Before the Numbers

Quantitative Research Design
from Question to Dataset

VIVEK H. PATIL

MARGIN OF ERROR MEDIA

Table of contents

Copyright 3

Preface 5

Introduction: How Research Works 7

I Foundations 15

1 From a Question to a Measure 17

2 Where Error Comes From 27

II Design and Measurement 37

3 Designs That Support Causal Claims 39

4 Measurement and Scales 49

III Instruments and Samples 59

5 Writing Questions 61

6 Sampling and Coverage 71

IV Data, Analysis, and Reporting 81

7 Collecting and Cleaning Data 83

8 The Analysis Plan and the Report 93

Index of Tools and Data 103

Before the Numbers

For everyone who has been handed a number and asked what it means.

Copyright

Before the Numbers: Quantitative Research Design from Question to Dataset

© 2027 Vivek H. Patil, Ph.D. All rights reserved.

Published by Margin of Error Media LLC, Spokane, Washington.

No part of this book may be reproduced, distributed, or transmitted in any form without the prior written permission of the publisher. Brief quotations in reviews or scholarly work, with full attribution, are the exception.

For permissions, classroom adoption, or bulk purchase, write to patilv@gmail.com.

First edition, 2027.

ISBN, paperback: [ISBN-PENDING]

ISBN, Kindle: [ISBN-PENDING]

Data and Figures

The author drew every figure in this book in R with ggplot2. Each caption names its source. The data files are published with the online edition, and a reader can check any number against them.

Seven cases use public data. Four sources are federal agencies. They are the Department of Housing and Urban Development, the Department of Homeland Security, the Department of Agriculture, and the Census Bureau. Three are research groups. They are the Pew Research Center, the Migration Policy Institute, and the Center for Migration Studies. The resume data are from Bertrand and Mullainathan (2004), and the pay data are from the 1985 Current Population Survey. None of these sources has endorsed this book.

The panel file in Chapter 7 is simulated. The item responses in Chapter 4 are simulated. Both were mathematically modeled to reflect patterns reported in published research, and both are labeled where they appear.

A Note on AI

I wrote this book with a language model as a working partner. I used Claude, made by Anthropic, to draft and revise passages, to check reasoning, and to write the R code behind the figures. It also helped build the tools that go with each chapter.

The judgment in the book is mine. I have checked every number against its source. My views on research design come from years of teaching the topic and from designing studies for people who had to use the results. Where the book is wrong, the fault is mine, and I would like to hear about it at the address above.

Preface

Most courses on data begin with a dataset. This book ends with one.

A dataset is the product of many decisions. A researcher decides what to study and how to define it. The researcher decides whom to ask, what to ask, and how to record the answers. By the time the file is ready for analysis, most of the choices that affect the result have been made. A statistical test cannot repair a question that was worded badly or a sample that left people out.

This book is about those earlier decisions. It follows the order in which a researcher makes them.

How the Book Is Organized

An introduction comes first. It explains what research is, the steps it follows, and where this book fits. Readers who are new to research should start there.

Eight chapters follow in four parts. The parts follow the order of the work.

1. **Foundations.** Chapter 1 turns a question into a measure. Chapter 2 lists the ways a survey number can be wrong.
2. **Design and Measurement.** Chapter 3 covers experiments and other designs. Chapter 4 covers scales, reliability, and validity.
3. **Instruments and Samples.** Chapter 5 covers question writing. Chapter 6 covers sampling and coverage.
4. **Data, Analysis, and Reporting.** Chapter 7 covers data collection and cleaning. Chapter 8 covers the analysis plan and the report.

Each chapter opens with a public case and uses that case throughout. The cases cover homelessness, immigration, hiring discrimination, food insecurity, community health, the census, online panels, and pay. These are topics where a number affects how money is spent or how a group is treated.

What Each Chapter Contains

Each chapter has the same parts. An opening case comes first, followed by the main sections. Key terms are in bold where they first appear, and a glossary at the end of the chapter defines them. Two kinds of boxes appear in each chapter.

- An **Ethics Moment** describes a choice that affects the people being studied.
- An **AI Reality Check** describes what a language model can and cannot do for that step of the work.

Each chapter also has a section called **Work With the Data**. It appears at the point where the idea is taught and is part of the chapter. In it you use a tool on the chapter's data, with numbered steps and a note on what to look for. In the online edition the tool sits on the page. Readers of the print edition can open it in any web browser.

Each chapter ends with exercises at three levels. *Check Your Understanding* covers concepts. *Apply It* asks for a calculation or a design. *Think Deeper* asks for judgment, and those questions do not have one right answer.

What You Need

You need no software to read this book. The tools run in a browser. The data files are plain CSV files, and any spreadsheet program can open them. Readers who use R can find the code for every figure in the online edition.

Some statistics appear in Chapters 4, 6, and 8. Each is explained where it is used. A reader who wants more can turn to *Margin of Error*, the statistics book in this series.

For Instructors

The eight chapters can fill a short course or form one part of a longer course. An instructor guide is available on request. It has a session plan for each chapter, an activity for each tool, student worksheets, and answers to the exercises. Instructors can write to patilv@gmail.com.

The Companion Volume

This book covers research that produces numbers. Its companion, *Beyond the Numbers*, covers research that produces words, such as interviews, focus groups, and written comments. The two books can be read alone or together. Many research questions need both.

Introduction: How Research Works

This book assumes no background in research. This introduction explains what research is, the steps it follows, and where the eight chapters fit. Readers who have conducted a study before can skim it. Readers who are new to research should start here.

Research Starts With a Decision

People in organizations make decisions every day without knowing all the facts. A store owner decides whether to stay open later. A school board decides which program to cut. A county decides where to place a new clinic.

Some of these decisions are small and easy to reverse. Others cost a great deal and affect many people. For the second kind, a guess is a weak basis. The decision is only as good as the evidence behind it.

One example runs through this introduction. It is invented for this book. A community health clinic finds that patients miss 22 percent of their appointments. Each missed visit costs the clinic money and leaves another patient waiting. The director must decide what to do.

She has several ideas. The clinic could send text reminders, add evening hours, or offer rides. Each idea costs money, and she cannot afford all three. She does not know why patients miss their visits, so she does not know which idea would help. She needs information before she can choose.

What Research Is

Research is a planned and careful way of producing information to support a decision. Three features separate it from ordinary fact-finding.

- **It is planned.** The researcher decides what to learn and how to learn it before collecting anything.
- **It is systematic.** The same steps are applied to every case, so one vivid story does not outweigh the rest.
- **It is open to checking.** The researcher records what was done, so another person can judge the work or repeat it.

The clinic director could ask the front desk staff why patients miss visits. That is useful, and it is not yet research. The staff hear from the patients who call to cancel. They do not hear from the patients who never call. A planned study would reach both groups.

Research that serves one decision for one organization is **applied research**. The clinic study is an example. Research that aims to add to general knowledge, with no single decision in view, is **basic research**. A university study of why people forget appointments is an example. Both kinds use the same methods, and this book serves both.

The Research Process

Research is a sequence of steps. Textbooks number the steps in slightly different ways, and the logic is the same in all of them. This book uses eight steps, adapted from Churchill and Iacobucci (2015) and from Zikmund and colleagues (2013).

1. **Formulate the problem.** State the decision and the question that the research must answer.
2. **Estimate the value of the information.** Decide whether the answer is worth the cost of getting it.
3. **Choose the research design.** Decide what kind of study fits the question.
4. **Design the method and the forms.** Decide how to gather the data, and write the questionnaire or other form.
5. **Plan the sample.** Decide whom to study and how to select them.
6. **Collect the data.** Carry out the plan and check the quality of the work as it proceeds.
7. **Prepare, analyze, and interpret the data.** Clean the file, run the analysis, and decide what the results mean.
8. **Write the report.** Tell the decision-maker what was found and how far it can be trusted.

