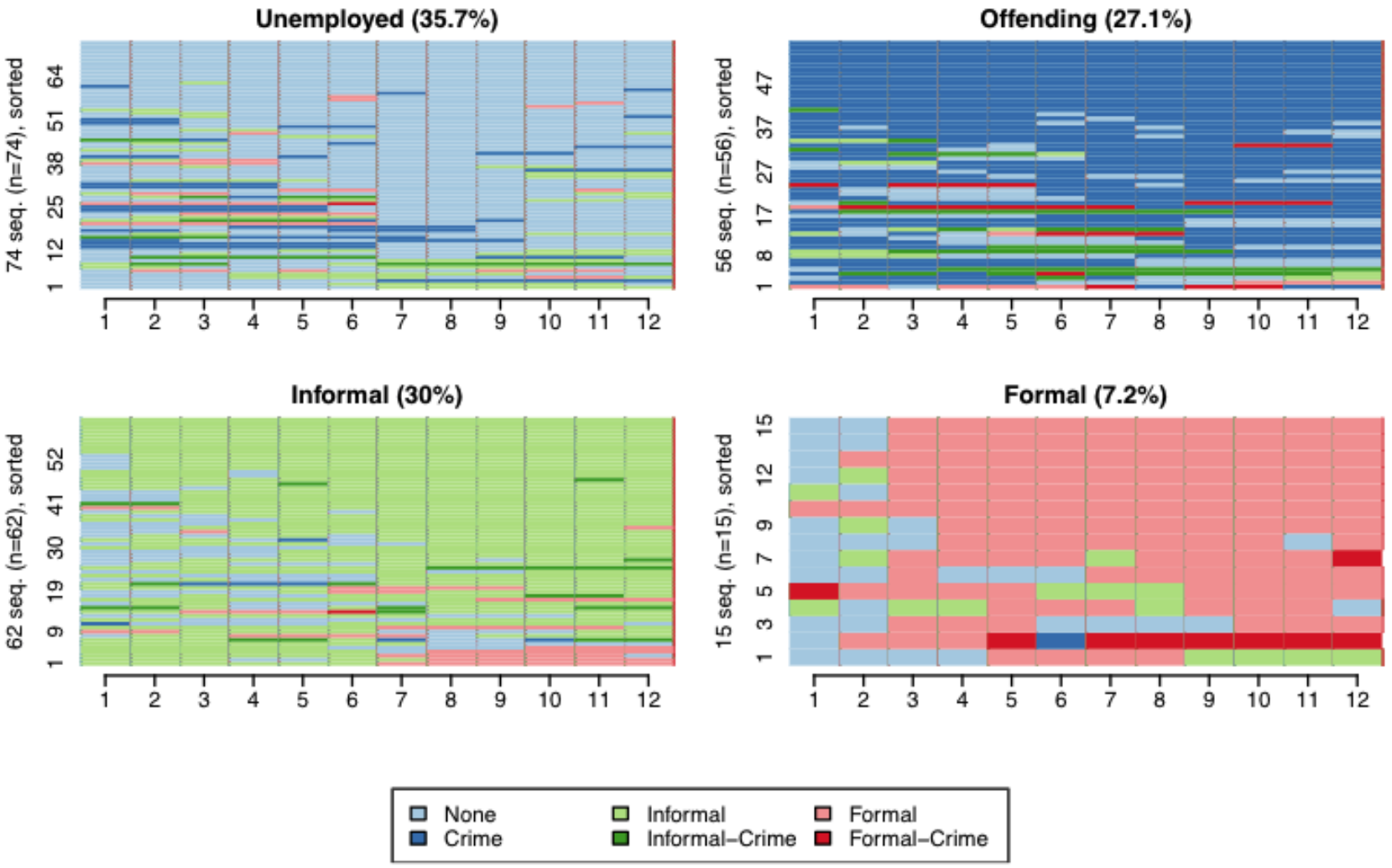
Let's look at a questionnaire

Reentry Study Response Rates

	Baseline	Week 1	Month 2	Month 6	Month 12
Interviews	225	181	177	197	200
No contact	-	26	31	19	21
Contacted	-	18	17	9	4
Response rate (%)	-	80,4	78,7	87,6	88,9

Work trajectories across women 🤓

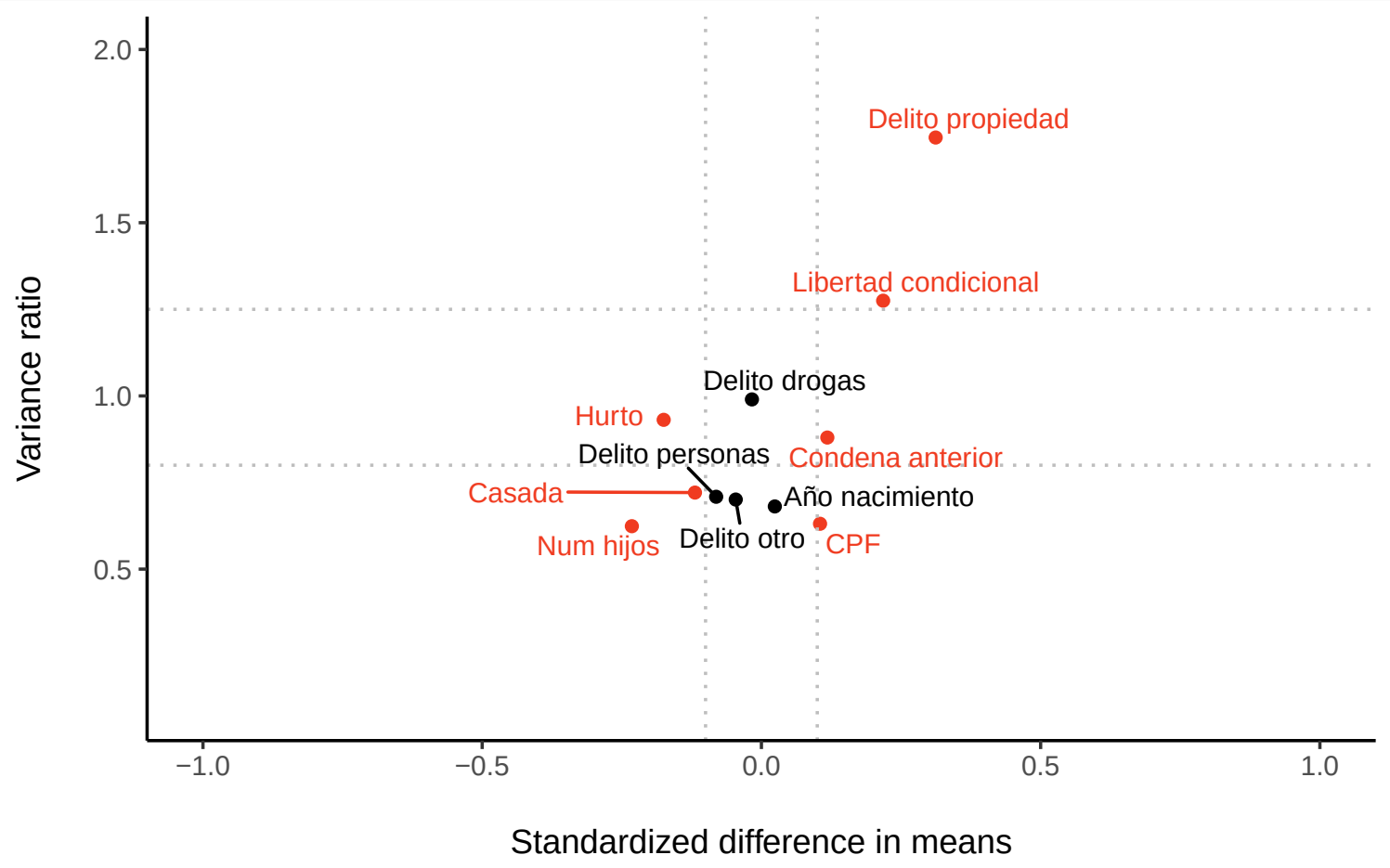
Figure S2: Sequences job-crime categories of women inmates during the first 12 months following their release by four employment-crime clusters (N = 207)



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)



Hypothetical decisions

Error sources from design decisions 🤔

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*

Error sources from design decisions 🤔

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*

Error sources from design decisions 🤔

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*

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? 🤔

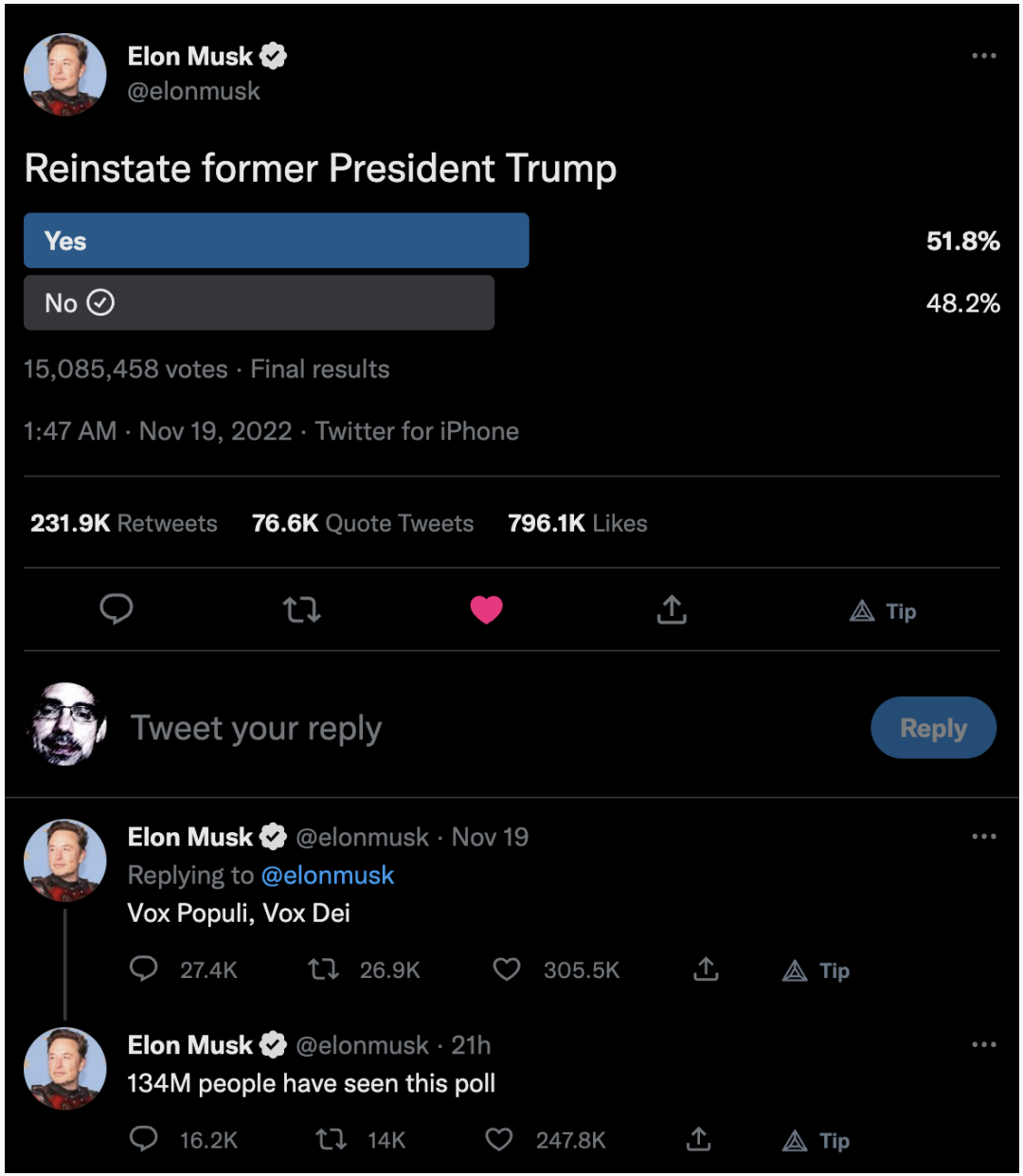
706 902 17.8K

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


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 Error

- Coverage error
- Query error
- Interpretation error

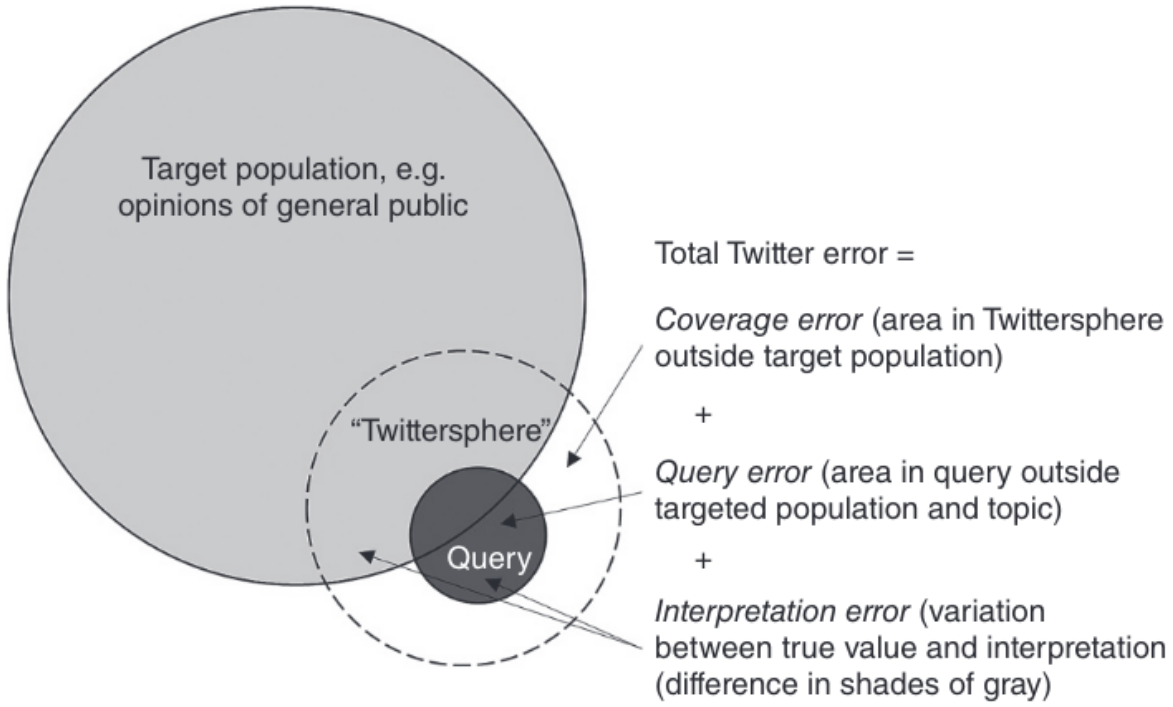


Figure 2.1 Theoretical spaces of Twitter data error.

An example using Twitter data

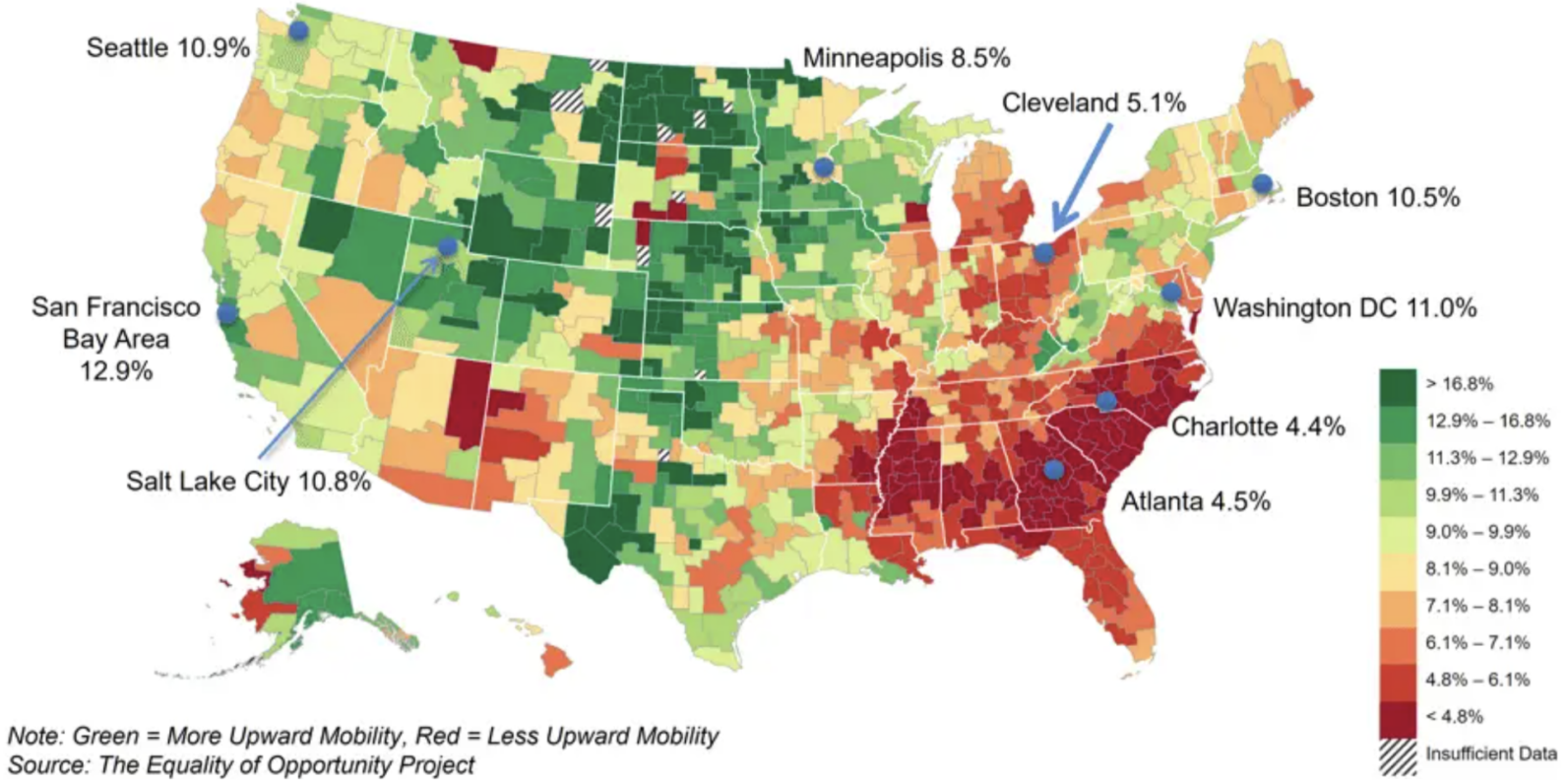
https://sdaza-capstone.herokuapp.com/

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

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

Measurement

Total survey error

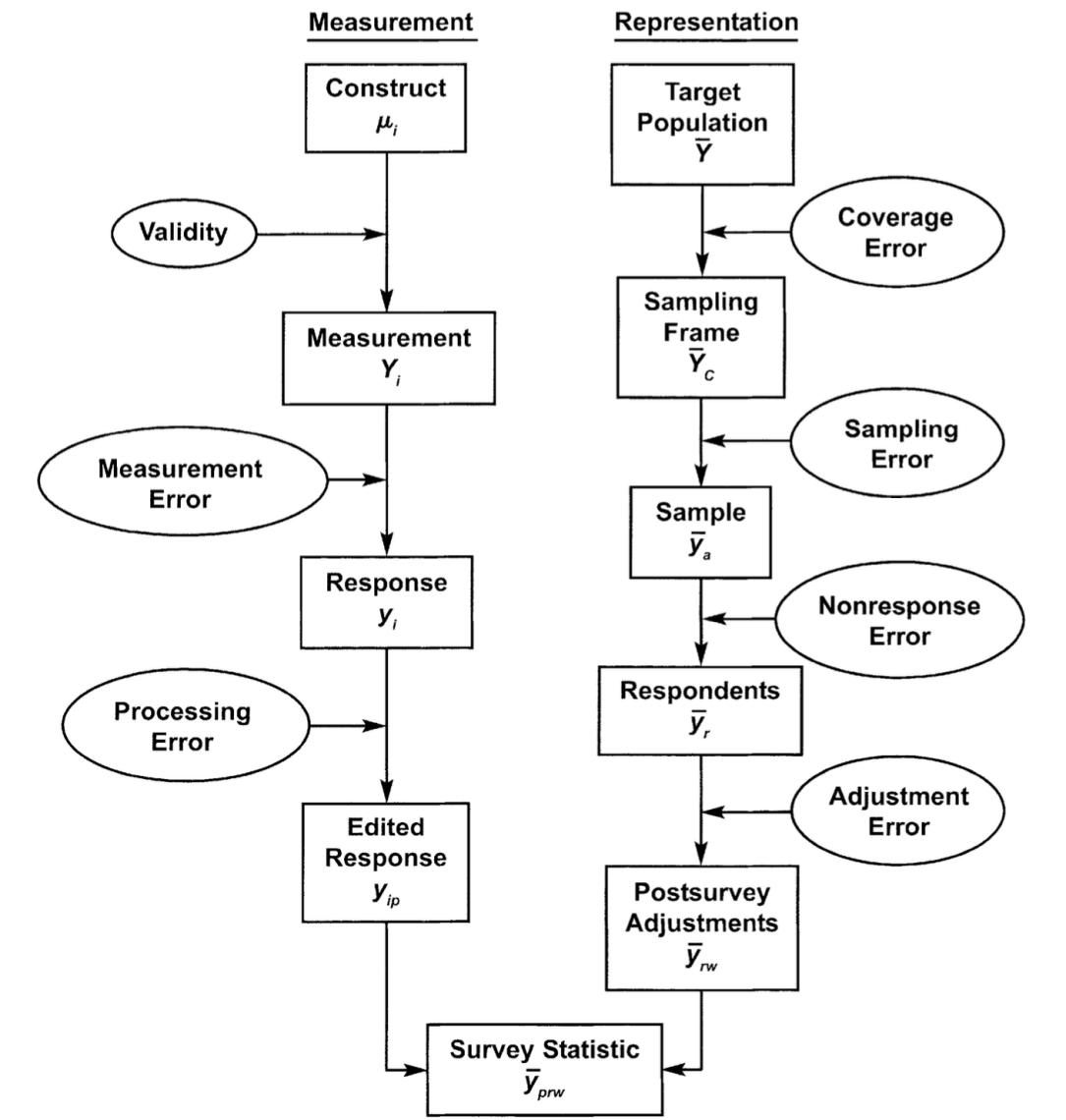


Figure 2.5 Survey life cycle from a quality perspective.