Each step limits the steps that follow. A question that was worded badly in step 4 cannot be repaired by statistics in step 7. A sample that left people out in step 5 cannot be completed in step 8. For this reason, most of the damage to a study is done early, and most of this book is about the early steps.

The steps also loop back. A trial run of a questionnaire can show that the research question itself is unclear. The researcher then returns to step 1. Careful researchers return to earlier steps often.

Figure 1 shows the eight steps and where this book covers each one.

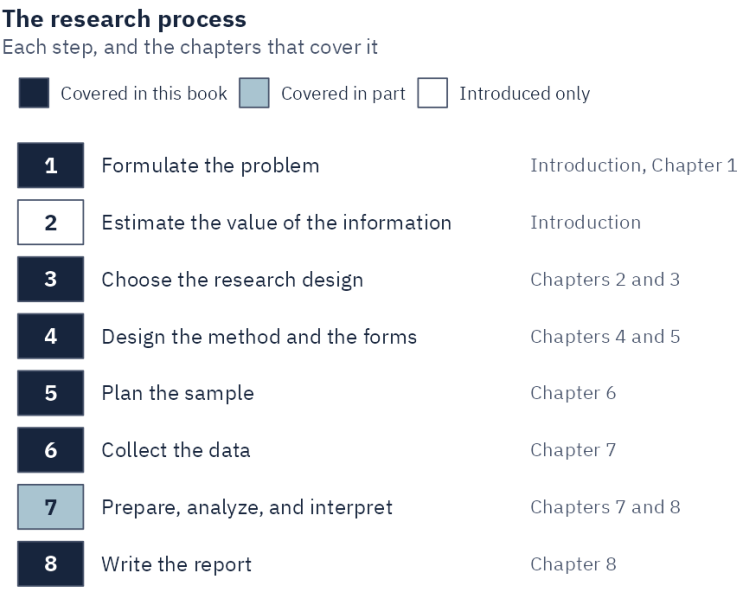


FIGURE 1. *The eight steps of the research process and where this book covers each one.*

From a Decision to a Research Question

The first step is the one most often rushed. A decision-maker usually arrives with a symptom. The clinic's symptom is the 22 percent of visits that are missed. A symptom is a sign that something needs attention. It does not say what the cause is.

The researcher and the decision-maker work from the symptom toward a question that a study can answer. Two statements come out of that work.

A **decision statement** says what the decision-maker has to choose. A **research objective** says what information would help with that choice. Table 1 shows both for the clinic.

TABLE 1. *A decision statement and four research objectives for the clinic example.*

Decision statement	Research objectives
Should the clinic spend its budget on text reminders, evening hours, or rides?	Estimate the share of patients who missed a visit in the past year. Describe which patients miss visits most often. Measure how often each reason for a missed visit occurs. Test whether a text reminder reduces missed visits.

A useful test for an objective is to ask what the decision-maker would do differently with the answer in hand. If no answer would change the decision, the objective can be dropped.

Chapter 1 continues from this point. It shows how each word in a research question has to be defined before anything can be counted.

Is the Research Worth Doing

Research costs time and money. Before starting, the researcher asks whether the information is worth its cost. Four considerations help.

- **The stakes.** A costly decision justifies more research than a minor one.
- **The uncertainty.** Research adds the most when the decision-maker does not know which way to go.
- **The reversibility.** A decision that is hard to undo deserves more evidence than one that can be changed next month.
- **The information that already exists.** Someone may have collected the answer already.

The last consideration deserves a note. Data that someone else collected for another purpose are **secondary data**. Census tables, agency reports, and an organization's own records are examples. Data collected for the study in hand are **primary data**. A careful researcher looks for secondary data first, because they cost less and take less time. The clinic's appointment records already show how many visits are missed and by whom. They do not show why.

This book is mainly about primary data. Several of its cases use published data, and those chapters show how to judge numbers that someone else produced.

Three Kinds of Design

A **research design** is the overall plan for a study. The right design depends on how much is already known about the problem. There are three main kinds.

Exploratory research is for a problem that is not yet understood. The researcher does not know which questions to ask. The methods are flexible, and the samples are small. The clinic could begin by interviewing 15 patients who missed visits.

Descriptive research is for a problem that is understood well enough to measure. It answers questions about how many, how often, and who. The methods are structured, and the samples are large enough to represent a population. The clinic could survey 400 patients to learn how common each reason is.

Causal research tests whether one thing produces a change in another. Its main tool is the experiment. The clinic could send reminders to a randomly chosen half of its patients and compare missed visits in the two halves.

Many projects use all three in that order. The interviews show which reasons exist. The survey shows how common each one is. The experiment shows whether a remedy works. Chapter 3 covers designs in detail.

Words and Numbers

Research data come in two broad forms. **Qualitative research** collects words, such as what people say in an interview. **Quantitative research** collects numbers, such as answers to a survey with fixed choices. Table 2 compares them.

Table 2. *Qualitative and quantitative research compared.*

	Qualitative research	Quantitative research
Typical question	Why, and how	How many, and how much
Questions asked	Open, with follow-up	Fixed, the same for everyone
Sample	Small, chosen for what each person can explain	Large, chosen to represent a population
Data	Words	Numbers
Analysis	Reading, coding, and comparison	Counting and statistics
Common use	Exploratory research	Descriptive and causal research

Neither form is better than the other. Each answers a different kind of question. The clinic needs interviews to learn the reasons for missed visits, and it needs a survey to learn how common each reason is. A study that uses both is a **mixed methods** study.

This book covers quantitative research. Its companion, *Beyond the Numbers*, covers qualitative research. Each book can be read without the other.

The People Involved

Three parties take part in most studies, and each has duties to the others.

- **Participants** give their time and their information. They are owed an honest account of the study, a free choice about taking part, and protection of what they share.
- **The researcher** owes the sponsor careful work and an accurate report, including results the sponsor did not hope for.
- **The sponsor** is the person or organization that asked for the study. The sponsor owes the researcher an honest statement of the decision and must not ask for a result fixed in advance.

Every chapter has an Ethics Moment that returns to these duties at the step where they apply.

What This Book Covers

This book follows the research process from the question to the dataset. Its title, *Before the Numbers*, refers to the decisions that come before any statistic is computed.

- **Foundations.** Chapter 1 turns a question into something that can be measured. Chapter 2 explains the ways a survey number can differ from the truth.
- **Design and Measurement.** Chapter 3 covers experiments and other designs. Chapter 4 covers how to measure something that cannot be seen directly.
- **Instruments and Samples.** Chapter 5 covers how to write questions. Chapter 6 covers whom to ask.
- **Data, Analysis, and Reporting.** Chapter 7 covers collecting and cleaning the data. Chapter 8 covers the analysis plan and the report.

Three topics are left to other books. Statistical tests belong to *Margin of Error*, the statistics book in this series. Interviews, focus groups, and the analysis of text belong to *Beyond the Numbers*. A literature review is a search of published studies. This book mentions it and does not teach it.

Each chapter opens with a note that names the step of the process it covers. Each chapter then works through one public case, and Chapter 1 begins with a count of homelessness on a single night in January.

Key Terms

- **Applied research.** Research that serves a specific decision.
- **Basic research.** Research that aims to add to general knowledge.
- **Causal research.** Research that tests whether one thing produces a change in another.
- **Decision statement.** A statement of what the decision-maker has to choose.
- **Descriptive research.** Research that measures how many, how often, and who.
- **Exploratory research.** Research on a problem that is not yet understood.
- **Mixed methods.** A study that uses qualitative and quantitative research together.
- **Primary data.** Data collected for the study in hand.
- **Qualitative research.** Research that collects words.

- **Quantitative research.** Research that collects numbers.
- **Research.** A planned and careful way of producing information to support a decision.
- **Research design.** The overall plan for a study.
- **Research objective.** A statement of the information that would help with a decision.
- **Secondary data.** Data that someone else collected for another purpose.

Further Reading

- Churchill, G. A., and Iacobucci, D. (2015). *Marketing Research: Methodological Foundations.*
- Zikmund, W. G., Babin, B. J., Carr, J. C., and Griffin, M. (2013). *Business Research Methods* (9th ed.). Cengage.

PART I

Foundations

CHAPTER 1

From a Question to a Measure

Where this chapter fits. Step 1 of the research process is to formulate the problem. The Introduction describes all eight steps. This chapter takes the research question from step 1 and defines each term in it so that it can be measured.

1.1 One Night in January

On one night in the last ten days of each January, volunteers and outreach workers across the United States count people who are homeless. They visit shelters. They also walk through parks, check cars and campgrounds, and look under bridges. The US Department of Housing and Urban Development, known as HUD, collects the local totals and reports a national number to Congress. This is the **point-in-time count**.

The count for January 2024 was 771,480 people. It was the highest total since the counts began in 2007. The count for January 2025 was 745,652. Figure 1.1 shows five of the annual totals.

Most news reports treat this number as the number of homeless people in the country. It is a narrower figure than that. It is the number of people who met one definition, in the places that were searched, on one night. A second federal count uses a second definition and gives a larger total. Public schools report the number of enrolled students who lack a fixed, regular, and adequate place to sleep at night. The school count includes children whose families have moved in with relatives or friends after losing their housing. The HUD count leaves those children out. In recent school years, the school count has been above 1 million students.

Neither count is wrong. They answer different questions. A researcher who wants to study homelessness has to decide which question to ask before any counting begins. This chapter is about that decision.

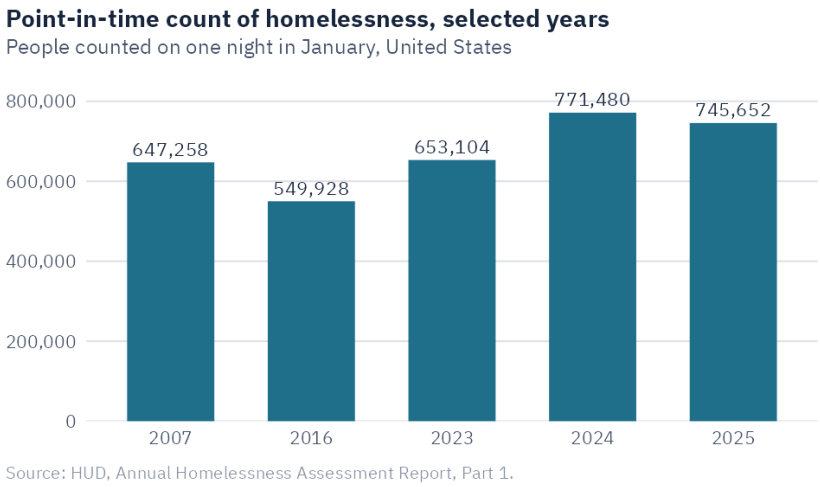


FIGURE 1.1. *People counted as homeless on one night in January, United States, selected years. The 2016 count was the lowest in the series and the 2024 count was the highest. Source: HUD, Annual Homelessness Assessment Report to Congress, Part 1.*

1.2 Start With the Decision

A study begins with a problem that someone needs to act on. A city council may need to decide how many shelter beds to fund. A school district may need to decide how many staff to assign to students without stable housing. A foundation may need to decide whether a housing program worked.

Each of those decisions needs a different number. The council needs to know how many people need a bed on a cold night. The district needs to know how many students lack stable housing during the school year. The foundation needs to know how many people in the program were housed a year later.

A **research question** states what the study will find out. A good research question names three things.

1. **The population.** Who or what the study is about.
2. **The concept.** What the study will measure or compare.
3. **The time and place.** When and where the answer applies.

"How many people are homeless?" names none of the three with enough detail to act on. "How many people in the county slept in a shelter or outdoors on the night of January 28?" names all three. The second question can be answered. The first one cannot be answered until someone makes it more specific.

A Test for a Research Question

Ask what you would do if the answer were high and what you would do if it were low. If the two actions are the same, the question does not help with the decision. Rewrite it until the answer would change what someone does.

1.3 Concepts and Definitions

"Homelessness" is a **construct**. A construct is an idea that cannot be observed directly. Poverty, health, satisfaction, and intelligence are constructs. A researcher can observe where a person slept last night. A researcher cannot observe "homelessness" without first deciding what the word covers.

That decision has two steps.

The first step is the **conceptual definition**. It states in words what the construct means. Federal law defines a homeless person as someone who lacks a fixed, regular, and adequate nighttime residence. That sentence is a conceptual definition. It is clear enough to discuss and too general to count with.

The second step is the **operational definition**. It states the procedure that will be used to decide whether a case fits. HUD's operational definition has two parts. A person is counted as *sheltered* if the person is staying in an emergency shelter, a transitional housing program, or a safe haven. A person is counted as *unsheltered* if the person is sleeping in a place that is not meant for living. Examples are a car, a park, and an abandoned building.

The operational definition is what produces the number. Two studies can share a conceptual definition and still report different results if their procedures differ.

Table 1.1 compares three operational definitions of homelessness that federal agencies use.

Table 1.1. *Three operational definitions of homelessness used by federal agencies.*

	HUD point-in-time count	HUD annual shelter estimate	Public school count
Who is counted	People in shelters and people sleeping in places not meant for living	People who used a shelter program	Enrolled students without a fixed, regular, and adequate place to sleep
Includes families staying with others	No	No	Yes
Includes people sleeping outdoors	Yes	No	Yes
Time covered	One night in January	Any time in one year	Any time in one school year

Who collects it	Local planning groups, with volunteers	Shelter record systems	School district staff

Each column is a reasonable way to count. Each one also leaves people out. The point-in-time count leaves out families staying with relatives. The shelter estimate leaves out people who never enter a shelter. The school count leaves out adults and children who are not enrolled.

1.4 A Count on One Night and a Count Over a Year

The point-in-time count and the annual shelter estimate differ in a second way. One covers a single night and the other covers a year.

A count at one moment is a **point-in-time measure**. It shows how many people are in a condition at that moment. A count over an interval is a **period measure**. It shows how many people were in the condition at any time during the interval.

The two can differ by a large amount. Suppose a shelter has 100 beds and every bed is full every night. A count on any single night finds 100 people. If the average stay is one month, about 1,200 different people use the shelter in a year. Both numbers are correct. The one-night count describes how much capacity the shelter needs. The annual count describes how many people the problem reaches.

A one-night count also contains a higher share of people with long stays. A person who is homeless all year is present on every night of the year. A person who is homeless for two weeks is present on only 14 of the 365 nights. A count on one night is therefore more likely to include the first person than the second. A reader who uses a one-night count to describe "the typical homeless person" will describe long stays as more common than they are.

1.5 Work With the Data: One Night or One Year

The shelter example above used one length of stay. The tool lets you change the stays and see both counts at once. It models a shelter that is full every night.

The tool runs in a web browser at quant.marginoferrormedia.com/tools/count-simulator.html. It needs no account and no software.

1. Leave the shelter at 100 beds. Read the two counts. The one-night count equals the number of beds.
2. Set the share of beds used by people with long stays to 50 percent. On one night, half the people counted have long stays. Compare that with the count over the year.
3. Shorten the short stay to 7 days. The count over the year rises. The one-night count does not change.