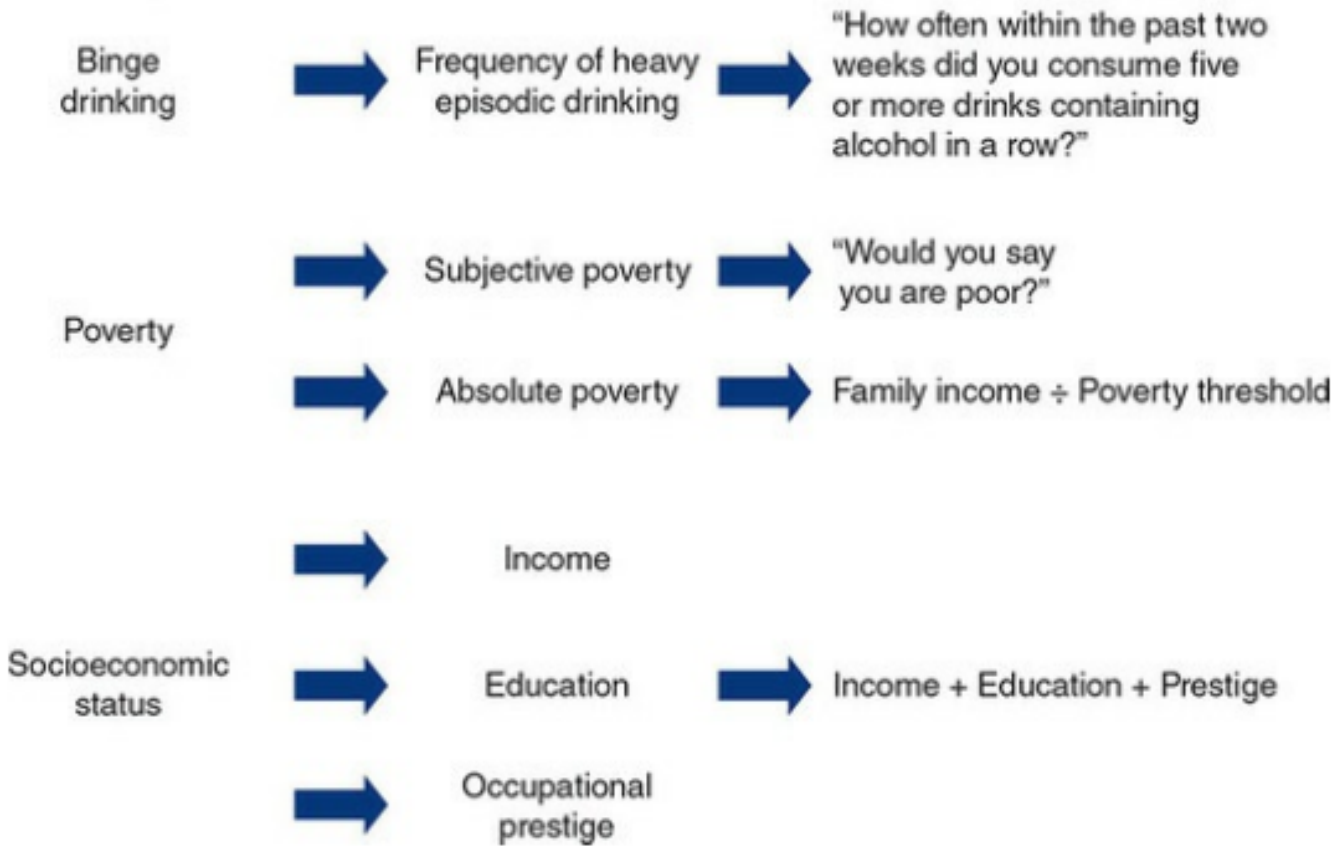
Conceptualization

- What do the main concepts mean in this research?
- How are the main concepts measured?

The process of specifying what we mean by a term

Operationalization

Exhibit 4.1 Concepts, Variables, and Indicators: Operationalizing Concepts.



From indicators to indexes or scales

Exhibit 4.2 Examples of Indexes: Short Form of the Center for Epidemiologic Studies Depression Index (CES-D) and "Negative Outlook" Index

CES-D Index			
At any time during the past week . . . (Circle one response on each line)	**Never**	**Some of the Time**	**Most of the Time**
a. Was your appetite so poor that you did not feel like eating?	1	2	3
b. Did you feel so tired and worn out that you could not enjoy anything?	1	2	3
c. Did you feel depressed?	1	2	3
d. Did you feel unhappy about the way your life is going?	1	2	3
e. Did you feel discouraged and worried about your future?	1	2	3
f. Did you feel lonely?	1	2	3
Negative Outlook Index			
How often was each of these things true during the past week? (Circle one response on each line)	**A Lot, Most, or All of the Time**	**Sometimes**	**Never or Rarely**
a. You felt that you were just as good as other people.	0	1	2
b. You felt hopeful about the future.	0	1	2
c. You were happy.	0	1	2
d. You enjoyed life.	0	1	2

Big five inventory (BFI) ~ personality traits 🤔

The Big Five Factors are (chart recreated from John & Srivastava, 1999):

Big Five Dimensions	Facet (and correlated trait adjective)
Extraversion vs. introversion	Gregariousness (sociable) Assertiveness (forceful) Activity (energetic) Excitement-seeking (adventurous) Positive emotions (enthusiastic) Warmth (outgoing)
Agreeableness vs. antagonism	Trust (forgiving) Straightforwardness (not demanding) Altruism (warm) Compliance (not stubborn) Modesty (not show-off) Tender-mindedness (sympathetic)
Conscientiousness vs. lack of direction	Competence (efficient) Order (organized) Dutifulness (not careless) Achievement striving (thorough) Self-discipline (not lazy) Deliberation (not impulsive)
Neuroticism vs. emotional stability	Anxiety (tense) Angry hostility (irritable) Depression (not contented) Self-consciousness (shy) Impulsiveness (moody) Vulnerability (not self-confident)
Openness vs. closedness to experience	Ideas (curious) Fantasy (imaginative) Aesthetics (artistic) Actions (wide interests) Feelings (excitable) Values (unconventional)

Which of these dimensions is the most important for job performance?

Big five inventory (BFI) ∼ personality traits

Which of these dimensions is the most important for job performance?

> *According to Essentials of Organizational Behavior (14th Edition), the big five personality dimension that has the biggest influence on job performance is conscientiousness.*

Self-assessment

Disagree strongly 1	Disagree a little 2	Neither agree nor disagree 3	Agree a little 4	Agree Strongly 5

I see Myself as Someone Who...

What are we measuring?

Conscientiousness

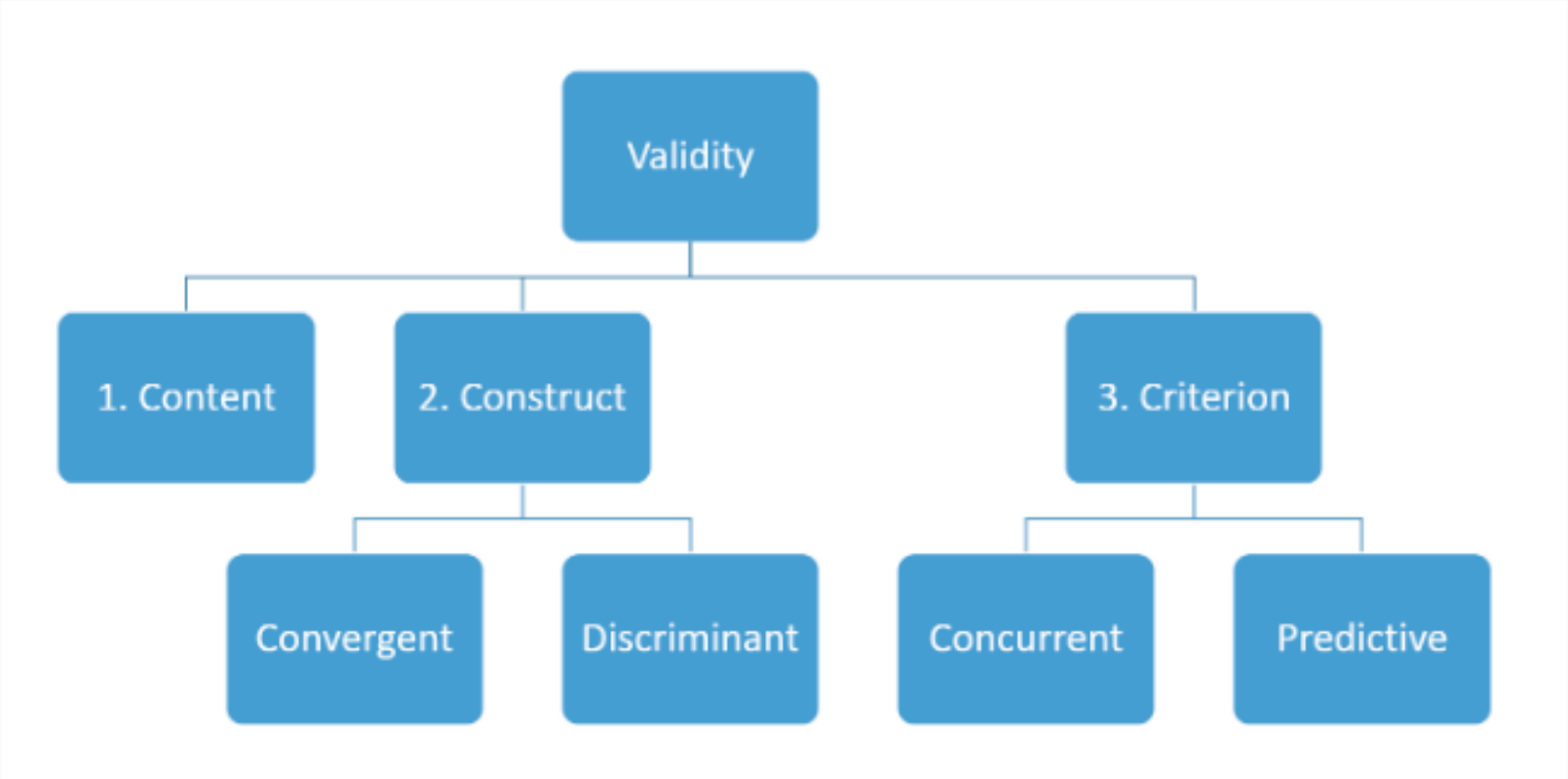
- *Is easily distracted*
- *Makes plans and follows through with them*
- *Does things efficiently*
- *Perseveres until the task is finished*
- *Tends to be lazy*
- *Is a reliable worker*
- *Can be somewhat careless*
- *Does a thorough job*
- *Tends to be disorganized*

Job performance

- **Supervisor?** Performance assessment?
- Any objective measure of performance?

Validity

The extent a survey measure accurately reflects the intended construct



Validity

$$Y_{it} = \mu_i + \epsilon_{it}$$

μ_i = true value of the construct for the ith respondent
Y_{it} = response to the measure by the ith respondent
ϵ_{it} = deviation from the true value of the construct

Validity can be defined as a measure of correlation:

$$cor(Y_i, \mu_i)$$

In practice, μ_i will be a *gold standard*

Validity

If there is no *gold standard* available...

- Correlation of the answers with answers of other survey questions with which, in theory, they ought to be highly related
- Comparison between groups whose answers ought to differ if the answers are measuring the intended construct
- Comparison of responses from comparable samples of respondents to alternative question wording or protocols for data collection (split-ballot studies) $\sim$ bias

Bias

What happens if there is a systematic deviation in response away from a true value?

Split-ballot experiment:

- *On days when you drink alcohol, how many drinks do you usually have - would you say one, two or three, or more?*
- *On days when you drink alcohol, how many drinks do you usually have - would you say one or two, three or four, five or six, or seven or more?*

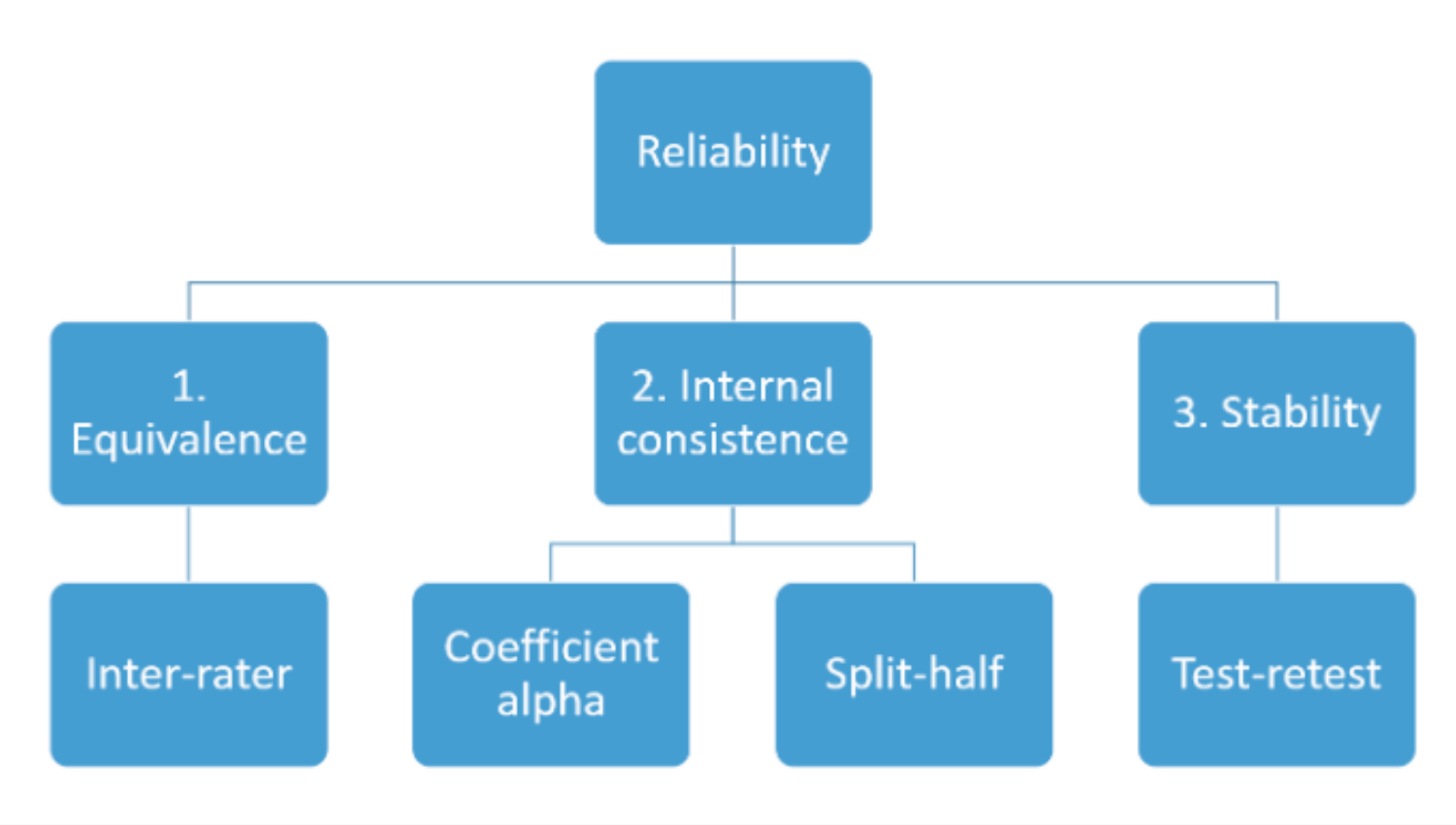
💻 Let's move to R...

Bias

- The notion of bias depends on the existence of a **true value**!
- Whether true values exist for subjective states, such as knowledge, opinions, or feelings, is controversial
- That's why the concept of bias technically applies to measure of **objectively verifiable** facts or events

Reliability

Consistency of measurement either across occasions or across items designed to measure the same construct



Two general approaches

- Repeated interviews with the same respondent
  - No changes in the underlying construct between interviews
  - Measurement protocol remains the same
  - No impact of the first measurement on the second responses (e.g., panel conditioning)
- Use multiple indicators of the same construct
  - All questions are indicators of the same construct
  - All questions have the same expected response deviation (their response variance is constant)
  - The measures of the items are independent

Cronbach's alpha

$$\alpha = \frac{k\bar{r}}{1 + (k-1)\bar{r}}$$

k = number of items

$\bar{r}$ = average inter-correlation

A higher α implies high reliability or low response variance, but can also indicate that one item affects the responses to another

Low α indicates low reliability or that the items do not really measure the same construct

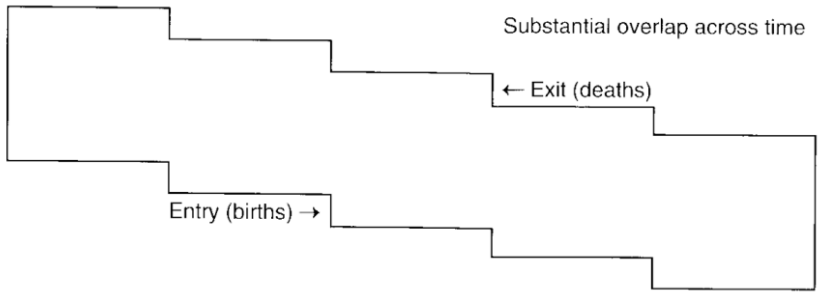
Panel conditioning

Hypotheses on highest impact

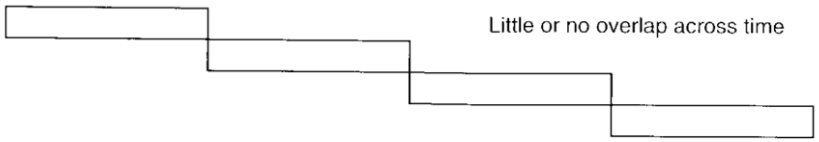
1. Attitudes are less crystallized
2. Increase respondents' knowledge
3. Provide socially non-normative or stigmatized responses
4. Comfort and trust with the survey experience
5. Learning (manipulate instrument or provide more accurate answers)
6. Short time between surveys

Longitudinal designs

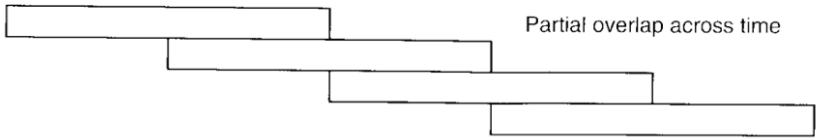
Total Population Design (Example: Census data)



Repeated Cross-Sectional Design (Example: NORC General Social Surveys)



Revolving Panel Design (Example: National Crime Victimization Survey)



Multiple Cohort Panel Design (Example: British Cohort Studies)

Age 11				Age 15
Age 12				Age 16
Age 13	Extensive overlap across time			Age 17
Age 14				Age 18
Age 15				Age 19
Age 16				Age 20

Panel conditioning

Suggestions / thoughts

- More research
  - Rotating panels
  - Experiments on large-scale surveys
  - General theory or framework
- May or may not bias substantive conclusions
  - Topic variability
  - Individual heterogeneity
- Corrections?
  - Similar to mode effects (What can we do?)
  - Ex-ante approaches
  - *We ignore panel conditioning at our own peril*

How reliable are survey measures?

- Survey content, facts versus non-facts
- Source of information, self-reports versus interviewer judgments
- Context of measurement, the position of item batteries
- Formal attributes of questions (type of rating scale, DK option)

Employment

Employment

- Acting as a thought partner for BCG case teams on market research and survey analytics
- Directly driving additional project volumes through broadening and deepening relationships with BCG's consulting team and topic experts
- Application of market research knowledge to interpret and discuss elements of survey design (sampling, quotas, methodology, questionnaire structure etc.)