4. Set the long-stay share to 0 percent, then to 100 percent. Note which count moves.

The one-night count depends only on the number of beds. The count over the year depends on how long people stay. This is why the two federal counts in this chapter can both be correct and still differ by hundreds of thousands.

The counts used in this chapter are in the file `ch01-pit-counts.csv` at quant.marginoferrormedia.com/data.

1.6 Variables, Units, and Levels of Measurement

Once the definition is set, the study records it as a **variable**. A variable is a characteristic that can take different values for different cases. Housing status on the night of the count is a variable. Its values are sheltered and unsheltered.

The **unit of analysis** is the thing each row of the dataset describes. In the point-in-time count, the unit is a person. HUD also reports results by household and by local area. A researcher has to choose the unit before collecting data, because a file of people cannot always be turned into a file of households later.

Variables differ in how much information their values hold. The standard classification has four **levels of measurement**.

1. **Nominal.** The values are names with no order. Sheltered and unsheltered are nominal. So are the names of states.
2. **Ordinal.** The values have an order, and the distances between them are unknown. “Less than a month, one to six months, more than six months” is ordinal.
3. **Interval.** The values have equal distances and no true zero. Temperature in degrees Fahrenheit is the usual example.
4. **Ratio.** The values have equal distances and a true zero. The number of nights a person spent in a shelter is a ratio variable.

The level of measurement limits what can be calculated. A mean requires interval or ratio data. A median requires at least ordinal data. Nominal data can only be counted. Collect each variable at the highest level that is practical. A researcher who records the exact number of nights can group the values later. A researcher who records only “short” or “long” cannot recover the exact number.

1.7 How the Procedure Shapes the Number

Figure 1.2 divides the two most recent counts into their two parts. The two parts are collected in different ways, and they are not equally accurate.

The sheltered count comes from records. Shelters know who slept there. The unsheltered count comes from a search. Volunteers have a few hours to find

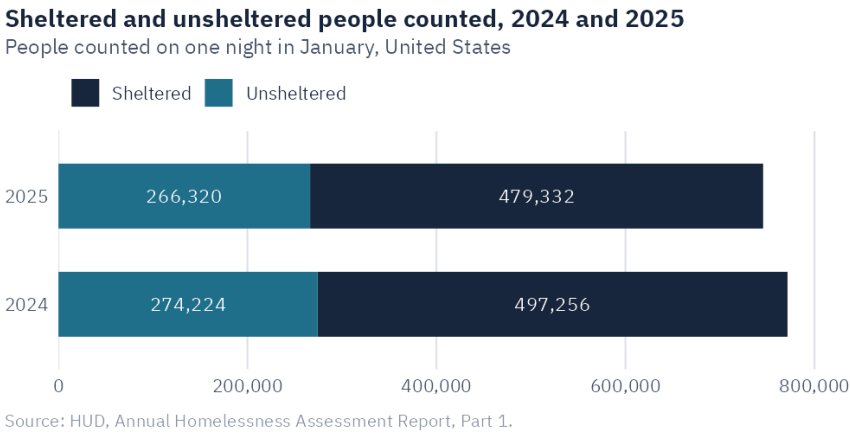


FIGURE 1.2. *People counted as sheltered and unsheltered in the 2024 and 2025 point-in-time counts. About 36 percent of the people counted in each year were unsheltered. Source: HUD, Annual Homelessness Assessment Report, Part 1.*

people who may not want to be found. HUD's own report says the true number of unsheltered people could be larger than the number reported.

The procedure also varies from place to place. HUD requires a sheltered count every year and an unsheltered count every second year. Most local areas conduct both every year. In 2024, 22 of almost 400 local areas did not conduct an unsheltered count. In 2025, 29 did not. In 2021, during the pandemic, only the sheltered count took place. A change in the national total can therefore have two causes. One is a change in the number of people. The other is a change in the procedure.

These details are part of the operational definition. A reader who knows them can use the number well. The count is a reasonable guide to how many shelter beds are in use. It is a weaker guide to how many people sleep outdoors. It gives no information about families who are staying with relatives.

A measure that stands in for a construct it does not fully capture is a **proxy**. Every count in Table 1.1 is a proxy for homelessness. The useful question is which part of the construct each proxy captures and which part it leaves out.

Ethics Moment: A Definition Decides Who Gets Help

Federal housing funds are distributed in part by the point-in-time count. A family that has moved in with relatives does not meet HUD's definition and is not counted. That family is also not eligible for some programs that the count helps to fund. A definition can therefore be an eligibility rule as well as a research choice. Before you adopt a definition, list who it leaves

out. Then ask whether those people would agree that they are outside the problem you are studying.

AI Reality Check: Ask Which Definition

A language model will define "homelessness" in a clear sentence if you ask. It will usually not tell you which agency's definition it used, and it may combine two of them. The same is true for poverty, unemployment, and food insecurity. Each has more than one official definition. Ask the model to name the source of the definition and the procedure used to count. Then read the source. A fluent sentence is not evidence that the definition matches the number you plan to cite.

1.8 Putting It All Together

A research question names a population, a concept, and a time and place. The concept is usually a construct that cannot be observed directly. A conceptual definition states what the construct means. An operational definition states the procedure for deciding whether a case fits, and the procedure is what produces the number.

The homelessness counts show how much the procedure matters. HUD's one-night count, its annual shelter estimate, and the public school count all measure homelessness. They cover different people over different lengths of time, and their totals differ by hundreds of thousands. None of them can be used well without knowing how it was collected.

1.9 Looking Ahead

This chapter treated each count as if it measured its own definition accurately. Chapter 2 drops that assumption. It uses four estimates of one population to show how a survey number can differ from the truth. It gives each kind of error a name.

1.10 Further Reading and References

- US Department of Housing and Urban Development (2024). *The 2024 Annual Homelessness Assessment Report to Congress, Part 1: Point-in-Time Estimates of Homelessness.* The source for the 2024 count and for the description of how the count is conducted.

- Perl, L. (2026). *Homelessness.* Congressional Research Service, In Focus IF12985. A two-page summary of the counts from 2007 to 2025 and of the federal definition.
- Babbie, E. (2020). *The Practice of Social Research* (15th ed.). Cengage. Chapter 5 covers conceptual and operational definitions in more depth.
- Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*(2684), 677 to 680. The original statement of the four levels of measurement.

1.11 Key Terms

- **Conceptual definition.** A statement in words of what a construct means.
- **Construct.** An idea that cannot be observed directly, such as homelessness or health.
- **Levels of measurement.** Nominal, ordinal, interval, and ratio. The level limits which statistics can be calculated.
- **Operational definition.** The procedure used to decide whether a case fits a construct.
- **Period measure.** A count of cases in a condition at any time during an interval.
- **Point-in-time count.** HUD's count of people who are homeless on one night in January.
- **Point-in-time measure.** A count of cases in a condition at one moment.
- **Proxy.** A measure that stands in for a construct it does not fully capture.
- **Research question.** A statement of what a study will find out, naming the population, the concept, and the time and place.
- **Unit of analysis.** The thing that each row of a dataset describes.
- **Variable.** A characteristic that can take different values for different cases.

1.12 Exercises

Check Your Understanding

1. State the difference between a conceptual definition and an operational definition. Give one example of each for "unemployment."
2. A shelter has 60 beds, all full each night, and the average stay is two months. About how many different people use it in a year? Which number would you give a fire inspector, and which would you give a funder?
3. Name the level of measurement for each variable. (a) Zip code. (b) Number of nights in a shelter. (c) Self-rated health on a scale of poor, fair, good, very good, and excellent. (d) Year of birth.
4. Explain why a one-night count includes a higher share of people with long stays than an annual count does.

Apply It

5. Use the file `ch01-pit-counts.csv`. Calculate the percent change in the total from 2016 to 2024 and from 2024 to 2025. Show each step.
6. Rewrite this question so that it names a population, a concept, and a time and place. "Is hunger a problem for college students?"
7. Write an operational definition of "regular exercise" that a survey could use. Then write a second one that would give a higher count. State which decision each would suit.

Think Deeper

8. A city reports that its unsheltered count fell by 20 percent in one year. List three explanations that involve the procedure and no change in the number of people. What would you ask the city before you repeated the result?
9. The school count includes families staying with relatives, and the HUD count does not. Argue for each choice. Which would you use to plan shelter capacity, and which to plan school services?
10. Choose a number you have seen in the news this month. Find the operational definition behind it. Describe one group the definition leaves out.

CHAPTER 2

Where Error Comes From

Where this chapter fits. Every step of the research process can add error to the final number. This chapter names the kinds of error, so that the later chapters can address them one at a time.

2.1 Four Groups, One Question

How many immigrants live in the United States without legal authorization? No agency keeps a list. The number has to be estimated, and four research groups publish estimates that reporters and lawmakers cite.

In August 2025, the Pew Research Center reported 14.0 million for July 2023. The Migration Policy Institute first reported 13.7 million for the middle of 2023 and later revised that figure to 14.7 million. The Center for Migration Studies reported 12.2 million for 2023. The most recent estimate from the Department of Homeland Security was 11.0 million, for January 2022. Figure 2.1 shows these estimates and three others from the same groups.

The estimates for 2023 differ by 2.5 million people. All four groups use the same basic method, and all four start from the same Census Bureau survey. The differences come from choices made at each step of the work.

This chapter uses the four estimates to introduce a framework. The framework lists every step at which a survey number can depart from the truth. Each later chapter in this book covers one or two of those steps.

2.2 How the Estimate Is Built

The four groups use a **residual method**. It has three steps.

1. Estimate the total number of foreign-born people living in the country. This figure comes from the American Community Survey, which the Census Bureau sends to about 3.5 million addresses a year.
2. Estimate the number of foreign-born people who are in the country legally. This figure comes from government records of naturalized citizens, permanent residents, refugees, and visa holders.

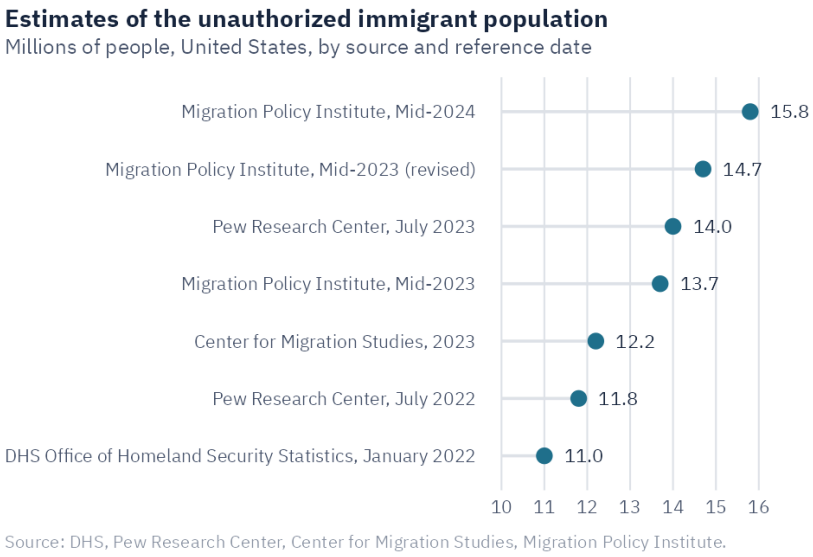


FIGURE 2.1. *Seven published estimates of the unauthorized immigrant population of the United States, by the date each estimate refers to. The estimates for 2023 range from 12.2 million to 14.7 million. Sources: Department of Homeland Security, Pew Research Center, Center for Migration Studies, and Migration Policy Institute.*

3. Subtract the second number from the first. The remainder, or residual, is the estimate of the unauthorized population.

The method is reasonable, and it is the best one available. It also depends on a survey, and surveys contain error. A small error in the first step becomes a large error in the residual. The total foreign-born population is about 50 million. Suppose the survey misses 2 percent of them. The first number is then low by 1 million, and the residual is low by the same 1 million.

2.3 Error Is Not the Same as a Mistake

In everyday speech, an error is a mistake. In survey research, **error** means the difference between an estimate and the true value. A study can be done with care and still contain error. Error cannot be removed. It can be reduced, measured, and reported.

Error has two forms, and they behave differently.

Variable error differs from one run of the study to the next. If the same survey were repeated with a new sample, the result would change a little each time. The changes go in both directions and tend to cancel out when many results are averaged. A larger sample reduces this kind of error.

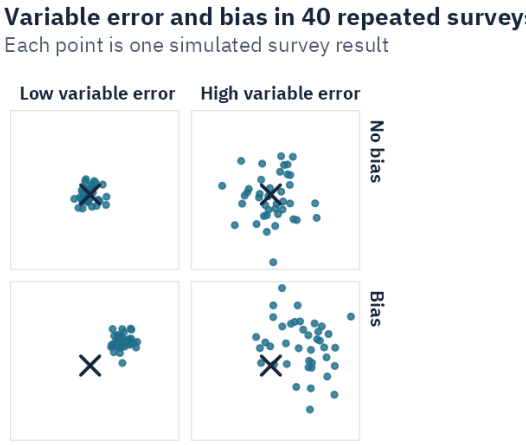


FIGURE 2.2. *Simulated results of 40 repeated surveys under four conditions. The cross marks the true value. Variable error spreads the results around a center. Bias moves the center away from the true value. The points are simulated for illustration.*

Bias pushes the result in one direction every time. If a survey cannot reach people without a fixed address, it will undercount them in every run. Repeating the survey does not help. A larger sample does not help either. It gives a more precise estimate of the wrong number.

Figure 2.2 shows the four combinations. The lower left panel is the most misleading case. The results agree closely with one another, and all of them are wrong. Agreement among results is not evidence that the results are accurate.

2.4 The Total Survey Error Framework

Survey researchers organize the sources of error into a framework called **total survey error**. Groves and colleagues (2009) present it as two paths. One path concerns *whom* the survey describes. The other concerns *what* the survey measures.

Whom the Survey Describes

A survey aims to describe a **target population**. For the American Community Survey, the target is every person living in the United States. Four kinds of error can enter between the target population and the final dataset.

1. **Coverage error.** The **sampling frame** is the list from which the sample is drawn. The American Community Survey draws from a list of addresses. People who are not linked to a listed address have no chance of selection. Coverage error is the difference between the target population and the

frame. Recent arrivals who share housing or move often are more likely to be missed.

2. **Sampling error.** A survey measures a sample and not the whole frame. A different sample would give a slightly different result. This is variable error, and it can be calculated.
3. **Nonresponse error.** Some people who are selected do not answer. If the people who do not answer differ from those who do, the result is biased. Immigrants without legal status may have reasons to avoid a government survey.
4. **Adjustment error.** Researchers apply weights to make the sample match known population totals. The weights depend on those totals. If the totals are wrong, the weights carry the error into every result.

What the Survey Measures

Three more kinds of error can enter between the concept and the recorded answer.

5. **Validity.** The question may not capture the concept the researcher has in mind. The American Community Survey asks where a person was born and whether the person is a citizen. It does not ask about legal status. Legal status has to be inferred, and the four groups infer it in different ways.
6. **Measurement error.** The answer a respondent gives may differ from the truth. A respondent may misunderstand the question, may not know the answer, or may prefer not to give it.
7. **Processing error.** Errors can enter when answers are coded, edited, or filled in. When a respondent skips a question, the Census Bureau estimates the missing value from similar respondents. That estimate can be wrong.