Employment

- Collaboration with survey programmers, third-party vendors, and partners for implementation of online surveys and data collection
- Quality review of online surveys before launch, data handling and management capabilities to validate and clean data prior to further processing
- Using specialized survey data analysis tools (SPSS) and application of statistical theoretical concepts (univariate, bivariate and multivariate methods) to deliver practical data analytics outcomes

Methods of data collection

Key questions on the mode of collection

Mode: Method or approach used for data collection

- What is the **most appropriate** method for a particular research question?
- What is the **impact** of a particular data collection method on survey **errors and costs**?

Survey technology

From paper to laptops, tablets, smartphones...

- **CAPI:** Computer-assisted personal interviewing
- **ACASI:** Audio computer-assisted self-interviewing
- **CATI:** Computer-assisted telephone interviewing
- **IVR:** Interactive voice response
- **SMS**
- **Web / Social media**

Survey technology

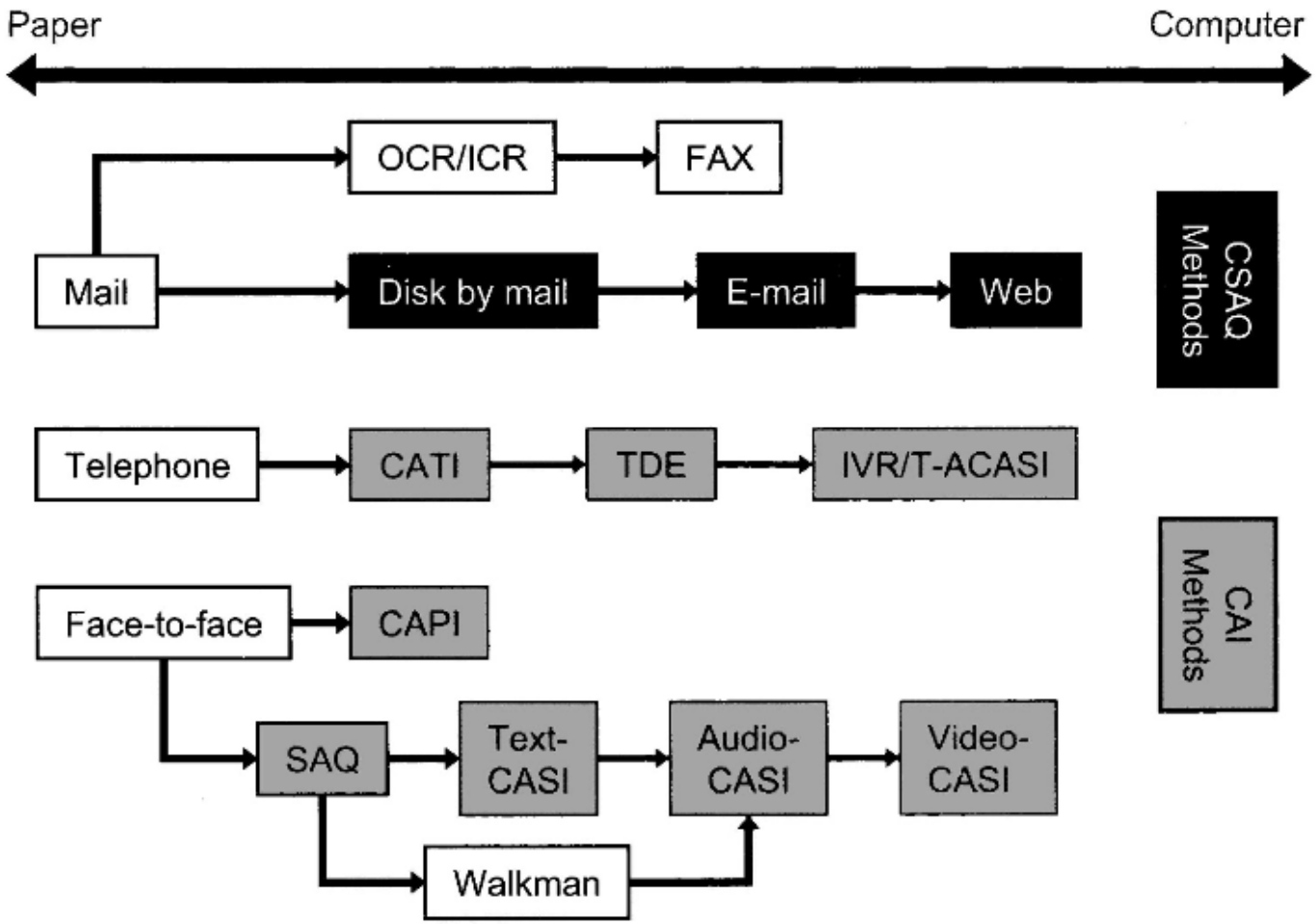
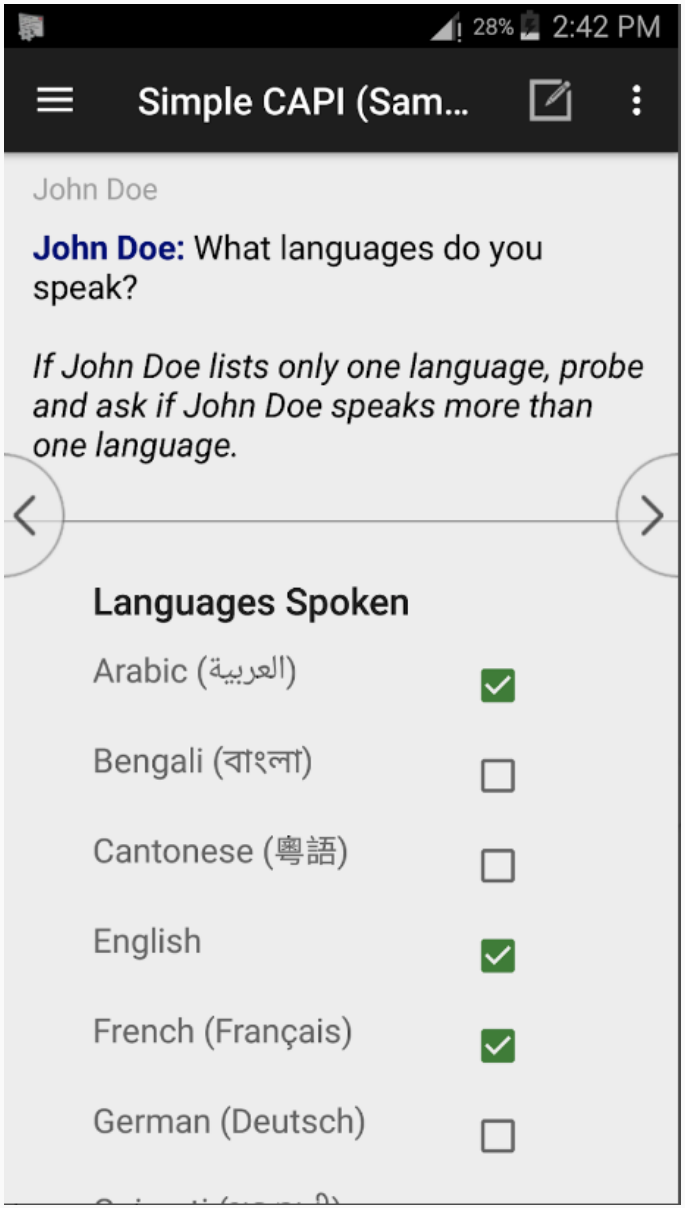


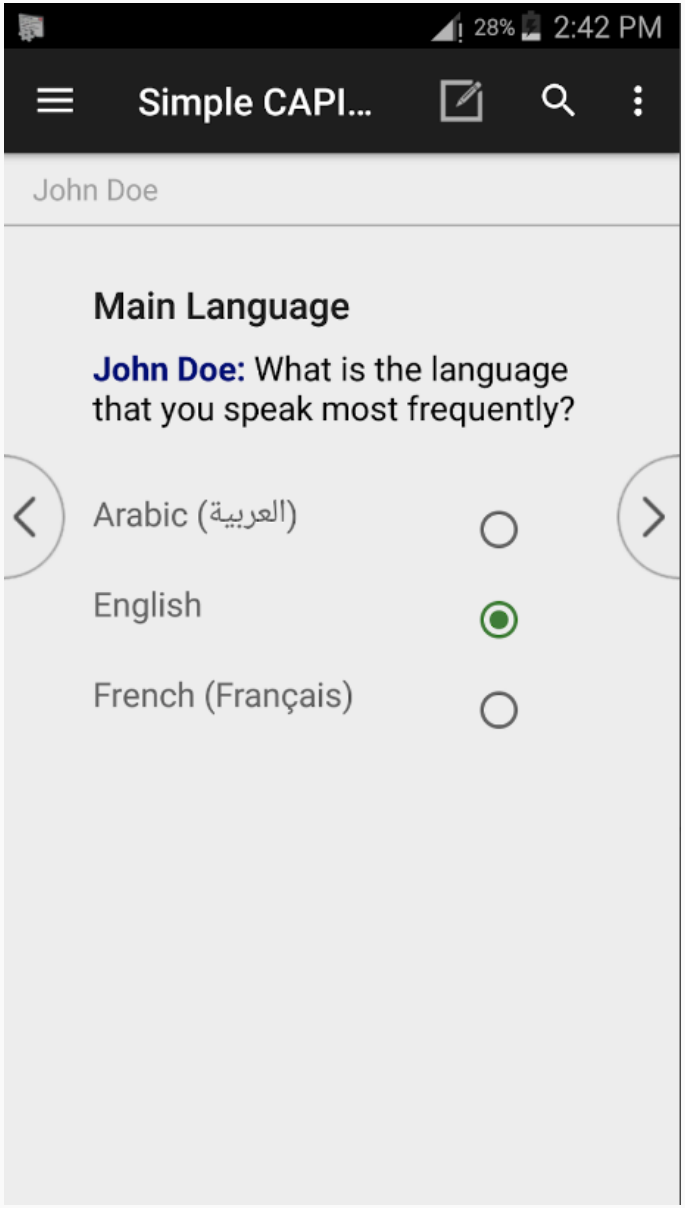
Figure 5.2 The evolution of survey technology.

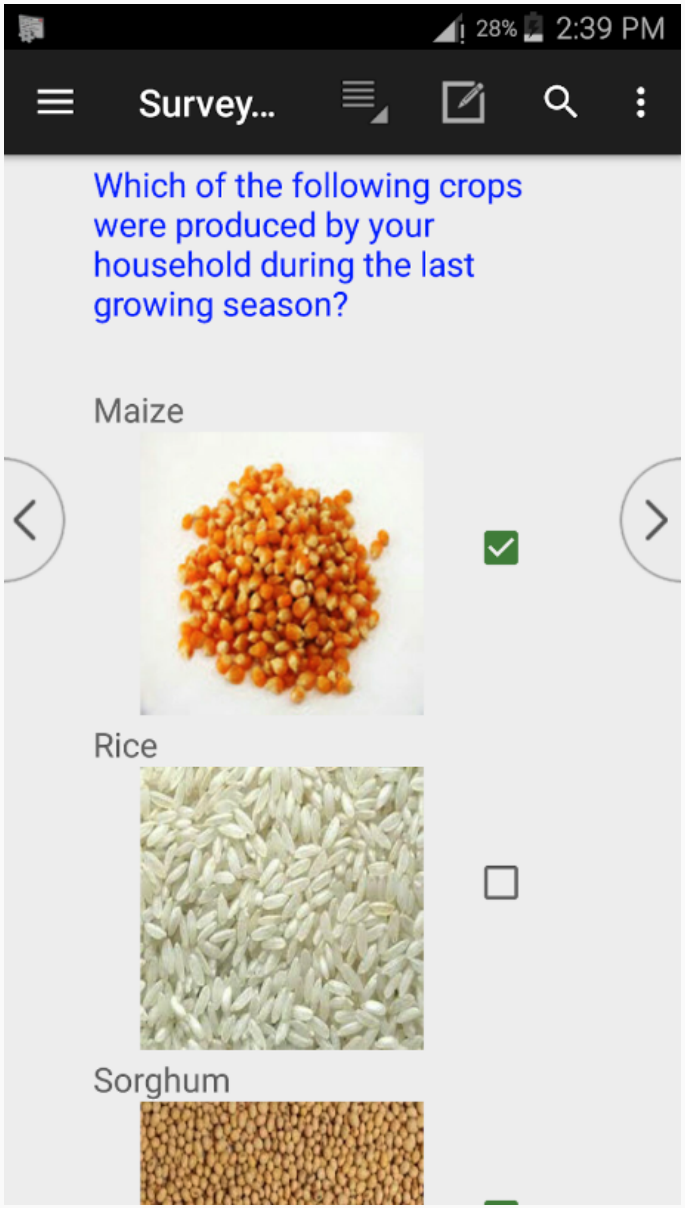
CAPI


28% 2:41 PM
Simple CAPI (Sam...
Enter the name of the next person in the household. Include everyone living or staying in this household.
Do not enter anything and press Next if all persons have been recorded.
First Name
John
Last Name
Doe

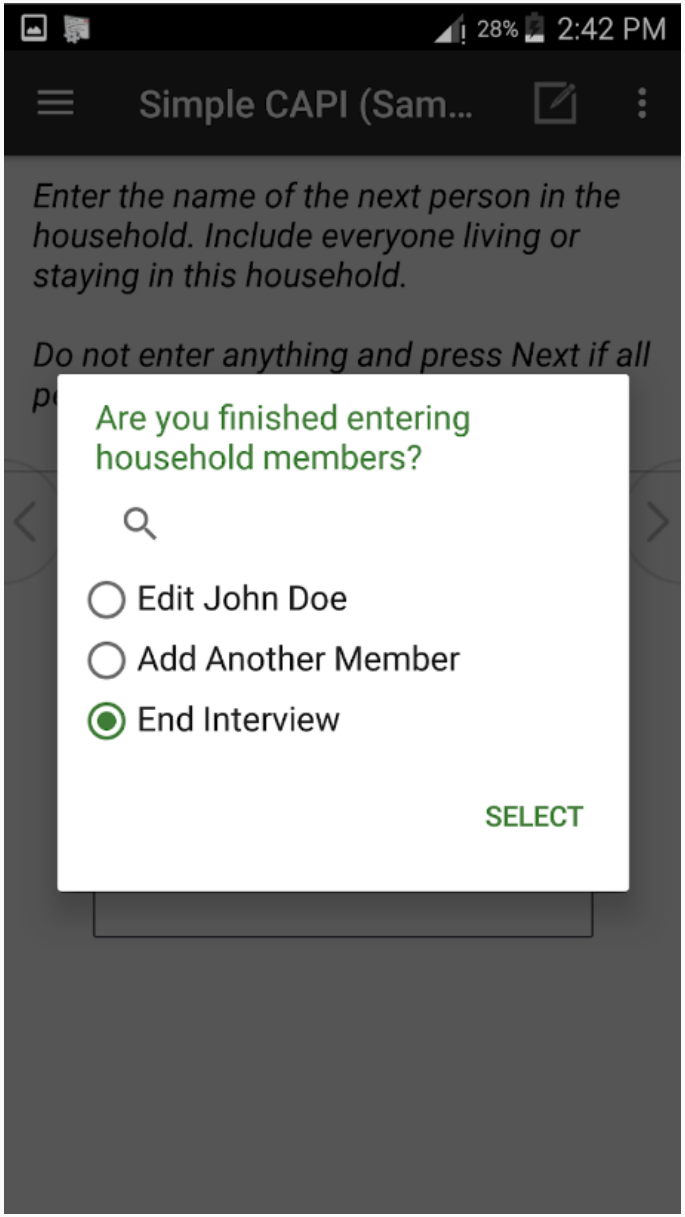

28% 2:42 PM
Simple CAPI (Sam...
John Doe
John Doe: What languages do you speak?
If John Doe lists only one language, probe and ask if John Doe speaks more than one language.
Languages Spoken
Arabic (العربية)
Bengali (বাংলা)
Cantonese (粵語)
English
French (Français)
German (Deutsch)

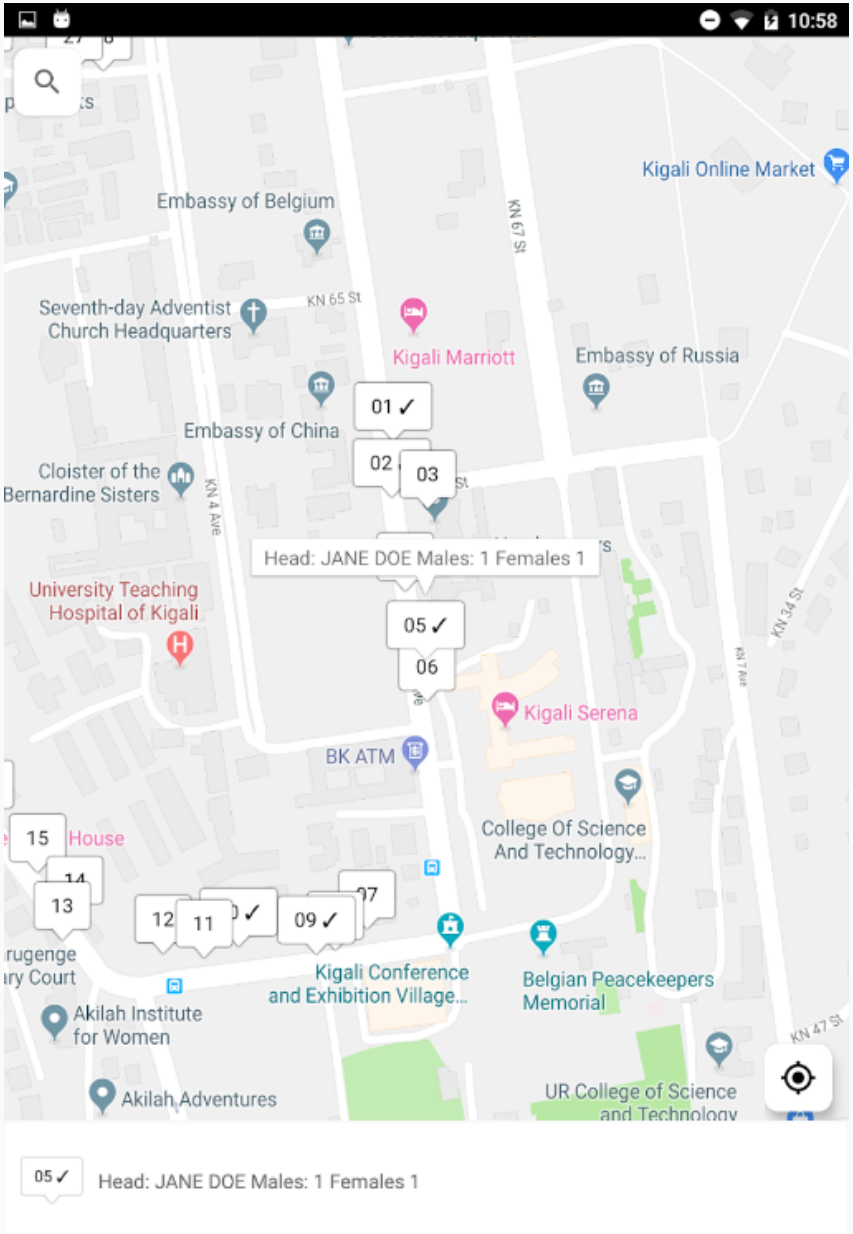
CAPI


28% 2:42 PM
Simple CAPI...
John Doe
Main Language
John Doe: What is the language that you speak most frequently?
Arabic (العربية)
English
French (Français)


28% 2:39 PM
Survey...
Which of the following crops were produced by your household during the last growing season?
Maize
Rice
Sorghum

CAPI

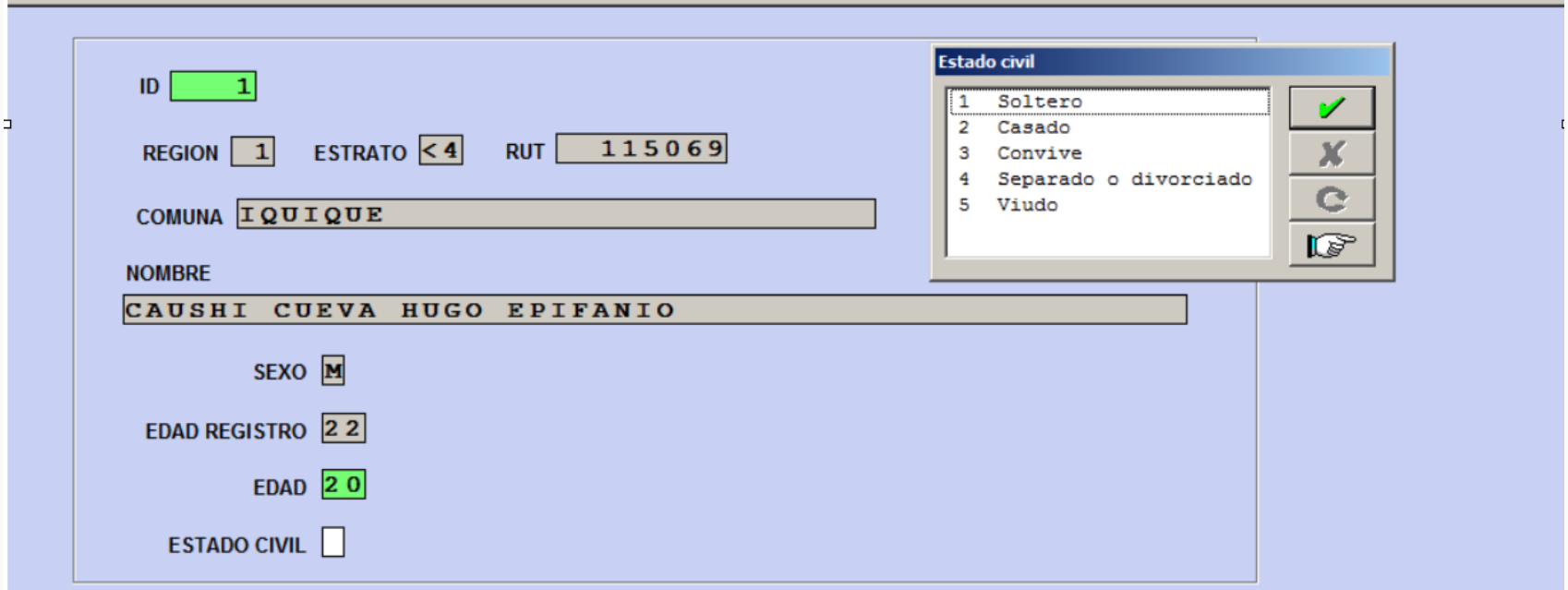

28% 2:42 PM
Simple CAPI (Sam...
Enter the name of the next person in the household. Include everyone living or staying in this household.
Do not enter anything and press Next if all
Are you finished entering household members?
Edit John Doe
Add Another Member
End Interview
SELECT


10:58
Embassy of Belgium
Kigali Online Market
Seventh-day Adventist Church Headquarters
KN 65 St
Kigali Marriott
Embassy of Russia
Embassy of China
01 ✓
02
03
Cloister of the Bernardine Sisters
KN 4 Ave
Head: JANE DOE Males: 1 Females 1
University Teaching Hospital of Kigali
05 ✓
06
Kigali Serena
BK ATM
College Of Science And Technology...
15
House
13
12
11
09 ✓
Kigali Conference and Exhibition Village...
Belgian Peacekeepers Memorial
Akilah Institute for Women
Akilah Adventures
UR College of Science and Technology
KN 47 St
05 ✓
Head: JANE DOE Males: 1 Females 1

CAPI

¿Cuál es tu estado civil?

1 Soltero
2 Casado
3 Convive
4 Separado o divorciado
5 Viudo

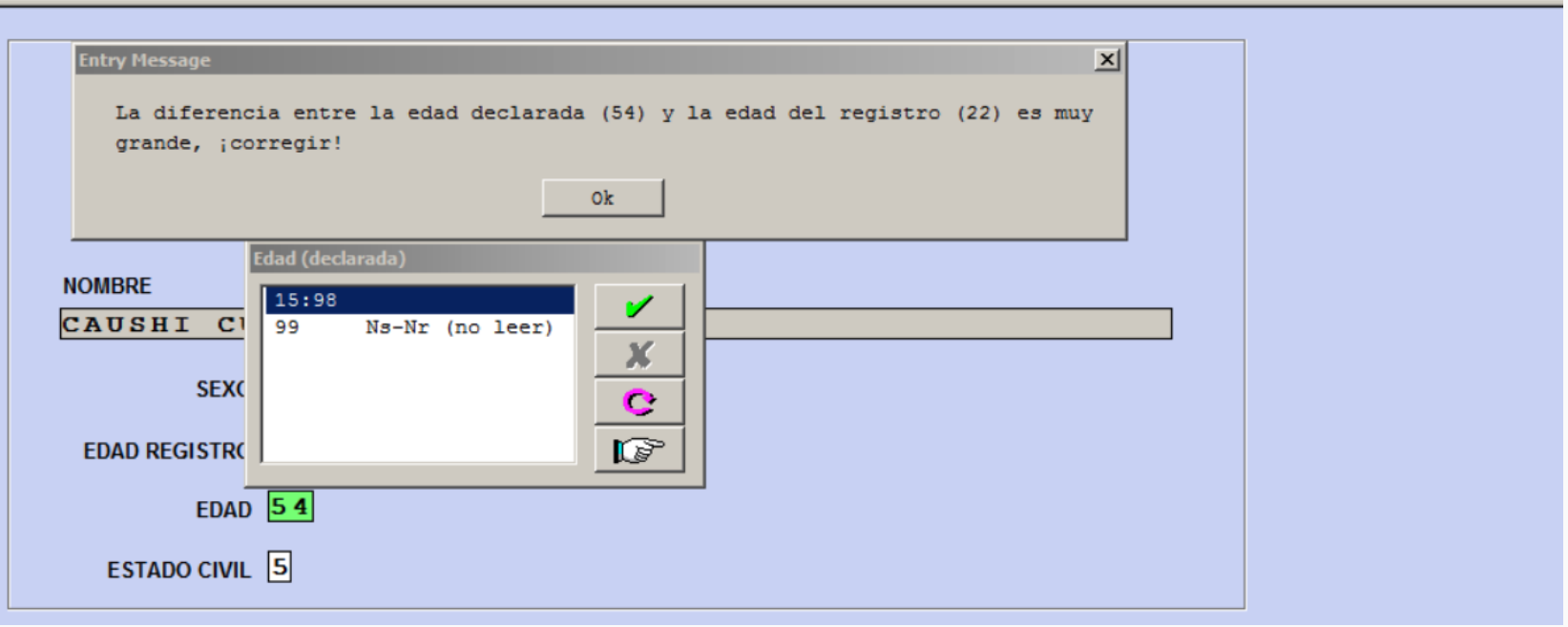


CAPI

Estimado(a) *CAUSHI CUEVA HUGO EPIFANIO*, INACAP se encuentra realizando un estudio sobre las principales razones y motivos por las sus alumnos dejan de estudiar en la institución. Para ello, ha seleccionado al azar a un grupo de ex-alumnos con el objetivo de aplicarles un breve cuestionario. Para nosotros es muy importante poder contar con tu opinión.

Te recordamos que tus respuestas serán absolutamente privadas y confidenciales. Muchas gracias por tu tiempo y colaboración.

Para empezar, ¿me podrías decir tu edad?



Dimensions affected by the mode of collection

- Degree of involvement of the interviewer
  - **Pros:** decrease of non-response, sampling recruitment, clarification of questions
  - **Cons:** costs, desirability bias
- Degree of interaction with respondents
  - **Pros:** researcher control of data collection
  - **Cons:** contextual effects
- Degree of privacy
  - Group application in schools (e.g., drug use)
  - Application in households (e.g., family violence)
  - Characteristics of the interviewer

Dimensions affected by the mode of collection

- Channels of communication
  - Visual, audio
  - Social presence
- Technology used
  - Training (CAPI / CATI)
  - Complex skips versus programming errors
  - Control of respondents (move backward, save)
  - Educational and literacy constraints

Design implications of mode

- Sampling frame and design
  - Area probability frames and face-to-face interviews
    - Gold standard (but with some limits)
  - Sample selection, screening $\rightarrow$ better done by interviewers
  - Panel studies: face-to-face $\rightarrow$ telephone

Design implications of mode

- Non-response
  - *face-to-face* $\rightarrow$ *telephone* $\rightarrow$ *mail* $\rightarrow$ *web*
  - Reasons for non-responding might also differ
  - How much info can we collect from the non-response unit

I received this survey some time ago... 🤔

After finishing the review of a scientific paper, I got a survey to assess the reviewing process. I didn't finish it. Do you guess why?

ELSEVIER

When considering whether or not to accept an invitation to review an article, how important are the following?

	Extremely important	Important	Neither important nor unimportant	Not very important	Not at all important	Don't know/ not applicable
Invitation to review (e.g. politeness, appropriateness etc)	○	○	○	○	○	○
Quality of the article	○	○	○	○	○	○
The article closely matches your area of expertise	○	○	○	○	○	○
Length of time given to review the article	○	○	○	○	○	○
Reputation of the journal	○	○	○	○	○	○
Instructions and criteria for review	○	○	○	○	○	○
Quality of online review system (e.g. it is reliable and easy to use)	○	○	○	○	○	○
Tools and support offered by publisher to help during review (e.g. help finding related articles)	○	○	○	○	○	○
Editor communication (e.g. assistance, feedback etc)	○	○	○	○	○	○
Recognition/rewards offered by the journal (e.g. reviewer certificates, credits)	○	○	○	○	○	○

I received this survey some time ago... 🤔

0% 25% 50% 75% 100%

We would like you to think about the article you most recently reviewed for Advances in Life Course Research. Please indicate the extent to which you agree/disagree with the following statements about your experience.

Invitation process

	Agree strongly	Agree	Neither agree nor disagree	Disagree	Disagree strongly	Don't know/ not applicable
The invitation was polite and courteous	○	○	○	○	○	○
The amount of information I received regarding the article was sufficient for me to decide to review	○	○	○	○	○	○
The article was relevant to my field of expertise	○	○	○	○	○	○
I had sufficient time to decide whether or not to accept the invitation to review	○	○	○	○	○	○
I found it easy to communicate my decision to accept/reject the invitation	○	○	○	○	○	○
I found it straightforward to register as a reviewer for the journal	○	○	○	○	○	○

<< >>

I received this survey some time ago... 🤔

0% 25% 50% 75% 100%

In reference to the article you most recently reviewed for Advances in Life Course Research, please indicate the extent to which you agree/disagree with the following statements about your experience.

Reviewing process

	Agree strongly	Agree	Neither agree nor disagree	Disagree	Disagree strongly	Don't know/ not applicable
It was clear to me by what criteria I should evaluate the article	○	○	○	○	○	○
The length of time given to complete the review was reasonable	○	○	○	○	○	○
I could read the manuscript and figures clearly with no technical problems	○	○	○	○	○	○
The quality of research in the article was sufficiently good to merit peer review	○	○	○	○	○	○
The standard of English in the article was reasonable	○	○	○	○	○	○
The review format and structure for review submission was helpful	○	○	○	○	○	○
There was a good level of communication with the Editor during the review process	○	○	○	○	○	○
The reminders I received were timely and useful	○	○	○	○	○	○
It was easy to submit my report/recommendations for the article	○	○	○	○	○	○

<< >>

I received this survey some time ago... 🤔

0% 25% 50% 75% 100%

In reference to the article you most recently reviewed for Advances in Life Course Research, please indicate the extent to which you agree/disagree with the following statements about your experience.

Performance of online review platform/ services/ support

	Agree strongly	Agree	Neither agree nor disagree	Disagree	Disagree strongly	Don't know/ not applicable
The system is reliable and robust	○	○	○	○	○	○
It is a fast system for review	○	○	○	○	○	○
The online review platform is available whenever I need to use it	○	○	○	○	○	○
The online help for reviewers is clear and easy to use	○	○	○	○	○	○
Tools offered by the publisher to help conduct the review were useful	○	○	○	○	○	○

<< >>

Design implications of survey mode

Measurement quality: social desirability bias

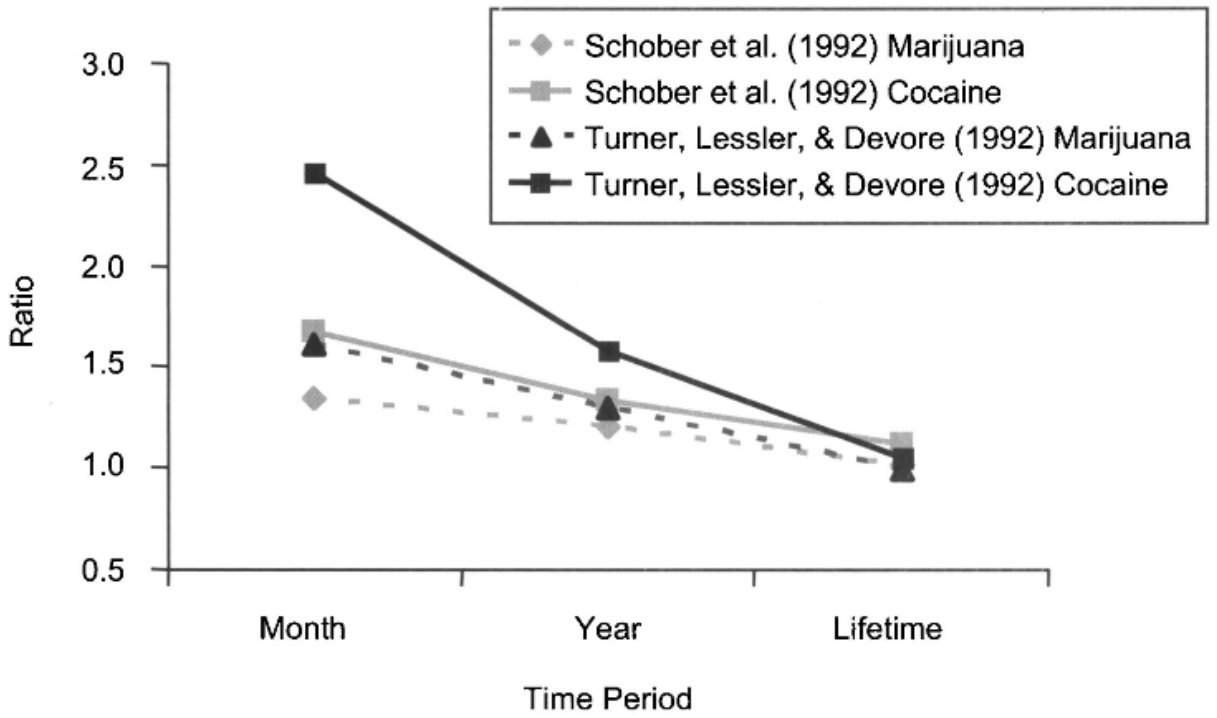


Figure 5.4 Ratio of proportion of respondents reporting illicit drug use in self-administered versus interviewer-administered questionnaires, by time period by drug.

Design implications of mode

Measurement quality: response effects 🤔

Primacy versus recency effects

P46STM.- Reasons why people are not treated equally
Reasons because people are not treated equally. Of all of the reasons for which people in this country are not treated equally, which one affects you the most?

0 No answer
1 For being youth
2 For being old
3 For being woman
4 For being a man
5 For skin color
6 For being Immigrant
7 For being homosexual
8 For religion
9 For being poor
10 For being handicapped
11 For not having enough education
12 For not having connections
13 For not being anyone
14 (Country) treats all equally
15 Not discriminated by anybody

Design implications of mode

Measurement quality: data completeness

- Fewer questions are answered in self-administered questionnaires
- Item-missing data
- Length opened-ended questions

Design implications of mode

Costs

- Training
- Supervision
- Interviewers
- Servers/databases
- Providers

Summary

Dimension	Paper	CAPI	CATI	Mail	e-mail	Web
Flexibility data collection	high	medium-high	medium-high	low	low	low
Interaction interviewer	yes	yes	yes	no	no	no
Question diversity	high	high	low	medium	medium-high	medium-high
Visual and audio stimuli	high	high	low	medium-low	medium-high	medium-high
Sample size per budget	low	low	medium	high	high	high
Sample control	medium	medium	high	low	medium-low	low
Control collection context	high	high	medium	low	low	low
Fieldwork control	medium-low	medium	high	no	no	no
Training interviewers	yes	yes	yes	no	no	no
Information per respondent	high	high	medium	medium	medium	medium
Response rate	high	high	medium	low	low	low
Privacy (perception)	medium	medium	medium	high	medium	high
Social desirability	high	high	medium	low	low	low
Getting sensitive information	low	low	medium-high	high	high	high
Interviewer bias	high	medium	medium	no	no	no
Speed	medium-low	medium-low	medium-high	low	high	high
Cost	high	high	medium	low	very low	very low

Which mode seems more suitable for this question? 🤔

P69. Lea con detención las siguientes frases e indique cuál de ellas se acerca más a la realidad de su empresa. Use la escala de 1 a 5, en donde el número 1 es más cercano a la frase A y 5 es más cercana a la frase B

	Frase A	← Escala →					Frase B
1	El **capital** con que cuento en la actualidad *no me alcanza* para realizar inversiones.	1	2	3	4	5	El **capital** con que cuento en la actualidad *me alcanza* para realizar inversiones.
2	El **acceso al financiamiento** *es* una dificultad real para mis inversiones.	1	2	3	4	5	El **acceso al financiamiento** *no es* una dificultad real para mis inversiones.
3	Los **costos de los créditos** *son* una dificultad real para mis inversiones.	1	2	3	4	5	Los **costos de los créditos** *no implican* una dificultad para mis inversiones.
4	Mi **endeudamiento** *es* una dificultad para realizar inversiones en mi empresa.	1	2	3	4	5	Mi **endeudamiento** *no presenta* una dificultad para realizar inversiones en mi empresa.
5	La **rentabilidad** de mi negocio *no me permite* realizar inversiones en mi empresa.	1	2	3	4	5	La **rentabilidad** de mi negocio *me permite* realizar inversiones en mi empresa.

How do you ask this question in a telephone survey?

Engagement in mining company from Chile 🤔

Codelco

- About 15000 workers
- Important differences between production workers and the rest of the staff
- Several units/department
- Public company, unions are very strong
- Six different regions in Chile

Which mode would you use?

Engagement in mining company from Chile 🤔

Codelco

Which mode would you use?

Mixed-mode surveys

- Key areas of impact
  - Reduce coverage error
  - Reduce non-response error
  - Lowering survey costs
- Limitations
  - Unified mode design
  - It can be time consuming

Mixed-mode surveys

- Typology
  - **Type I**: contact one mode, respond using a different mode
  - **Type II**: mostly one mode, specific question another mode (reduce social desirability effects) $\sim$ concurrent
  - **Type III**: Several modes for different respondents during the same survey period
  - **Type IV**: Use one mode on one occasion, and a different mode (follow-up)

Questions and answers

Model of survey response process

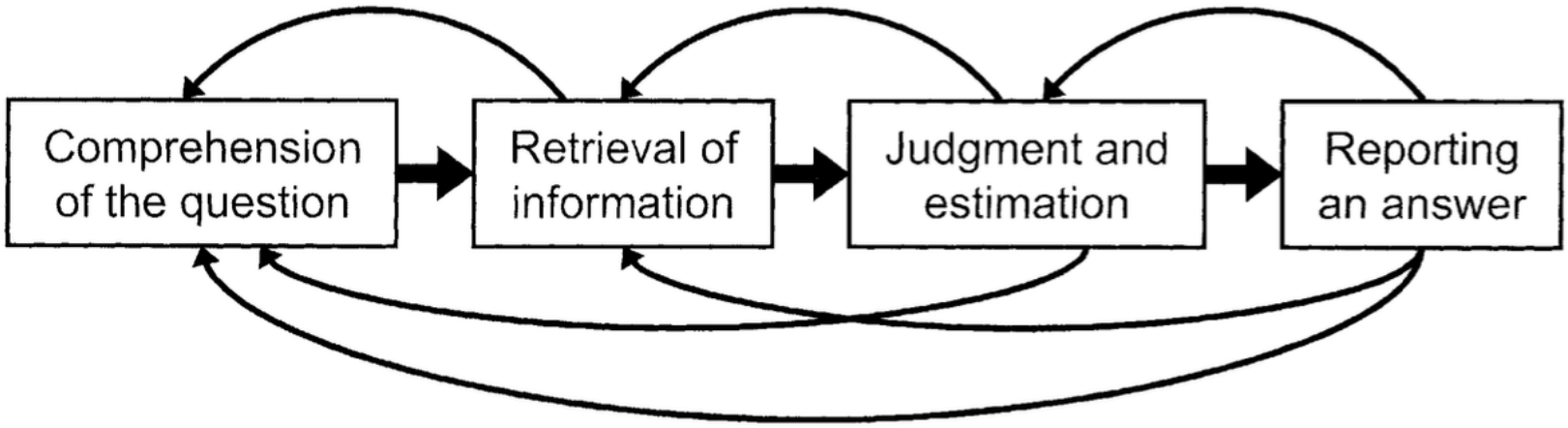


Figure 7.1 A simple model of the survey response process.

Problems in answering survey questions

- Failure to encode the information sought
- Misinterpretation of questions
- Forgetting and memory problems
- Flawed judgment or estimation strategies
- Problems in formatting an answer
- More or less deliberate misreporting
- Failure to follow instructions

Misinterpreting questions

- Grammatical ambiguity
- Excessive complexity
- Faulty presupposition
- Vague concepts
- Vague quantifiers
- Unfamiliar terms
- False inferences
- Cultural differences

Grammatical ambiguity + complexity 🤔

During the past 12 months, since [DATE], how many times have you seen or talked to a doctor or assistant about your health? Do not count any time you might have seen a doctor while you were a patient in a hospital, but count all other times you actually saw or talked to a medical doctor of any kind.