Table 2.1 summarizes the seven sources.

TABLE 2.1. *The seven sources of error in the total survey error framework, after Groves et al. (2009).*

Source of error	The question it raises	Where this book covers it
Coverage	Who had no chance of being selected?	Chapter 6
Sampling	How much would the result change with a new sample?	Chapter 6
Nonresponse	Who was selected and did not answer?	Chapters 6 and 7
Adjustment	What totals were the weights based on?	Chapter 6
Validity	Does the question capture the concept?	Chapters 1 and 4
Measurement	Did respondents answer accurately?	Chapters 4 and 5

Source of error	The question it raises	Where this book covers it
Processing	What happened to the answers after they were given?	Chapter 7

2.5 Tracing the Four Estimates

The framework explains most of the gap among the estimates in Figure 2.1.

Reference date. The estimates refer to different dates. The Department of Homeland Security figure is for January 2022. The Pew figure is for July 2023. The population grew quickly between those dates, and so the two figures should differ. This is a difference in the research question and not an error. A reader who compares them without checking the dates will see a disagreement that is partly a difference in timing.

Adjustment. In December 2024, the Census Bureau raised its estimates of immigration for 2021 through 2023. The survey weights for those years had been based on the earlier, lower totals. Pew reweighted the survey data to match the new totals. Its estimate for 2022 rose from 11.0 million to 11.8 million. No new data had been collected. The change came from the weights.

Coverage. The survey misses some immigrants, and each group adds people back to correct for the undercount. They use different correction rates. The Migration Policy Institute adopted a method that corrects more fully for people the survey missed. Its estimate for 2023 rose from 13.7 million to 14.7 million.

Validity. The category itself needs a definition. Pew reports that about 6 million of the 14 million people in its 2023 estimate had some protection from deportation. These include people with pending asylum claims and people with temporary protected status. All four groups count them as unauthorized. A reader who assumes that "unauthorized" means "subject to removal" will misread the total.

None of these differences is a mistake by the researchers. Each group documents its choices. A reader who cites one number without its date, its definition, and its correction for undercount has left out the information needed to interpret it.

2.6 Work With the Data: Survey Error Tree

The section above traced four estimates to their sources of error. The tool gives you ten more cases to sort. Each case describes a problem in a survey.

The tool runs in a web browser at quant.marginoferrormedia.com/tools/error-tree.html. It needs no account and no software.

1. Read the case. Before you choose, ask the seven questions in the table under the cases.

2. Choose the source of error that fits best. Read the explanation, including when your choice was correct.
3. Work through all ten cases. The tool reports your score at the end.
4. For each case you missed, name the question from the table that would have led you to the better answer.

Some cases fit more than one source. Case 9 is a question that asks two things at once. The wording is the cause, and the damage shows up as inaccurate answers. When two sources seem to fit, name the step in the survey where the error entered.

2.7 What a Margin of Error Covers

News reports often give a **margin of error** with a survey result. The margin of error describes sampling error. It describes how much the result would be expected to vary if the survey were repeated with new samples of the same size.

It leaves out the other six sources in Table 2.1. For the immigration estimates, sampling error is small, because the survey is very large. The differences among the four groups are many times larger than the sampling error of any one estimate.

This is a common pattern. In large surveys, bias from the other six sources is usually larger than sampling error. A report that gives only a margin of error has described the one source of error that is easiest to calculate.

A Rule for Comparing Two Numbers

Before you compare two estimates, check four things. They are the reference date, the definition, the source of the data, and any correction the authors applied. If any of the four differs, part of the gap between the numbers comes from that difference.

Ethics Moment: The Cost of Being Counted

The people in this case have reasons to avoid a government survey. A person who answers may fear that the answer will be used against the household. That fear lowers response and makes the estimate worse. It also places a duty on the researcher. Federal law bars the Census Bureau from sharing individual answers with any other agency. A researcher who collects sensitive data has to offer the same kind of protection, state it plainly, and keep the promise. Chapter 7 returns to this duty.

AI Reality Check: One Number, No Method

Ask a language model how many unauthorized immigrants live in the United States, and it will often give one number. It may not say which group produced the number or what date it refers to. Its answer may also predate the most recent revision. Ask for the source, the reference date, and the definition. Then ask what other groups report for the same date. A single figure with no method attached should be treated as a starting point for a search.

2.8 Putting It All Together

Error is the difference between an estimate and the true value. Variable error changes from one sample to the next and falls as the sample grows. Bias pushes every result in the same direction and does not fall as the sample grows.

The total survey error framework lists seven sources. Four concern whom the survey describes. They are coverage, sampling, nonresponse, and adjustment. Three concern what the survey measures. They are validity, measurement, and processing. The margin of error covers only sampling.

The four estimates of the unauthorized population differ by millions. The differences come from reference dates, weights, corrections for undercount, and definitions. Sampling error accounts for very little of the gap.

2.9 Looking Ahead

A survey describes a population. It can show that two things occur together, and it usually cannot show that one causes the other. Chapter 3 covers the design that can support a causal claim. The case is an experiment in which researchers mailed thousands of resumes to employers.

2.10 Further Reading and References

- Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2009). *Survey Methodology* (2nd ed.). Wiley. Chapter 2 presents the total survey error framework used in this chapter.
- Passel, J. S., and Krogstad, J. M. (2025). *U.S. Unauthorized Immigrant Population Reached a Record 14 Million in 2023.* Pew Research Center. The report describes the residual method and the reweighting of the 2022 estimate.
- Migration Policy Institute (2026). *Following Years of High Border Arrivals, the U.S. Unauthorized Immigrant Population Hits New Peak.* The source for the revised estimate for 2023 and the estimate for 2024.

- Warren, R. (2025). *The Undocumented Population in the United States Increased to 12.2 Million in 2023.* Center for Migration Studies.
- Baker, B., and Warren, R. (2024). *Estimates of the Unauthorized Immigrant Population Residing in the United States: January 2018 to January 2022.* Department of Homeland Security, Office of Homeland Security Statistics.
- Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly, 74*(5), 817 to 848.

2.11 Key Terms

- **Adjustment error.** Error that enters when weights are based on incorrect population totals.
- **Bias.** Error that pushes every result in the same direction.
- **Coverage error.** The difference between the target population and the sampling frame.
- **Error.** The difference between an estimate and the true value.
- **Margin of error.** A measure of sampling error. It does not cover the other six sources.
- **Measurement error.** The difference between the answer a respondent gives and the truth.
- **Nonresponse error.** Bias that arises when people who do not answer differ from those who do.
- **Processing error.** Error that enters when answers are coded, edited, or filled in.
- **Residual method.** An estimate made by subtracting one count from another.
- **Sampling error.** Variation in results that comes from measuring a sample and not the whole frame.
- **Sampling frame.** The list from which a sample is drawn.
- **Target population.** The group a survey aims to describe.
- **Total survey error.** A framework that lists every source of error in a survey estimate.
- **Variable error.** Error that differs from one run of a study to the next.

2.12 Exercises

Check Your Understanding

1. Explain why a larger sample reduces variable error and does not reduce bias.
2. A phone survey calls only landlines. Name the source of error and the group most likely to be left out.
3. Which of the seven sources does a margin of error describe?

4. Pew's estimate for 2022 changed from 11.0 million to 11.8 million without new data. Which source of error does this illustrate?

Apply It

5. Use the file `ch02-unauthorized-estimates.csv`. Calculate the range of the estimates that refer to 2023. Express the range as a percent of the lowest estimate.
6. The foreign-born population is about 50 million. Suppose the legally present population is 36 million, and the survey misses 3 percent of all foreign-born people. Calculate the residual with and without a correction for the undercount. Show each step.
7. A customer survey is emailed to everyone who bought a product in the past year. Twelve percent respond. List one possible problem under each of four sources of error.