What is being asked?

Faulty presupposition

How much do you agree or disagree with this sentence: Family life often suffers because men concentrate too much on their work

False inference (intention)

Are there any situations you can imagine where you would approve of a policeman striking an adult male citizen?

Any thoughts regarding this question? 🤔

P28NF.- Questioning the leaders /authority
Which of the following statements is closest to your view? Choose statement

A. We should be more active in questioning the actions of our leaders.

B. As citizens, we should show more respect for authority.

0 No answer
1 Agree strongly with A
2 Agree with A
3 Agree with B
4 Agree strongly with B
7 None
8 Don't know

Cultural differences

- EcoSocial Study (Social Cohesion in LA)
- Corporación de Estudios para Latinoamérica (CIEPLAN)
- 7 countries, 2007

Cultural differences

95% of people in Guatemala declare to believe in God

Do you consider yourself a person?

Very religious
Quite religious
Somewhat religious
Not very religious
Not religious at all

Recap on quirky question

How much do you agree or disagree with this sentence: Family life often suffers because men concentrate too much on their work

What do you want to measure?

- Macho/conservative stance
- Agree or disagree with what (suffering or working too much)?

Forgetting and memory problems

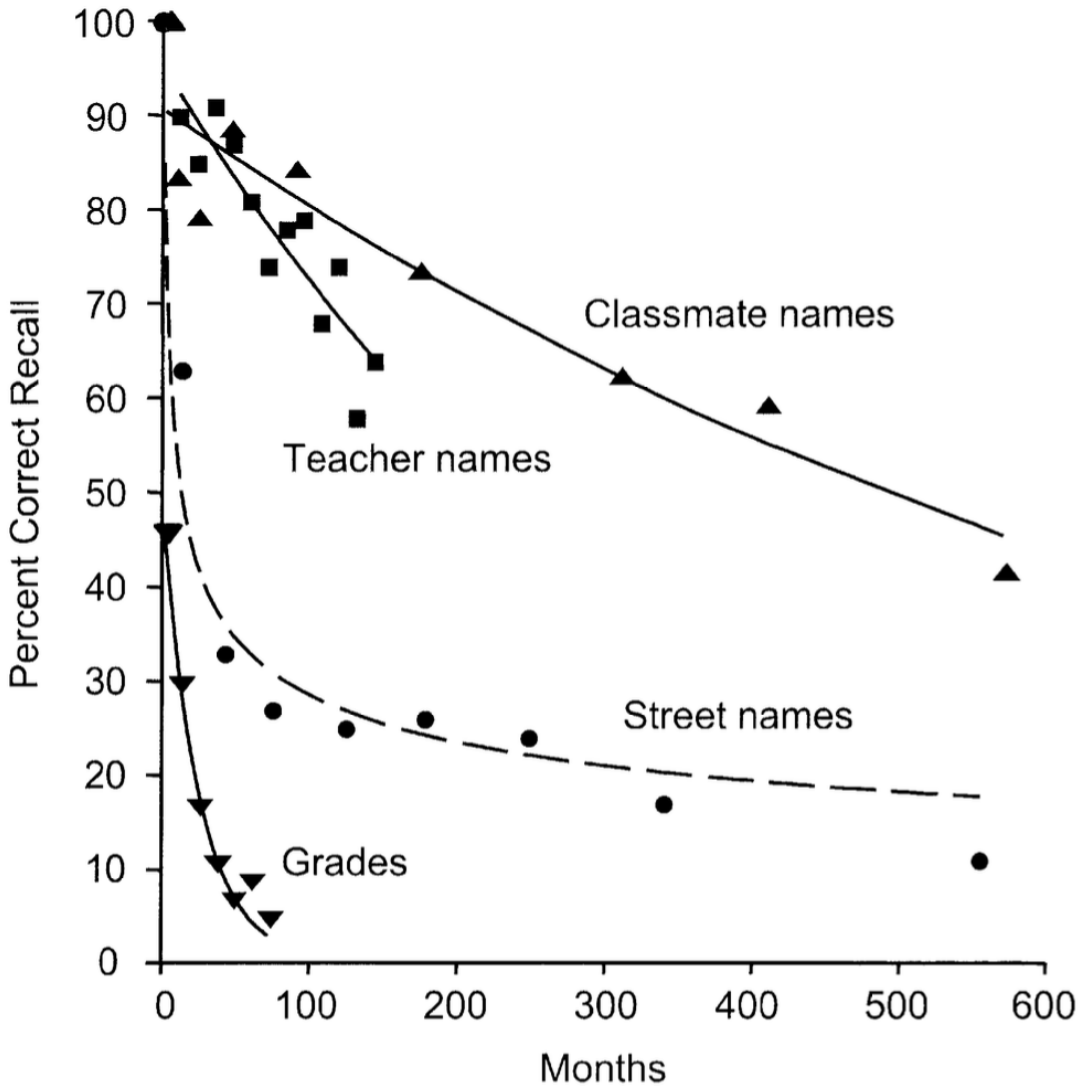


Figure 7.2 Recall accuracy for types of personal information. (Source: Tourangeau, Rips, and Rasinski, 2000.)

Forgetting and memory problems

Reentry study

8. ¿Ha habido algún evento importante para ti? **Anotar la última entrevista realizada como evento en el primer mes del calendario, para marcar el mes de inicio. (NOTA: Registrar con una X los meses en que ocurrió, y anotar descripción del evento).**

MES		1	2	3	4	5	6	7	8	9	10
Día/Mes/Año		__/__/__	__/__/__	__/__/__	__/__/__	__/__/__	__/__/__	__/__/__	__/__/__	__/__/__	__/__/__
Evento	Descripción literal										
Evento 1											
Evento 2											
Evento 3											
Evento 4											
Evento 5											
Evento 6											
Evento 7											
Evento 8											

Luego regresaremos a usar el calendario, sin embargo ahora revisaremos otras cosas sin utilizarlo….

Estimation process for behavioral questions

Now think about the past 12 months, from [DATE] through today. We want to know how many days you've used any prescription tranquilizer that was not prescribed to you or that you took only for the experience or feeling it caused during the past 12 months.

National Survey on Drug Use and Health (NSDUH)

Strategies

- Recall-and-count $\rightarrow$ telescoping $\rightarrow$ over-reporting
- Rate-based-estimation
- Impression-based-estimation

Judgment processes for attitude questions

*Do you think the US make a mistake in deciding to defend Korea or not? (**Gallup**)*

*Do you think the US was right or wrong in sending American troops to stop the Communist invasion of South Korea? (**NORC**)*

Which question did get more support to the US?

Formatting the answer

The most common will be:

- Open-ended questions that call for numerical answers
- Closed questions with ordered response scales
- Closed questions with categorical response options

Formatting the answer 🤔

Would you say that in general your health is (*read options*):

(1) Excellent
(2) Very good
(3) Good
(4) Fair
(5) Poor

Behavioral Risk Factor Surveillance System (BRFSS)

Any issues here?

Formatting the answer 🤔

> *Now, thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?*
>
> *Behavioral Risk Factor Surveillance System (BRFSS)*

Any issues here? Better than the previous question? Pros and cons?

Ordinal scales

Let's check some examples here!

- Statements to be agreed or disagreed with is one of the most common formats
- They are simple to construct and answer
- But, they encourage a tendency to agree on irrespective of the item content (acquiescence)

Acquiescence

0% 25% 50% 75% 100%

In reference to the article you most recently reviewed for Advances in Life Course Research, please indicate the extent to which you agree/disagree with the following statements about your experience.

Reviewing process

	Agree strongly	Agree	Neither agree nor disagree	Disagree	Disagree strongly	Don't know/ not applicable
It was clear to me by what criteria I should evaluate the article	○	○	○	○	○	○
The length of time given to complete the review was reasonable	○	○	○	○	○	○
I could read the manuscript and figures clearly with no technical problems	○	○	○	○	○	○
The quality of research in the article was sufficiently good to merit peer review	○	○	○	○	○	○
The standard of English in the article was reasonable	○	○	○	○	○	○
The review format and structure for review submission was helpful	○	○	○	○	○	○
There was a good level of communication with the Editor during the review process	○	○	○	○	○	○
The reminders I received were timely and useful	○	○	○	○	○	○
It was easy to submit my report/recommendations for the article	○	○	○	○	○	○

<< >>

Acquiescence

How to offset the effect of acquiescence?

- Balancing direction of agree-disagree items 😐
- Do not use agree-disagree items, but forced-choice questions

Acquiescence (before)

Do you agree or disagree that most of men are better suited emotionally for politics than are most women?

Acquiescence (after)

Would you say that most of men are better suited emotionally for politics than are most women, that men and women are equally suited, or that women are better suited than men?

Acquiescence (before)

Management lets employees know how their work contributes to the company's goals

Acquiescence (after)

Some people think management lets employees know how their work contributes to the company's goals. Other people think management does not let employees know how their work contributes to the company's goals. Which comes closer to how you feel?

Meaning of frequency categories

Relative frequencies are not simple translations of absolute frequencies, they incorporate evaluative information

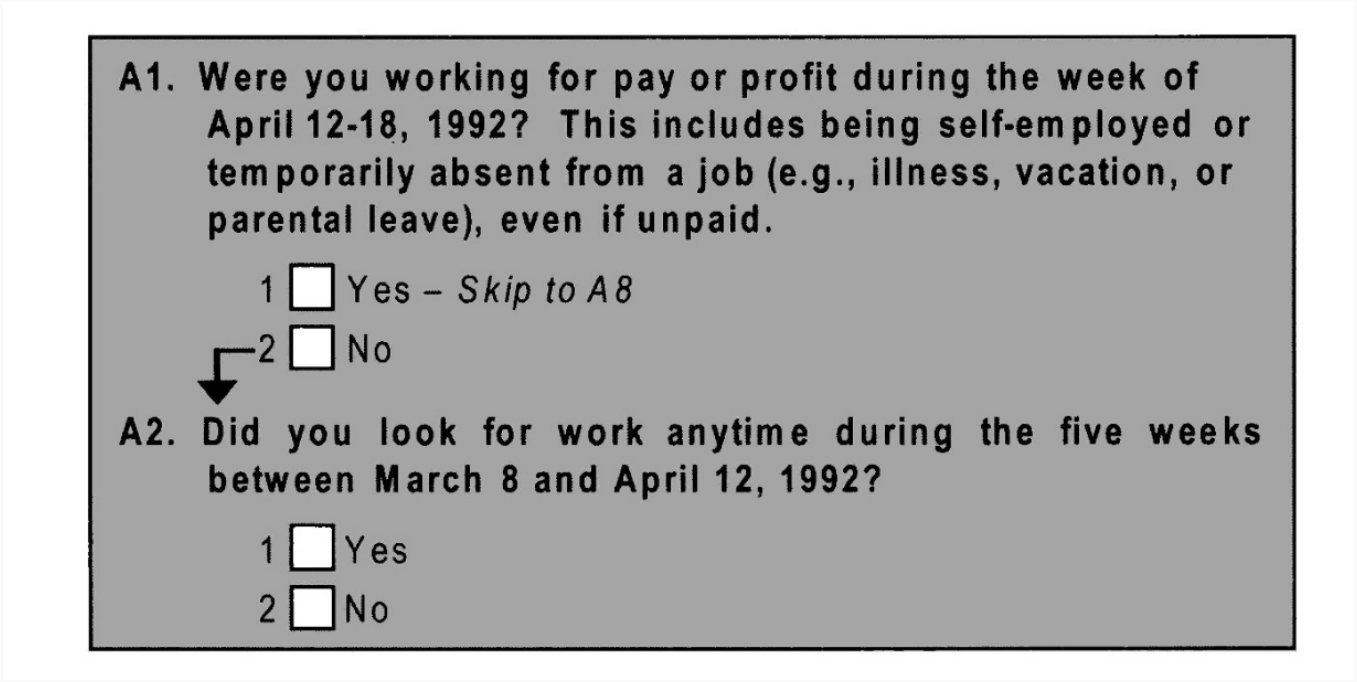
> *Both Annie and Alvie Singer report having sex three times a week, but she characterizes this as **constantly**, whereas his description is **hardly ever***
>
> *Woody Allen's Annie Hall*

Motivated mis-reporting

*Think especially about the last 30 days, from [DATE] up to and including today. During the past 30 days, on how many days did you use cocaine? (**NSDUH**)*

*In talking to people about elections, we often find that a lot of people were not able to vote because they were not registered, they were sick, or they just didn't have the time. How about your – did you vote in the elections this November? (**American National Election Study**)*

Navigational errors



A1. Were you working for pay or profit during the week of April 12-18, 1992? This includes being self-employed or temporarily absent from a job (e.g., illness, vacation, or parental leave), even if unpaid.

1 ☐ Yes – *Skip to A8*

2 ☐ No

A2. Did you look for work anytime during the five weeks between March 8 and April 12, 1992?

1 ☐ Yes

2 ☐ No

Double-negative

Please tell me whether you agree or disagree with the following statement about teachers in public schools: Teachers should not be required to supervise students in the halls, the lunchroom, and the school parking lot

Ranking questions

People look for different things in a job. Below is a list of six factors in job satisfaction. Please mark the items in the order of their importance to you, starting with 1 as the most important factor in job satisfaction, through 6 for the least important factor.

- Work that pays well
- Work that gives a sense of accomplishment
- Work where you can advance
- Pleasant working environment
- Steady place of work with minimal chances of being laid off
- Work that provides good social benefits

Some guidelines

Non-sensitive questions

- All reasonable possibilities in closed questions
- Specific questions!
- Words easy to understand
- Memory clues to improve recall or use aided recall
- For long recall periods use a life event calendar
- Diary, household records, bounded recall
- Use proxies when the cost is a factor

Some general guidelines

Sensitive questions

- Open rather closed questions (eliciting frequency of sensitive behaviors)
- Longer than shorted questions (loading)
- Familiar words
- Ask about long periods (distant events)
- Sensitivity flow
- Self-administrated methods
- Validation data

Some general guidelines

Attitude questions

- Specify the attitude object clearly
- Avoid double-barreled questions
- Measure the strength of attitude
- Use bipolar items
- Alternatives mentioned in the question have a big impact on answers
- Panel studies, measure the same questions each time

Some guidelines

Attitude questions

- Multiple items, start with the least popular
- Use five to seven response scales and label every scale point
- Use lever or analog devices to collect more detailed scale information
- Use ranking only if the respondents can see all alternatives, otherwise use pair comparisons
- Get ratings for every item of interest, not check all-that-apply

How would you evaluate these questions? 🤔

Do you think the government is spending too little, about the right amount, or too much on anti-terrorism measures?

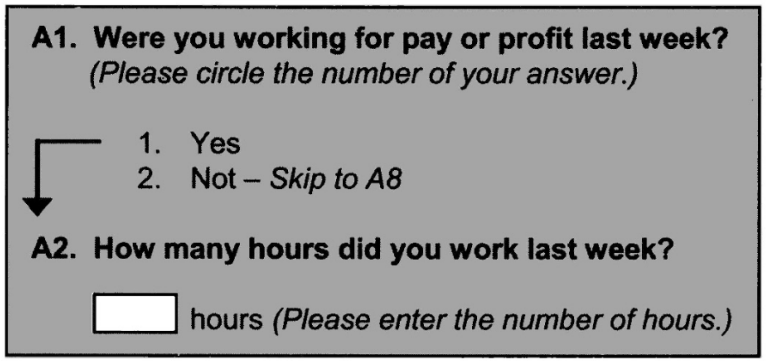
Do you favor legalized abortion because it gives women the right to choose?

The US Supreme Court has ruled that women should be able to end a pregnancy at any time during the first three months. Do you favor or oppose this ruling?

Some guidelines

Self-administered questions

- Use visual elements in a consistent way (path of questionnaire)
- Place directions where they are to be used and where they can be seen
- Ask one question at a time
- Simple and short questions



Let's review an example

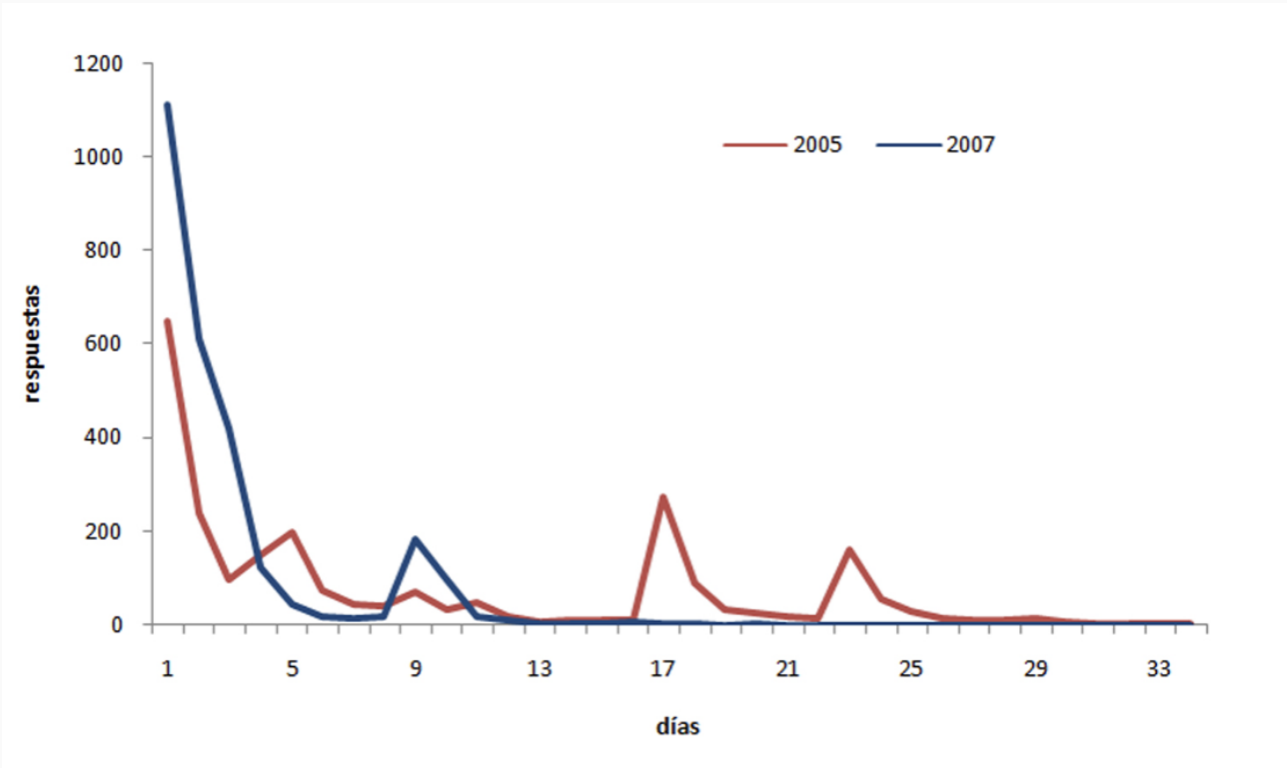
The U.S. Trans Survey (USTS) is the largest survey of trans people, by trans people, in the United States. The USTS documents the lives and experiences of trans and non-binary people ages 16+ in the U.S. and U.S. territories.

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

Sample validation Web surveys

- IP addresses
- Detect and mitigate bot attacks
- Anomaly detection

PUC Student survey on drug use (emails)



Assessing survey questions

Methods to evaluate draft questions

- Expert reviews
- Focus group (focused discussion)
- Cognitive interviews
- Field pretests
- Randomized or split-ballot experiments

Cognitive interviewing

- Concurrent think-alouds
- Retrospective think-alouds
- Confidence ratings
- Paraphrasing (own respondent's words)
- Definitions
- Probes (follow-up questions)

Field pretest

- **Small scale rehearsal**
  - Check CAPI/CATI/Web survey are working correctly
- Sample $\sim$ 100
- Debriefings with interviewers
- Data are tabulated
  - Rate of missing data
  - Distribution of answers
  - Reliability / consistency
- **Record interviews**
  - Behavior coding

Behavioral coding

Table 8.1. Examples of Behavior Codes for Interviewer and Respondent Behaviors

Code Category	Description
Interview Questioning Behaviors (choose one)	
1.	Reads question exactly as worded
2.	Reads questions with minor changes
3.	Reads questions so that meaning is altered
Respondent Behaviors (check as many as apply)	
1.	Interrupts question reading
2.	Asks for clarification of question
3.	Gives adequate answer
4.	Gives answer qualified about accuracy
5.	Gives answer inadequate for question
6.	Answers “don’t know”
7.	Refuses to answer

Create a questionnaire from scratch

Seven steps for questionnaire design

1. Define conceptual and construct variables according to the research objectives
2. Formulate preliminary survey items according to the above constructs
3. Examine preliminary questionnaire item
4. Write an introduction and instructions
5. Run an empirical examination in a small representative study (pilot study)
6. Correct and rephrase items according to findings from the previous stage
7. Make any final adjustments and modifications

Operationalization

Let’s check an example...