Think Deeper

8. A news report says two estimates "contradict" each other. One refers to January 2022 and the other to July 2023. Write two sentences that describe the relationship between them more accurately.
9. All four groups count people with pending asylum claims as unauthorized. Give one argument for this choice and one against it. How should a report word the total so that readers do not misread it?
10. Find a survey result in the news that reports a margin of error. Which of the other six sources is most likely to matter for that result, and why?

PART II

Design and Measurement

CHAPTER 3

Designs That Support Causal Claims

Where this chapter fits. Step 3 of the research process is to choose the research design. This chapter covers the designs that can support a claim about cause, and what the other designs can support.

3.1 The Same Resume With Two Names

In 2001 and 2002, two economists answered help-wanted ads in Boston and Chicago newspapers. Marianne Bertrand and Sendhil Mullainathan sent 4,870 resumes in response to more than 1,300 ads. The jobs were in sales, administrative support, clerical work, and customer service.

The resumes were fictitious. The researchers built them from real resumes and then assigned a first name to each one at random. Half of the resumes received a name that is more common among white Americans, such as Emily or Greg. The other half received a name that is more common among Black Americans, such as Lakisha or Jamal. The researchers then recorded which resumes received a call or an email from an employer.

Resumes with white-sounding names received a callback 9.65 percent of the time. Resumes with Black-sounding names received a callback 6.45 percent of the time. Figure 3.1 shows the two rates.

A difference of 3.2 percentage points may look small. In relative terms, a resume with a white-sounding name was about 50 percent more likely to receive a callback. An applicant with a white-sounding name needed to send about 10 resumes for each callback. An applicant with a Black-sounding name needed to send about 15.

The size of the gap is one reason the study is widely cited. The design is the other reason. Because the names were assigned at random, the study supports a claim that most studies of discrimination cannot support. The name on the resume caused the difference in callbacks. This chapter explains why the design allows that claim.

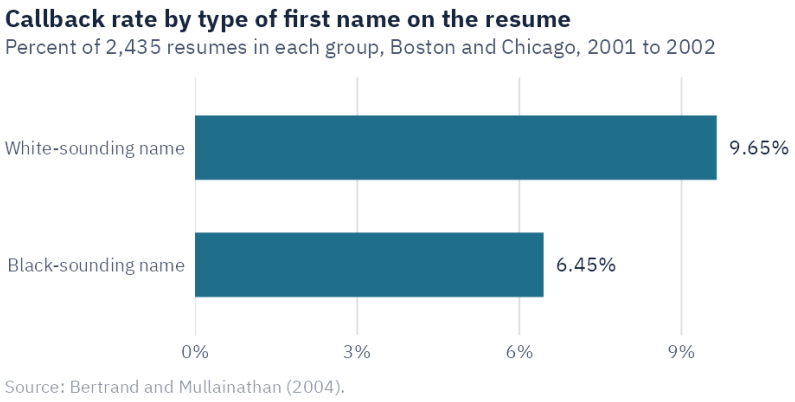


FIGURE 3.1. *Share of resumes that received a callback, by the type of first name assigned to the resume. Each group contains 2,435 resumes. Source: Bertrand and Mullainathan (2004).*

3.2 Why Comparing Groups Is Not Enough

Suppose a researcher took a different approach. The researcher collects records of real job applicants and compares callback rates for Black and white applicants. Suppose the white applicants received more callbacks.

That result would be an **association**. Two variables are associated when the values of one tend to go with certain values of the other. An association does not show that one variable causes the other.

Real applicants differ in many ways besides race. They differ in education, work history, the neighborhoods they live in, and the jobs they apply for. Any of those differences could account for a gap in callbacks. A variable that is related to both the supposed cause and the outcome is a **confounder**. With records of real applicants, the researcher cannot rule out confounders, because no two applicants are identical in everything except race.

Researchers generally require three conditions before they accept that one variable causes another.

1. **Association.** The two variables are related.
2. **Time order.** The cause comes before the effect.
3. **No alternative explanation.** No third variable accounts for the relationship.

A comparison of real applicants can meet the first two conditions. It rarely meets the third.

3.3 What Random Assignment Does

An **experiment** is a study in which the researcher controls who receives which condition. The variable the researcher controls is the **independent variable**, also called the **treatment**. The outcome the researcher measures is the **dependent variable**. In the resume study, the type of name is the independent variable and the callback is the dependent variable.

Random assignment means that chance alone decides which unit receives which condition. In the resume study, a computer assigned the names. The quality of a resume had no influence on which name it received.

Random assignment addresses the third condition. Because chance decided the names, the two groups of resumes are alike, on average, in every other respect. They have similar education, similar experience, and similar formatting. The same is true for characteristics the researchers did not think to measure. Only the name differs in a consistent way. If callbacks differ between the groups by more than chance would produce, the name is the only remaining explanation.

The data file allows a check. Table 3.1 compares the two groups on three resume characteristics.

TABLE 3.1. *Characteristics of the resumes in the two name groups. Random assignment produced groups that are nearly identical. Source: Bertrand and Mullainathan (2004).*

Name group	Resumes	Mean years of experience	With a college degree	Rated high quality
Black-sounding name	2,435	7.8	72.3%	50.2%
White-sounding name	2,435	7.9	71.6%	50.2%

The two groups match closely on all three. This is what random assignment is expected to produce when the groups are large.

3.4 Work With the Data: Random Assignment Lab

The study found a gap of 3.2 percentage points in callbacks. Chance alone produces some gap between any two groups. The tool shows how large a gap chance produces here. It uses the callback totals from the study.

The tool runs in a web browser at quant.marginoferrormedia.com/tools/assignment-lab.html. It needs no account and no software.

1. Press "Shuffle 1,000 times." The tool deals the 392 callbacks into two groups of 2,435 at random and plots the gap from each shuffle.
2. Read how many of the 1,000 shuffles gave a gap of 3.2 points or more.
3. Read the largest gap that chance produced. Compare it with the dark line at 3.2 points.

4. Press the button several more times. Note how little the answer changes.

In most runs, no shuffle reaches 3.2 points. Chance is therefore a poor explanation for the gap in the study. Random assignment is what permits this reasoning. The names were assigned by chance, so chance and the names are the only two explanations left.

The full file of 4,870 resumes is `ch03-resume-audit.csv` at quant.marginoferrormedia.com/da

Random Assignment and Random Sampling

These two terms sound alike and do different jobs. Random *sampling* decides who is in the study. It supports the claim that the results describe a larger population. Random *assignment* decides who receives which condition. It supports the claim that the treatment caused the outcome. A study can have one, both, or neither. The resume study had random assignment. It did not have a random sample of all employers.

3.5 Internal and External Validity

Two questions apply to any study that claims a cause.

Internal validity concerns whether the treatment caused the outcome within the study. Random assignment gives the resume study strong internal validity. Few alternative explanations remain for the gap in callbacks.

External validity concerns whether the result holds for other people, places, and times. Here the resume study is more limited. It covered two cities, four job categories, and ads placed in newspapers in 2001 and 2002. It measured callbacks, which are the first step in hiring, and it did not measure job offers or pay.

The two kinds of validity often compete. A tightly controlled experiment has strong internal validity and may not resemble ordinary conditions. A large survey of real workers resembles ordinary conditions and cannot rule out confounders. No single study settles a question. Confidence grows when studies with different strengths reach the same conclusion.