Useful steps/questions to construct questionnaires

First: Review relevant literature and previously tested tools

- Who is your target population?
- Do you fully understand the issues and concepts to be examined?
- What are the main variables of interest?
- What do you want to learn about the above variables from your target population?
- How will the questionnaire be administered (mode)?

Useful steps/questions to construct questionnaires

Second: Write the items to be included in your questionnaire

- Have you clearly identified the concept and construct of the variables of interest?
- Have you examined other relevant and related questionnaires?
- Have you consulted others to make sure your items are clear?

Useful steps/questions to construct questionnaires

Second: Write the items to be included in your questionnaire

- Are the questions simply presented?
- Do the questions cover the scope of issues intended?
- Are items too complex and convoluted and potentially confusing to potential participants?
- Do any items use double negatives?
- Are any questions leading or loaded?

Useful steps/questions to construct questionnaires

Third: Design the layout and overall questionnaire

- Does the questionnaire have a title, clear introduction, and section directions?
- Do the title and introduction promote interest in the research?
- Do the title and introduction promote participation and completion?
- Are items logically arranged?
- Are directions in each section clear and easy to follow?

0% 25% 50% 75% 100%

ELSEVIER

In reference to the article you most recently reviewed for Advances in Life Course Research, please indicate the extent to which you agree/disagree with the following statements about your experience.

Performance of online review platform/ services/ support

	Agree strongly	Agree	Neither agree nor disagree	Disagree	Disagree strongly	Don't know/ not applicable
The system is reliable and robust	○	○	○	○	○	○
It is a fast system for review	○	○	○	○	○	○
The online review platform is available whenever I need to use it	○	○	○	○	○	○
The online help for reviewers is clear and easy to use	○	○	○	○	○	○
Tools offered by the publisher to help conduct the review were useful	○	○	○	○	○	○

<< >>

2022 USTS

English

Please make an ID. The research team will use your ID for their analysis. It will not be used to identify you. Please answer the following questions to make your ID:

What are the first and last letters of your first name?
(Example: Sam = SM)

What are the last two digits of your primary phone number?
(Example: 123-456-7890 = 90)

What are the last two digits of the zip code where you reside?
(Example: 01234 = 34)

What are the first and last letters of your mother's **first** name?
(Example: Alex = AX) (Enter XX if you do not know your mother's first name)

Useful steps/questions to construct questionnaires

Fourth: Empirical examination and pilot test

- How long does it take participants to complete the questionnaire?
- What items are difficult for them to answer?
- Are there items that were not clear to the participants?
- What items are left unanswered?

Useful questionnaire design steps

Fourth: Empirical examination and pilot test

- Is there any information that pilot group participants added?
- Are items reliable and valid?
- Do the reliability and validity of the data provide evidence that the questionnaire examines the trait intended by the research question and variables?

Useful questionnaire design steps

Removing items!

ELSEVIER

We would like you to think about the article you most recently reviewed for Advances in Life Course Research. Please indicate the extent to which you agree/disagree with the following statements about your experience.

Invitation process

	Agree strongly	Agree	Neither agree nor disagree	Disagree	Disagree strongly	Don't know/ not applicable
The invitation was polite and courteous	○	○	○	○	○	○
The amount of information I received regarding the article was sufficient for me to decide to review	○	○	○	○	○	○
The article was relevant to my field of expertise	○	○	○	○	○	○
I had sufficient time to decide whether or not to accept the invitation to review	○	○	○	○	○	○
I found it easy to communicate my decision to accept/reject the invitation	○	○	○	○	○	○
I found it straightforward to register as a reviewer for the journal	○	○	○	○	○	○

<< >>

Standardization and styles of interviewing

Set of practices commonly used in surveys conducted to describe a population

- Read questions as written
- Probe inadequate answers nondirectively
- Record answers without discretion
- Be interpersonally non-judgmental regarding the substance of answers

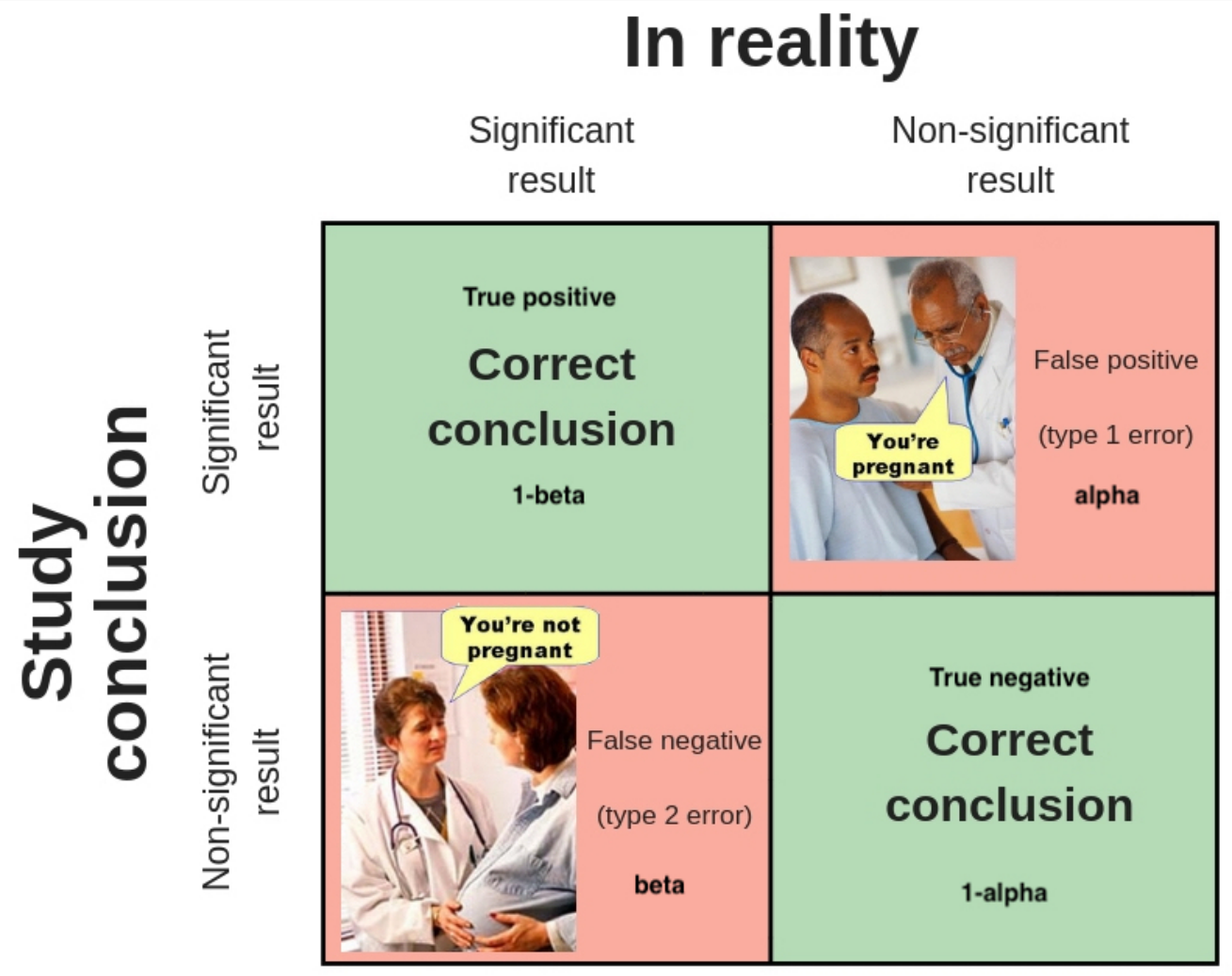
Social desirability by mode

- Sample: recent alumni of a single university
- Interactive Voice Response (IVR)

Table 9. False Negative and False Positive Rates, by Item and Mode

	CATI		IVR		Web		All respondents
True status	False negative (%)	False positive (%)	False negative (%)	False positive (%)	False negative (%)	False positive (%)	Yes (record) (%)
GPA < 2.5	**83.3**	0.0	**69.2**	0.8	**61.5**	0.7	12.1
At least one D or F	**33.0**	3.3	**28.3**	5.4	**19.9**	2.2	61.1
Dropped a class	**34.3**	6.3	**34.2**	9.1	**31.6**	7.6	68.4
Warning or probation	**33.3**	9.1	**33.3**	11.8	**25.0**	12.4	2.4
GPA > 3.5	16.7	**7.4**	19.1	**1.9**	6.8	**6.0**	21.4
Honors	5.3	**5.2**	0.0	**5.7**	2.8	**6.4**	12.1
Donations ever	31.5	**24.3**	25.4	**19.2**	30.5	**20.3**	40.7
Donations in last year	8.8	**25.6**	22.2	**25.9**	20.0	**23.3**	15.7
Member of Alumni Association	2.0	**10.7**	11.1	**10.1**	3.2	**8.1**	16.2

Social desirability by mode



Techniques for sensitive questions

- Indirect question (loading)
- Randomized response technique (RRT)
- Unmatched count technique (UCT)
- Open questions
- Sensitive questions at the end of the questionnaire
- Embedding sensitive items with items even more threatening (from most to least severe)

Field experiences

EcoSocial Study

- EcoSocial Study (Social Cohesion in LA)
- Corporación de Estudios para Latinoamérica (CIEPLAN)
- 7 countries, 2007

Fieldwork

Country	Company	Sample design	Language	Training	Manual	Data entry	Output
Argentina	University	Quota	Spanish	Supervisors and interviewers	Yes	CsPro	Delayed
Brazil	Ipsos	Quota	Portuguese	Supervisors and interviewers	Yes	CsPro	On time
Chile	University	Quota	Spanish	Supervisors and interviewers	Yes	CsPro	On time
Peru	University	Quota	Spanish	Supervisors and interviewers	Yes	CsPro	On time
Colombia	Centro Nacional de Consultoría	Quota	Spanish	Supervisors and interviewers	Yes	CsPro	Delayed
Guatemala	Borges & Asociados	Quota	Spanish	Supervisors and interviewers	Yes	CsPro	On time
Mexico	Ipsos	Quota	Spanish	Supervisors and interviewers	Yes	CsPro	Delayed

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

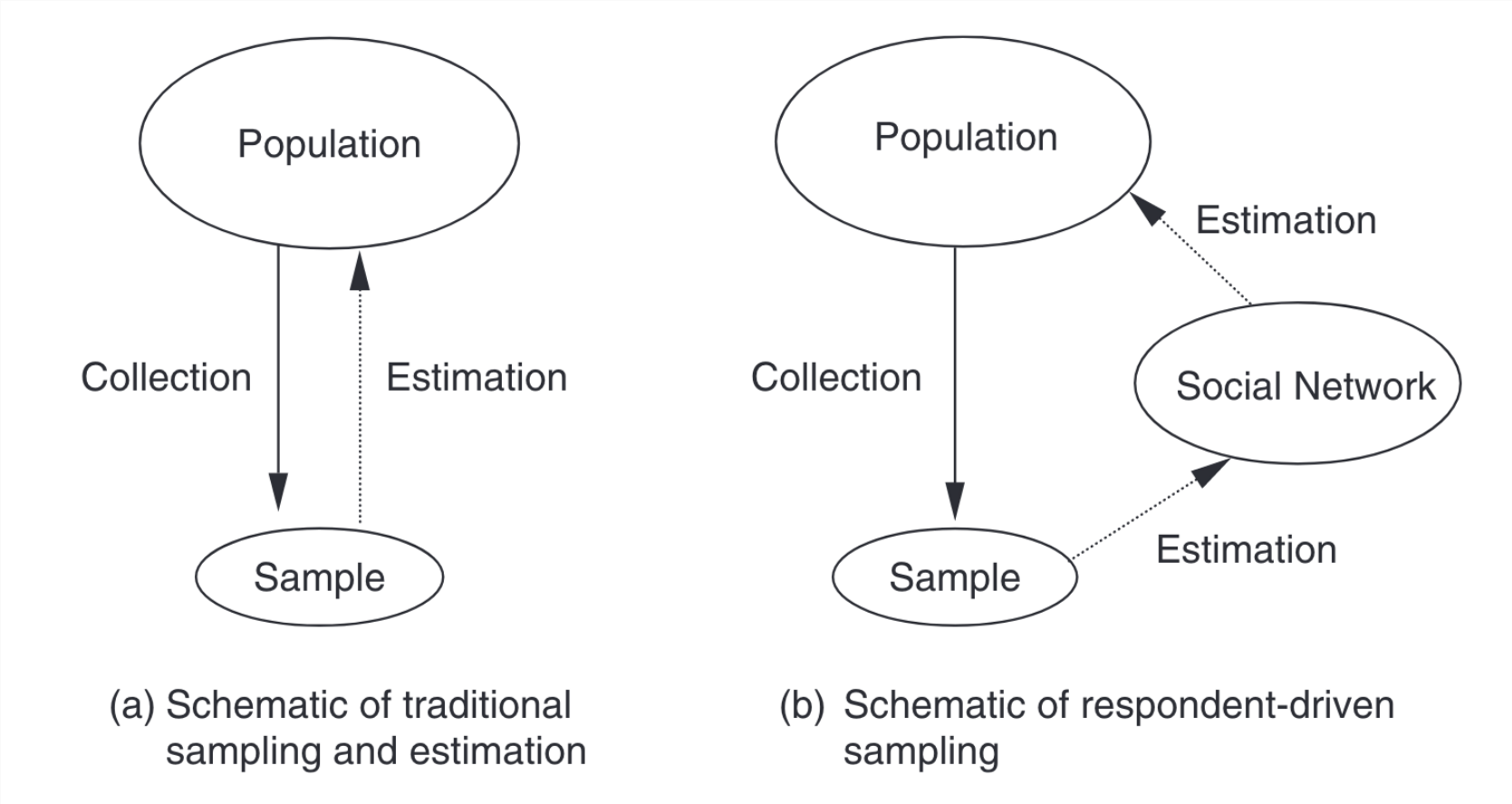
- Convenience sampling
- Quota sampling
- Judgmental or purposive sampling
- Snowball sampling

For instance, street corner interviews will be...

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)



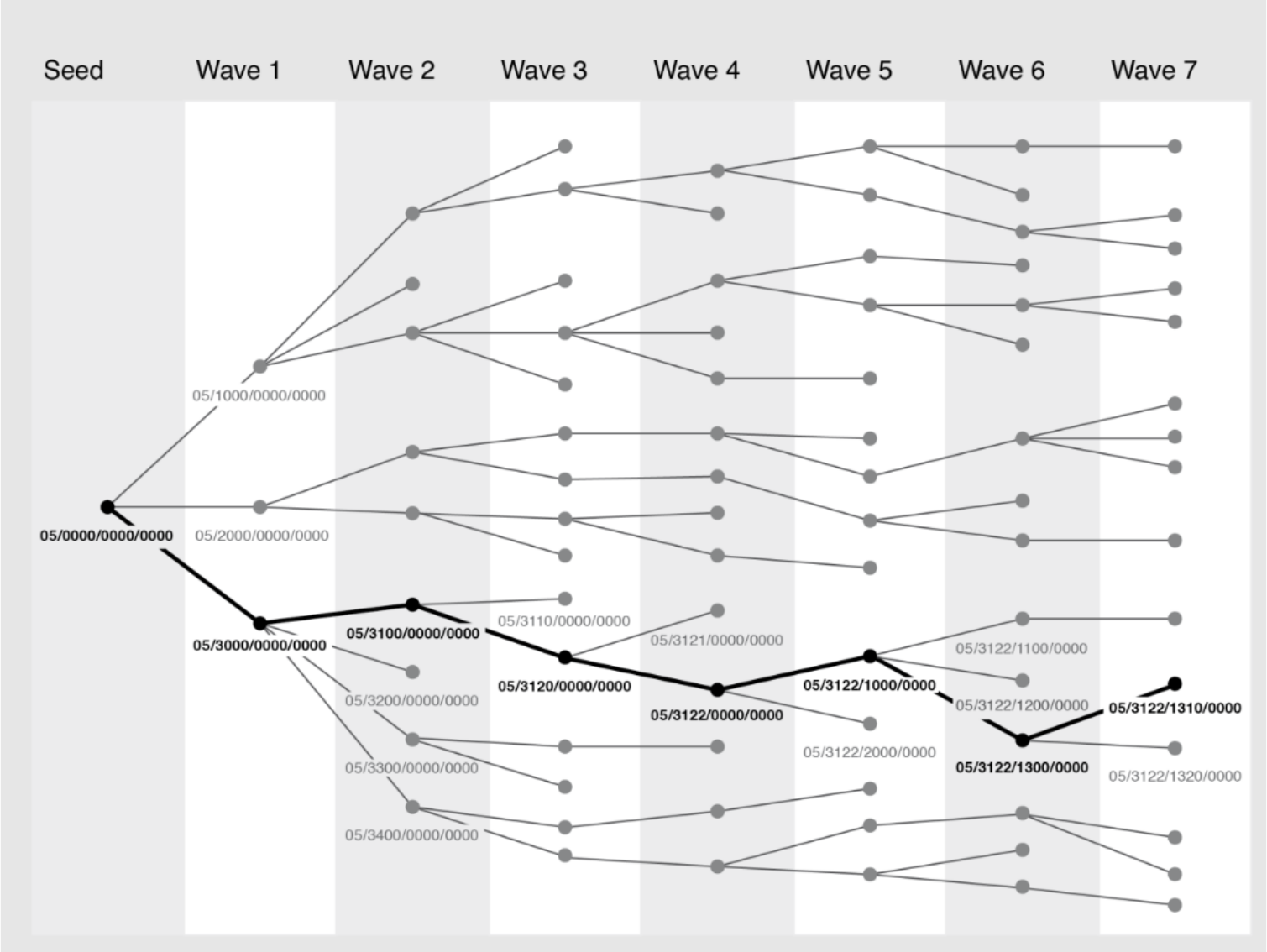
(a) Schematic of traditional sampling and estimation

(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
- Starting with an initial convenience sample (seeds), researchers incentives respondents to recruit their peers to also participate in the survey

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 biased 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 drawn with probability proportional to their degree (the number of connections a node has to other nodes)

RDS - Comparisons

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

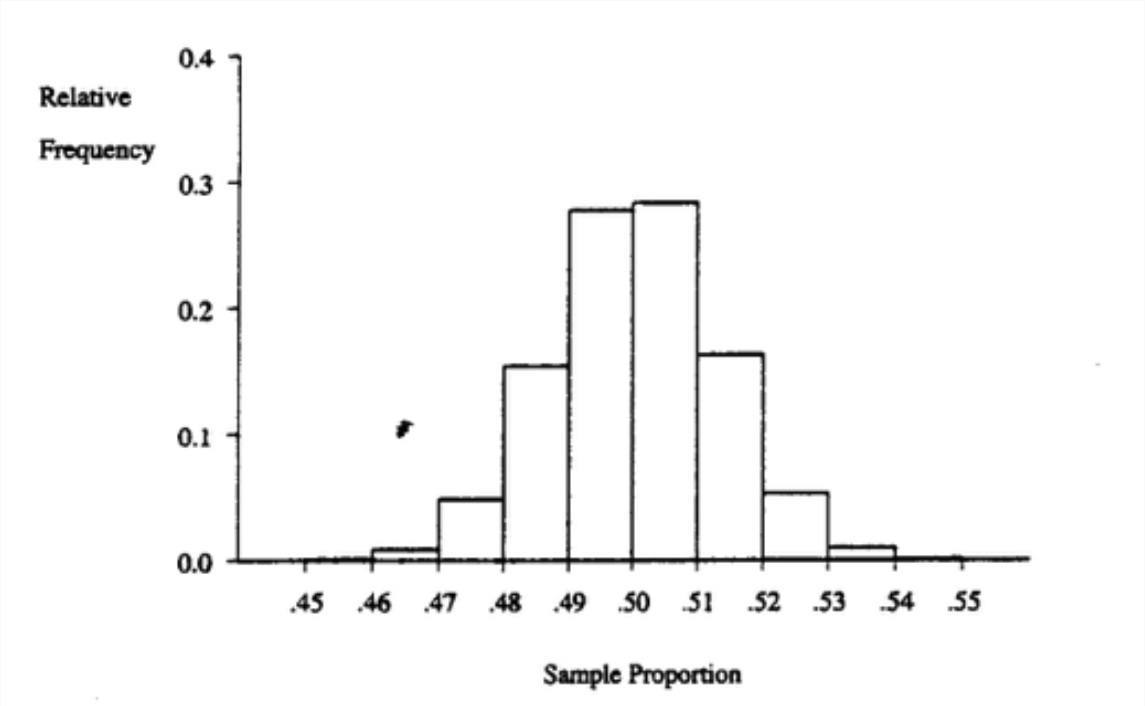
Jazz musicians in New York			
Characteristic	Union Sample ($n = 415$)	Chain-Referral Sample ($n = 251$)	RDS Estimate ($n = 251$)
Union member	100	39.6	25.4
Female	16.1	26.8	23.7
Solo only	19.9	8.8	11.6
Received airplay	79.6	81.6	74.9
Jazz musicians in San Francisco			
Characteristic	Union Sample ($n = 237$)	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.

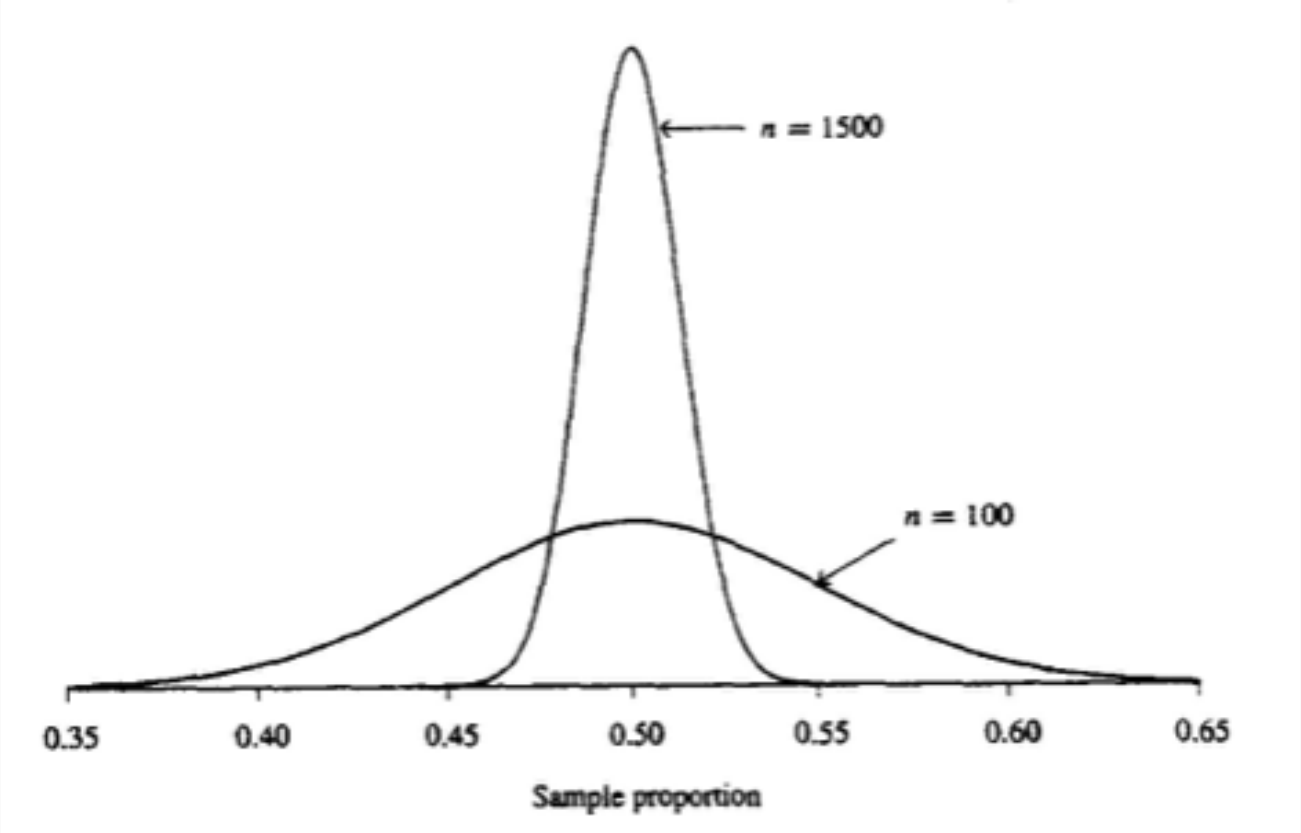
Frequentist inference refreshment

Sampling distribution

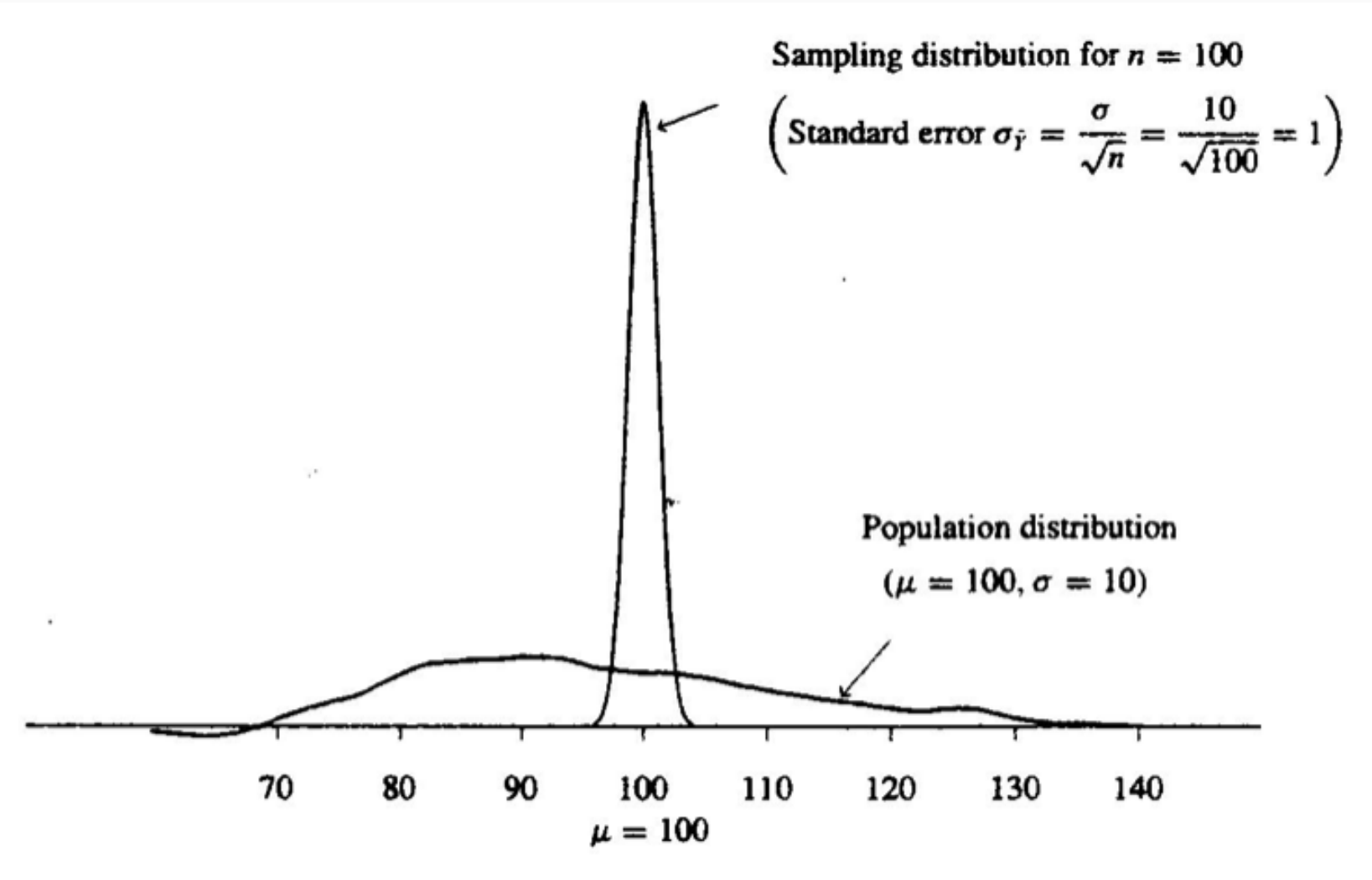
We repeat (random) sampling to get a theoretical set of possible values. But in practice, we only get one sample. We rest on statistical theory and the properties of repeating sampling infinitely



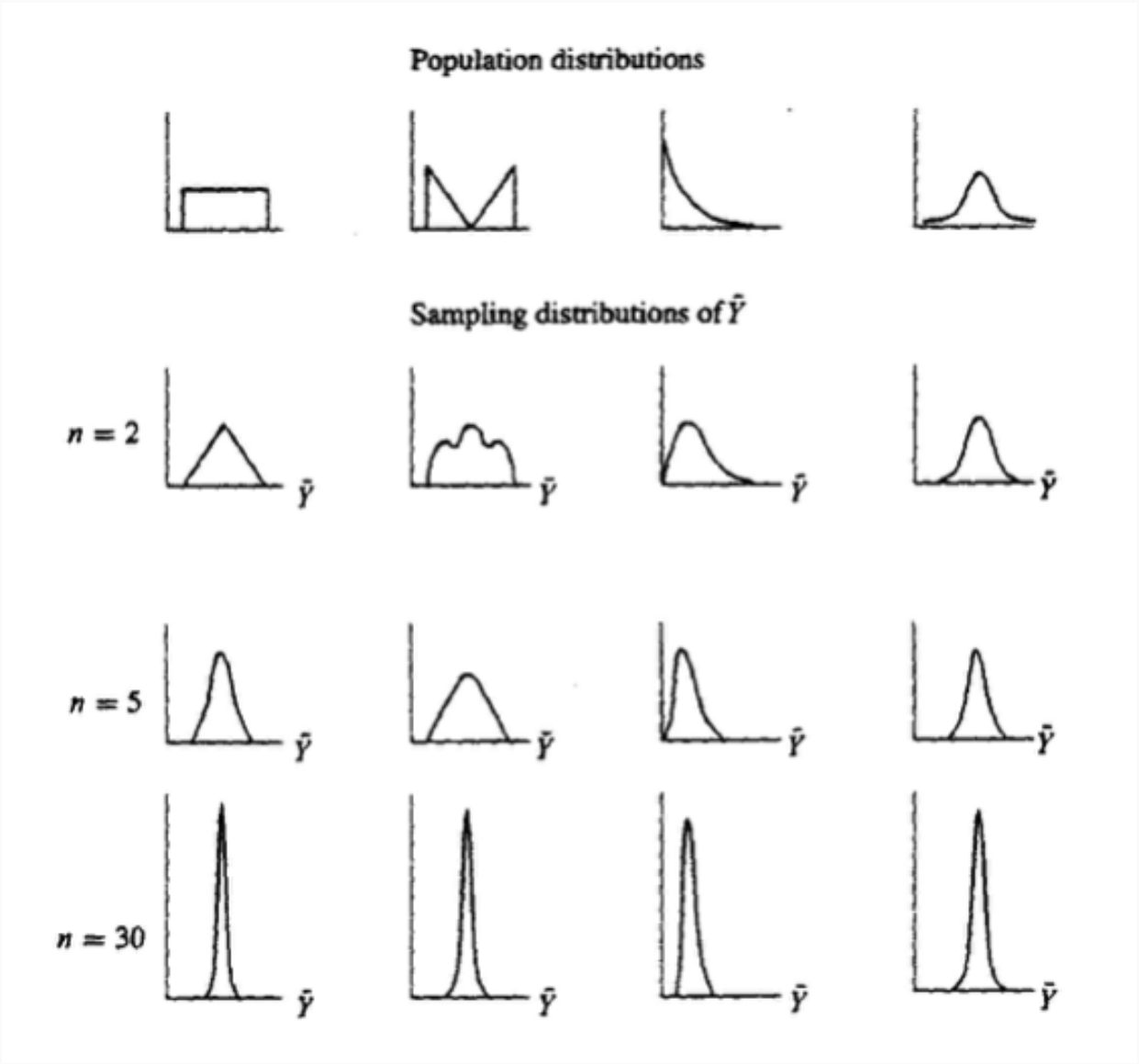
Sampling distribution



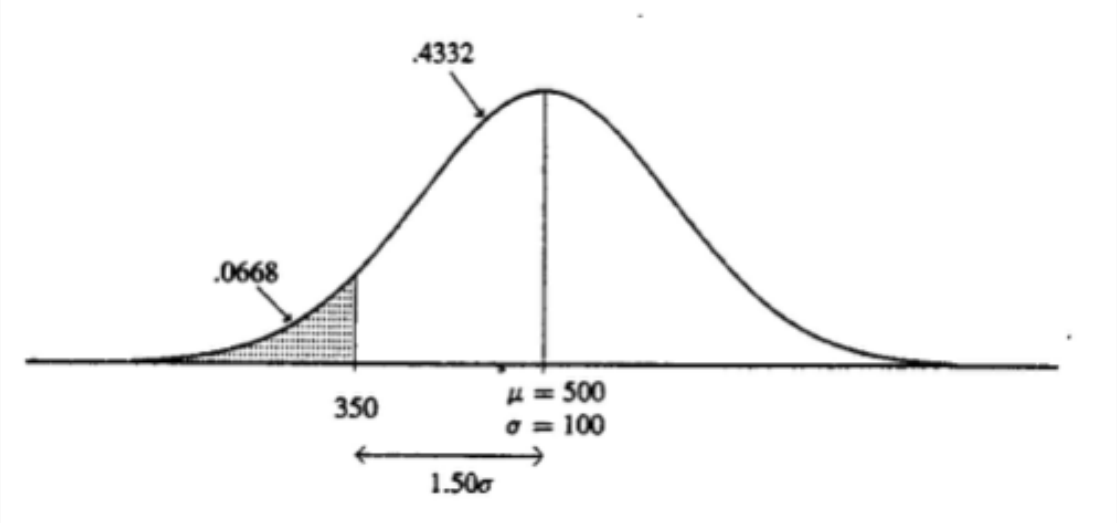
Central limit theorem



Central limit theorem

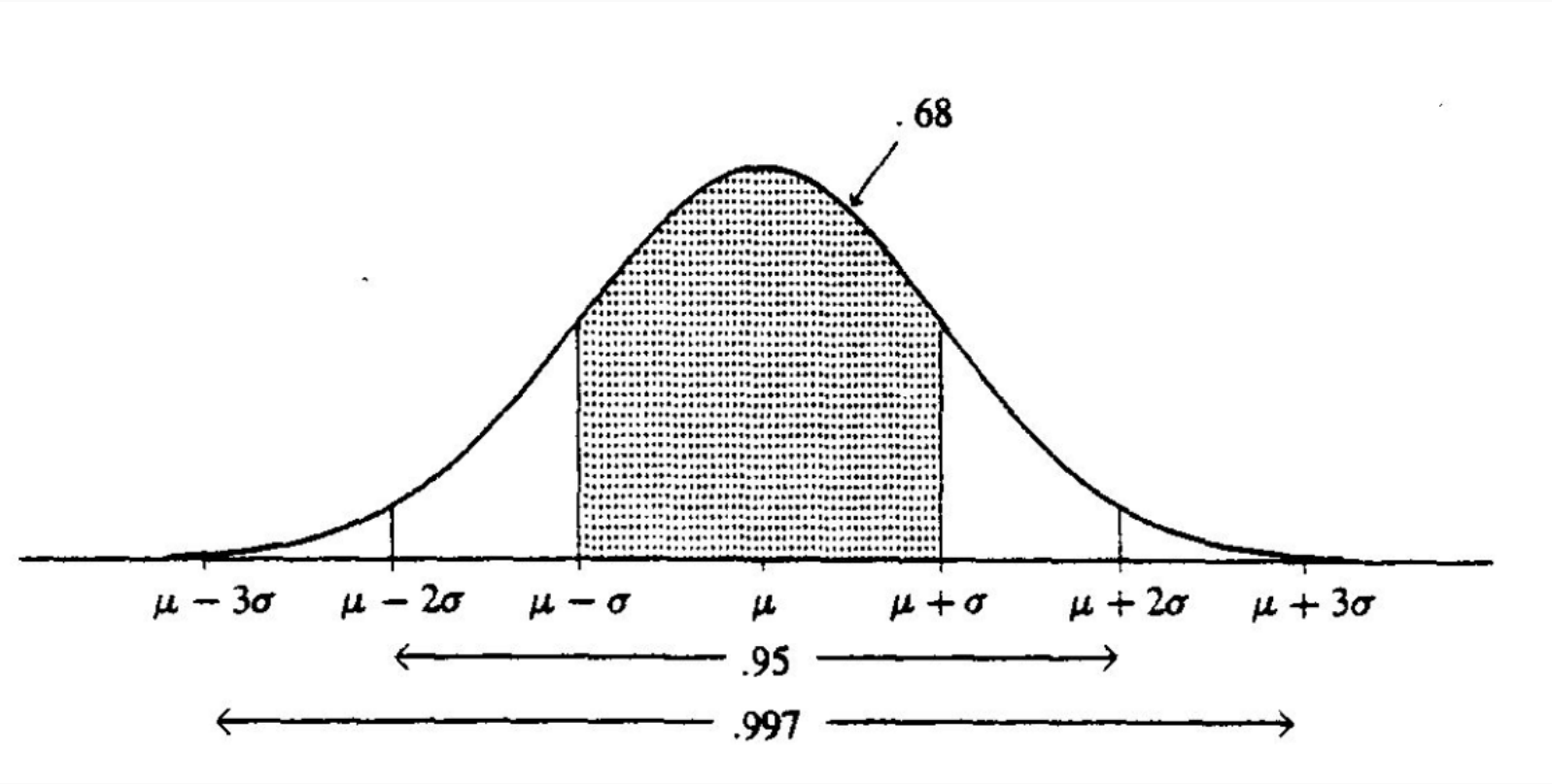


Central limit theorem

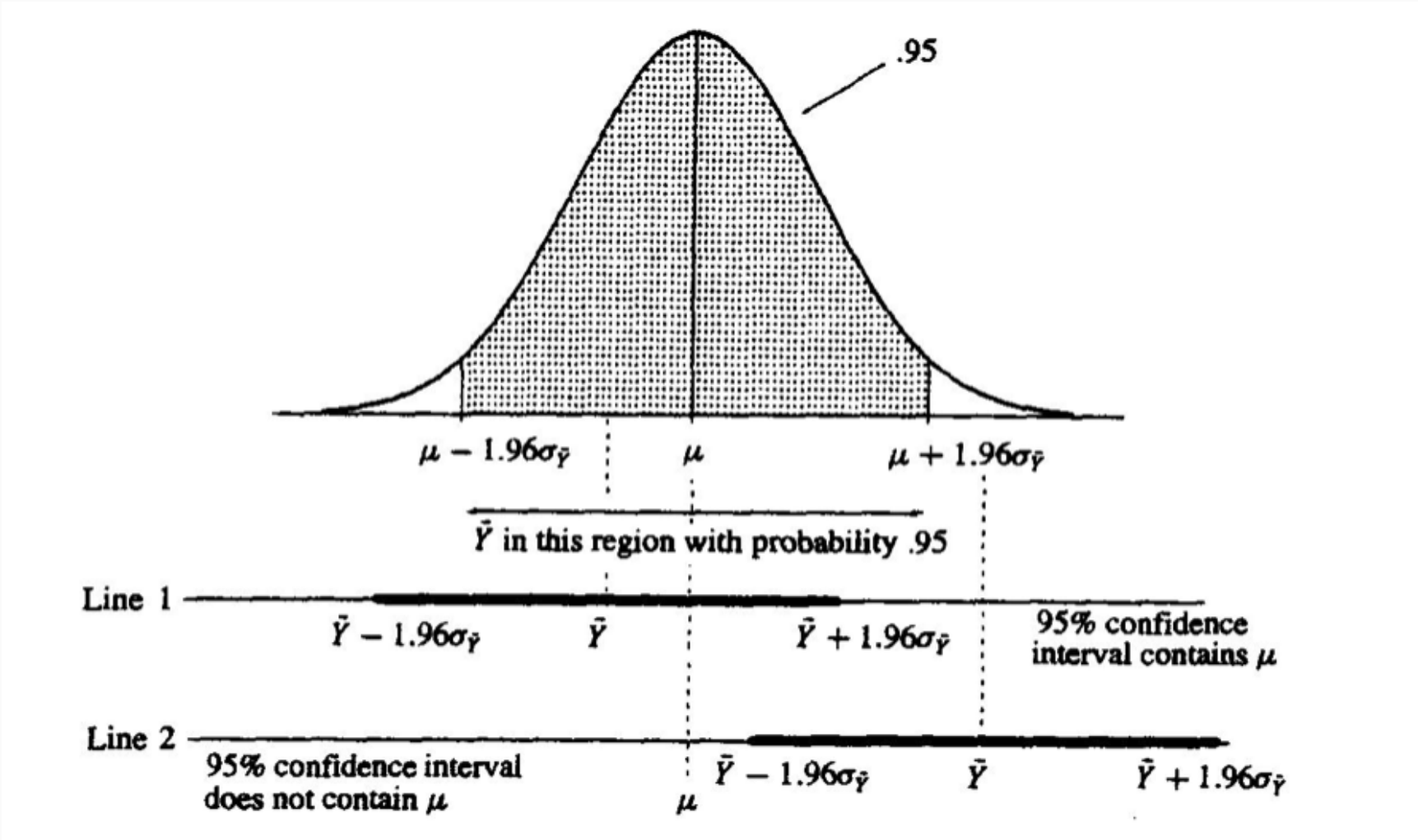



```
z = (350-500)/100
print(z)
[1] -1.5

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

Normal distribution



Confidence interval



Margin of error (MOE)

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

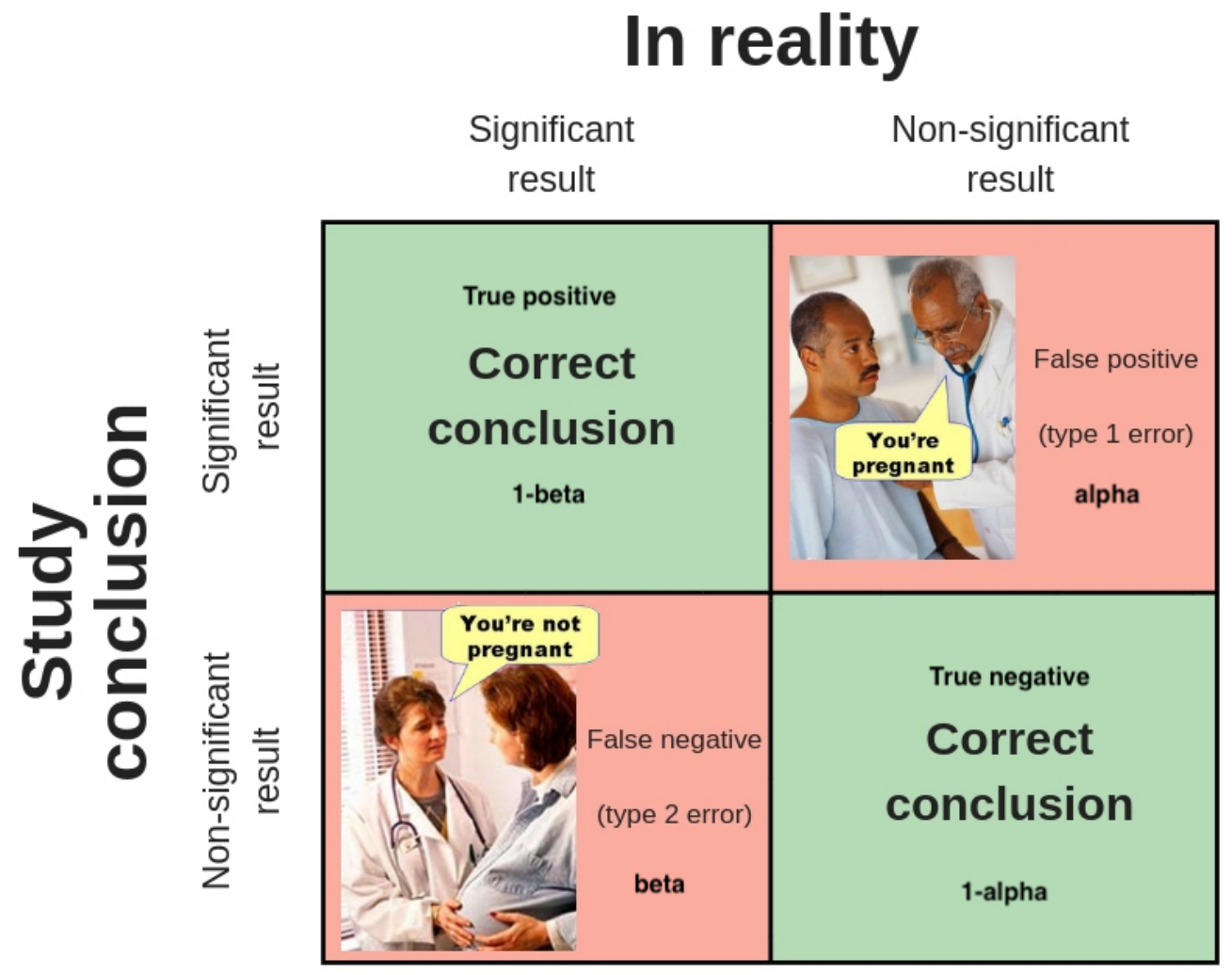
$$MOE = z * SE$$

What is the z-value for 95% confidence?

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

qnorm(0.975)
[1] 1.959964
```

Let’s move a bit to R

Beyond power calculations

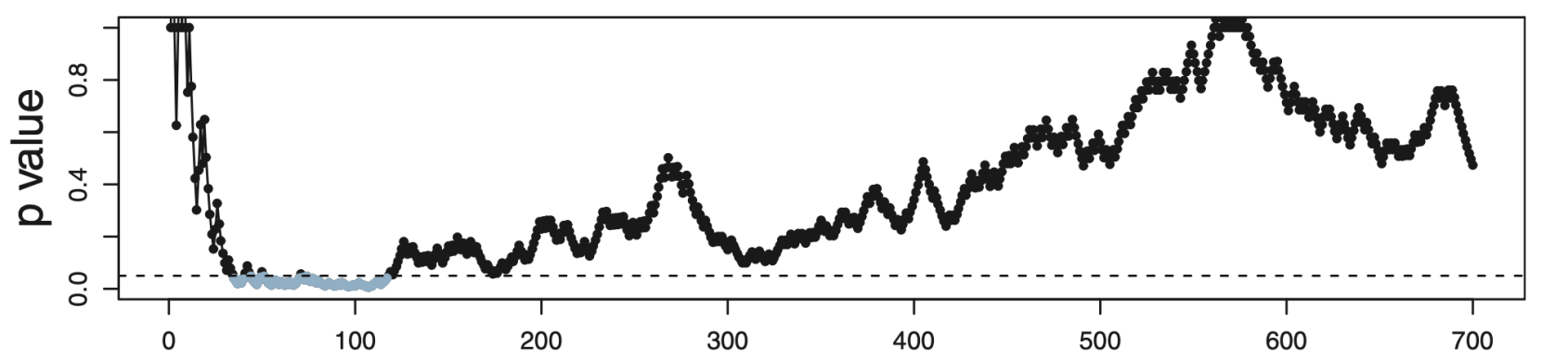


Beyond power calculations

- 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*, i.e., correctly rejecting the null hypothesis. **Power of the test**.

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 $\alpha = 0.05$



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).

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

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
sigma = p * (1-0.5)
print(sigma)
0.25
n = (1.96^2 * sigma) / 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)**

CI's 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

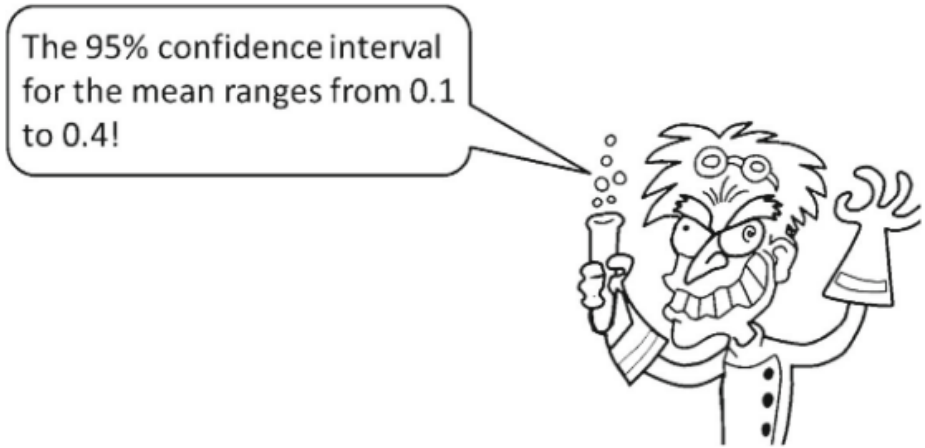
CI's clarification

- CIs do not provide for a statement about the parameter as it relates to the particular sample at hand
- They provide for 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

CI's 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): _____

CI's 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

CI's 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*

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

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

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)	$w_i = \frac{N_r}{N_t}$	$w_i \times n_t$	$\frac{n_t}{r}$
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

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

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? Always!

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 points about are almost always not met

Let’s move a bit to R...

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.

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

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

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
- Two-stage (sub-sampling)
  sample of clusters $\rightarrow$ sample of units
- 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):

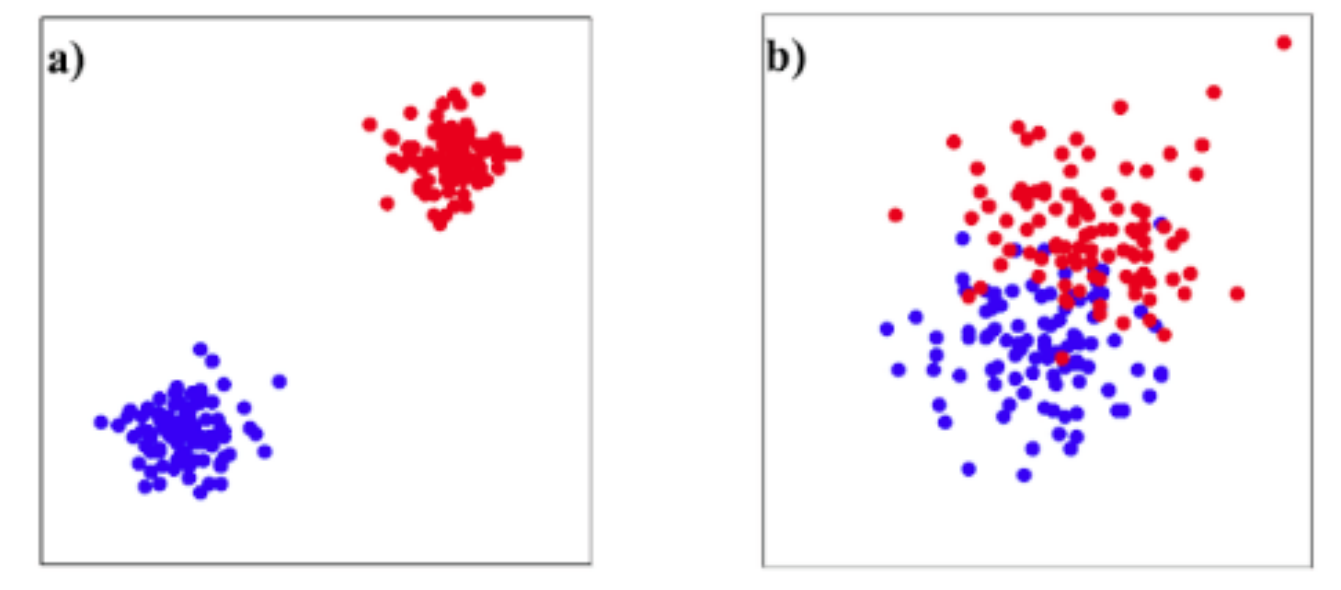
$$Deff = 1 + (n_c - 1)\rho$$

- 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  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[, p := size / sum(size)]
blocks[sample(.N, 3, prob=p)]
blocks[, sp := 3 * p]
```

Let’s move a bit to R...

Weighting and distributions

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 data

```
a = paste0('A', 1:3)
b = paste0('B', 1:3)

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

Create data with joint distributions

```
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(sa=la,sb=lb, joint=joint)
djoint[, total_joint := sum(joint)]

dat = merge(dat, djoint, by=c('sa', 'sb'))
dat[, N := .N, .(sa, sb)]

head(dat, 4)
```

```
       sa     sb joint     N
1:     A1     B1    10    42
2:     A1     B1    10    42
3:     A1     B1    10    42
4:     A1     B1    10    42
```

Weights to emulate joint distribution?

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

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

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

$$w_{a_1 b_1} = 0.5537$$

Weights to emulate joint distribution?

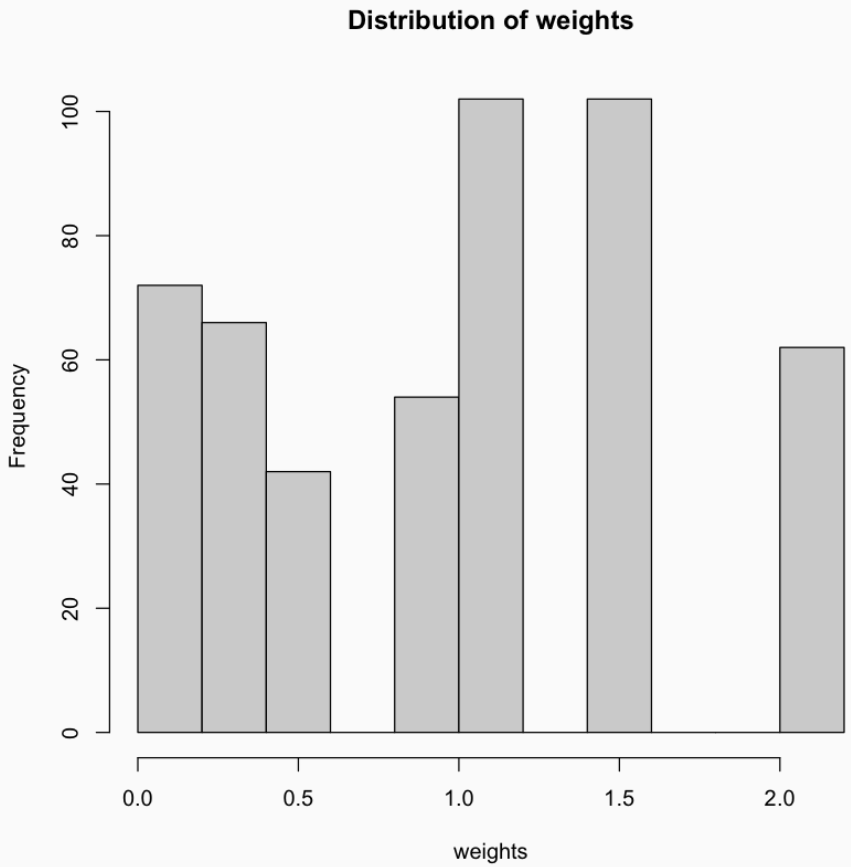
```
dat[, w := joint/total_joint * total_sim/N]

head(dat[, .(sa, sb, w), 4)
       sa     sb         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$sa)
   A1    A2    A3
0.282 0.336 0.382
wpct(dat$sa, dat$w)
       A1        A2        A3
0.3023256 0.4744186 0.2232558

wpct(dat$sb)
   B1    B2    B3
0.332 0.356 0.312
wpct(dat$sb, 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

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

```
PracTools::deffK(dat$w)
```

```
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

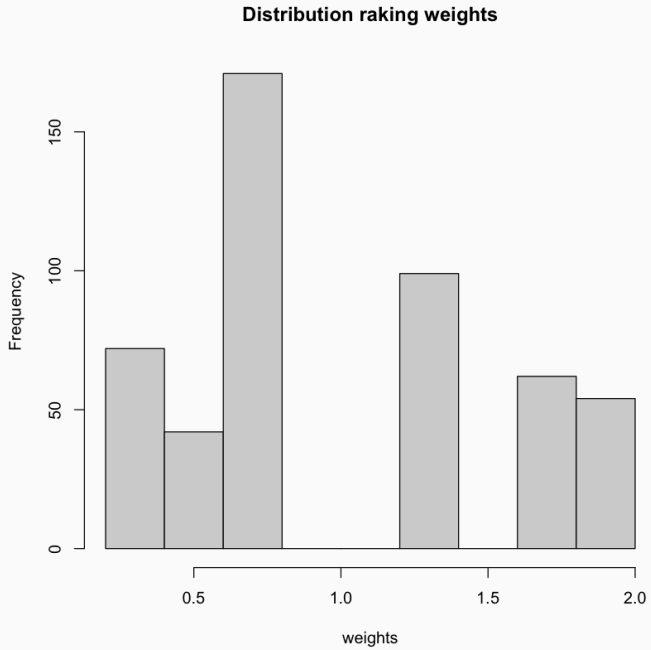
```r
devtools::install_github("sdaza/autumn-adjustments")
library(autumn)

# define targets
target = list(
    sa = c(A1=0.302, A2=0.474, A3=0.223),
    sb = 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
sa	A1	0.282	0.30230	0.30230
sa	A2	0.336	0.47447	0.47447
sa	A3	0.382	0.22322	0.22322
sb	B1	0.332	0.17234	0.17234
sb	B2	0.356	0.41382	0.41382
sb	B3	0.312	0.41382	0.41382

Distribution of weights with raking

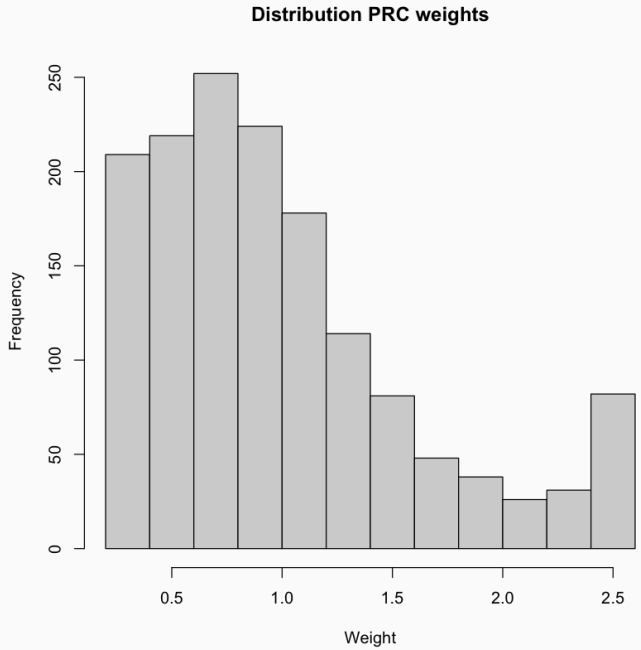



```
design_effect(result$weights)
1.24

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
```

Let’s move a bit to R...

News / Next steps

New paper out!

Social Science Research 110 (2023) 102844

Contents lists available at ScienceDirect

Social Science Research

journal homepage: www.elsevier.com/locate/ssresearch

From prison to work? Job-crime patterns for women in a precarious labor market☆

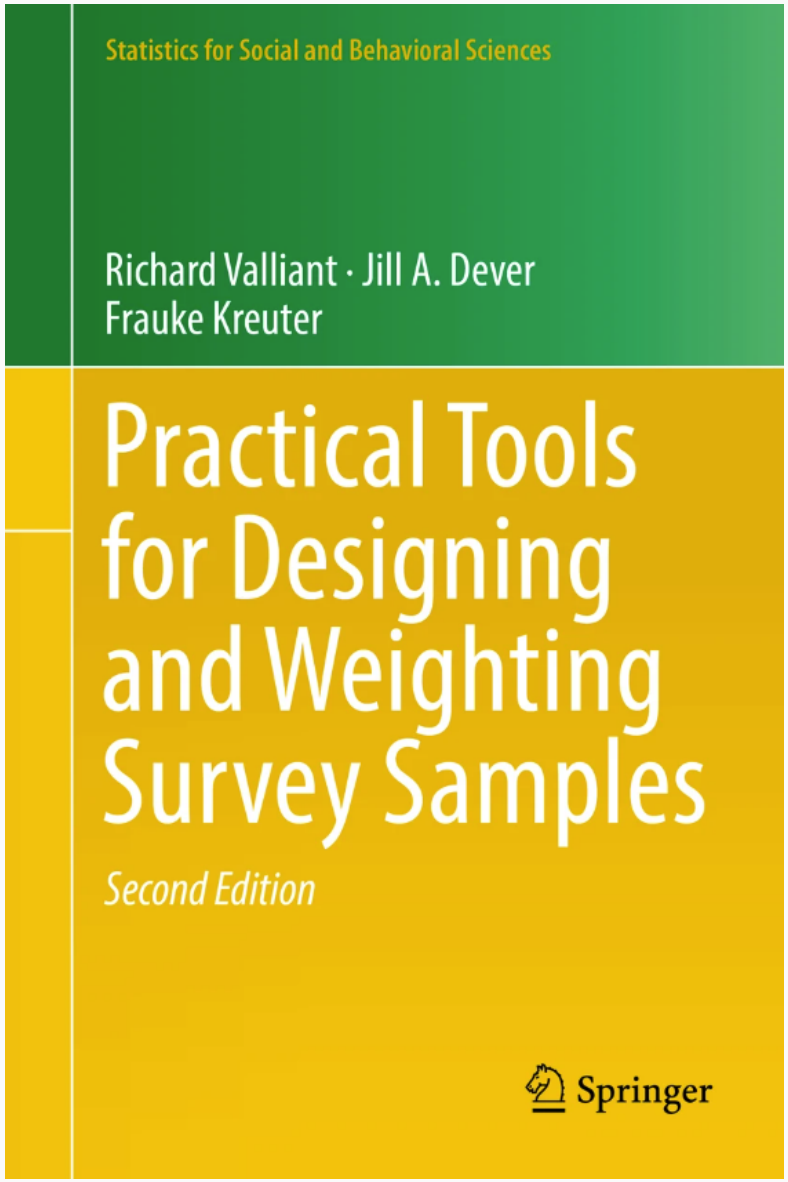
Pilar Larroulet [a,b,*], Sebastian Daza [c,d], Ignacio Bórquez [d,e]

Resources

- GitHub repo will be available and updated (if there are courses)
- Zotero group will be down next month, so please, get books you are interested in)

Training opportunities!

Check out this book!



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

sebastian.daza@gmail.com