Later research has addressed the external validity of the resume study. Kline, Rose, and Walters (2022) sent more than 83,000 fictitious applications to 108 of the largest employers in the United States. Applications with distinctively Black names were contacted about 2 percentage points less often. The gap was smaller than in the original study, and it was concentrated in a subset of firms. The direction of the result was the same twenty years later, with different employers and online applications.

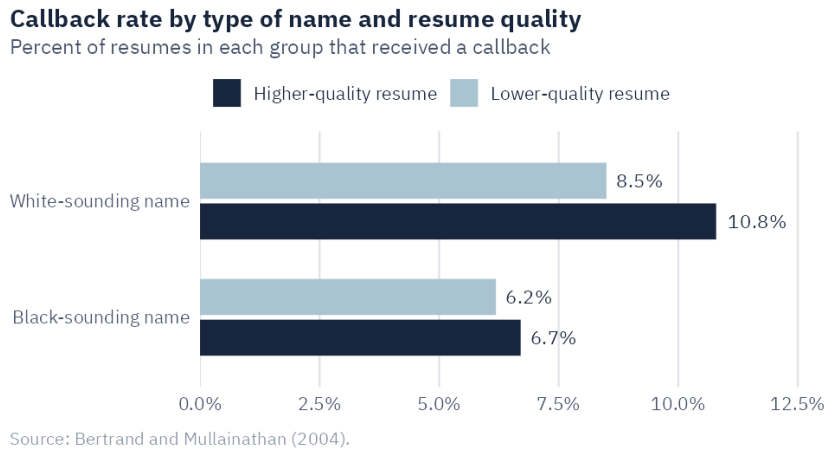


FIGURE 3.2. *Callback rate by type of name and resume quality. A higher-quality resume raised the callback rate by 2.3 percentage points for white-sounding names and by 0.5 points for Black-sounding names. Source: Bertrand and Mullainathan (2004).*

3.6 A Second Factor

The researchers varied a second characteristic. Each resume was written as either higher quality or lower quality. The higher-quality resumes listed more experience, fewer gaps in employment, and more skills. This gave the study four groups in place of two. A design that crosses two independent variables is a **factorial design**.

Figure 3.2 shows the four rates. A better resume helped applicants with white-sounding names more than it helped applicants with Black-sounding names. When the effect of one variable depends on the level of another, the two variables **interact**. A study with only one factor could not have found this result.

3.7 What the Study Does Not Show

A careful reader separates what a design supports from what it does not.

The study shows that the name on a resume changed the callback rate. It does not show why employers responded as they did. The study cannot separate a preference against Black applicants from an assumption about applicants with those names.

Several researchers have also asked what a name signals. Fryer and Levitt (2004) found that distinctively Black names are more common among children of mothers with less education. An employer could read a name as a signal of social class as well as of race. Gaddis (2017) tested this directly and found that people do not perceive all such names in the same way. The treatment in the

study was a name. The conclusion concerns race. The step from one to the other is a validity question of the kind raised in Chapter 2.

The study also measured the first step in hiring only. A gap in callbacks may grow or shrink at the interview and the offer.

These limits do not weaken the central result. They mark where it ends.

3.8 When an Experiment Is Not Possible

Many questions cannot be studied with an experiment. A researcher cannot assign people to lose their housing or to grow up in a poor neighborhood. In those cases, researchers use other designs.

- A **cross-sectional study** measures a group once. Most surveys are cross-sectional. They describe a population and show associations.
- A **longitudinal study** measures the same people more than once. It can show that a change in one variable came before a change in another, which addresses time order.
- A **quasi-experiment** compares groups that were formed by an event the researcher did not control. A new law that applies in one state and not in a neighboring state is an example. The groups were not assigned at random, and so the researcher has to argue that they were similar before the event.

Each design supports some claims and not others. Table 3.2 summarizes them.

TABLE 3.2. *Four research designs and the conditions for a causal claim that each can meet.*

Design	Researcher assigns the treatment	Association	Time order	Rules out confounders
Cross-sectional survey	No	Yes	Rarely	No
Longitudinal study	No	Yes	Yes	Partly
Quasi-experiment	No	Yes	Yes	Partly, with an argument
Experiment with random assignment	Yes	Yes	Yes	Yes, on average

The resume study is a special kind of experiment called an **audit study**. An audit study sends matched testers or matched applications into a real market and records how they are treated. It combines the control of an experiment with the conditions of a real market.

Ethics Moment: An Experiment Without Consent

The employers in the resume study did not know they were in a study. They spent time reading applications from people who did not exist. Research ethics normally require informed consent, and this study did not obtain it. Review boards allow a waiver of consent when three conditions hold. The study could not be done if participants knew about it. The risk to participants is small. The knowledge gained is important. The researchers also did not publish the names of the employers. Reasonable people disagree about whether these conditions were met. A researcher who plans a study that involves deception has to make the case to a review board before the study begins.

AI Reality Check: A Model Can Describe a Design and Cannot Assign One

A language model can explain random assignment and can suggest a design for your question. It cannot make your data come from that design. If you give a model a file of survey responses and ask what causes what, it will often answer in causal language. The language does not change what the data support. Before you accept a causal sentence from a model, ask how the treatment was assigned. If no one assigned it, treat the sentence as a statement about association.

3.9 Putting It All Together

An association between two variables does not show that one causes the other. A causal claim requires an association, the right time order, and no alternative explanation. Random assignment addresses the third condition by making the groups alike, on average, in everything except the treatment.

The resume study assigned names at random to 4,870 resumes. Resumes with white-sounding names received callbacks at a rate of 9.65 percent, and resumes with Black-sounding names at 6.45 percent. The design supports the claim that the name caused the gap. It does not show why employers responded as they did, and it measured only the first step in hiring.

Internal validity concerns whether the treatment caused the outcome within the study. External validity concerns whether the result holds elsewhere. When an experiment is not possible, longitudinal studies and quasi-experiments can meet some of the conditions for a causal claim.

3.10 Looking Ahead

The resume study measured a simple outcome. An employer called back or did not. Many research questions concern constructs that are much harder to measure, such as hunger, health, or trust. Chapter 4 covers how researchers build a measure from several questions and how they check that it works.

3.11 Further Reading and References

- Bertrand, M., and Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review, 94*(4), 991 to 1013. The source of the case and the data.
- Kline, P., Rose, E. K., and Walters, C. R. (2022). Systemic discrimination among large U.S. employers. *Quarterly Journal of Economics, 137*(4), 1963 to 2036. A much larger audit study of 108 large firms.
- Fryer, R. G., and Levitt, S. D. (2004). The causes and consequences of distinctively Black names. *Quarterly Journal of Economics, 119*(3), 767 to 805.
- Gaddis, S. M. (2017). How Black are Lakisha and Jamal? Racial perceptions from names used in correspondence audit studies. *Sociological Science, 4*, 469 to 489.
- Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Houghton Mifflin. The standard reference on validity and design.

3.12 Key Terms

- **Association.** A relationship in which the values of one variable tend to go with certain values of another.
- **Audit study.** An experiment that sends matched testers or applications into a real market.
- **Confounder.** A variable related to both the supposed cause and the outcome.
- **Cross-sectional study.** A study that measures a group once.
- **Dependent variable.** The outcome a researcher measures.
- **Experiment.** A study in which the researcher controls who receives which condition.
- **External validity.** Whether a result holds for other people, places, and times.
- **Factorial design.** A design that crosses two or more independent variables.
- **Independent variable.** The variable a researcher controls. Also called the treatment.

- **Interaction.** A pattern in which the effect of one variable depends on the level of another.
- **Internal validity.** Whether the treatment caused the outcome within the study.
- **Longitudinal study.** A study that measures the same people more than once.
- **Quasi-experiment.** A comparison of groups formed by an event the researcher did not control.
- **Random assignment.** The use of chance alone to decide which unit receives which condition.

3.13 Exercises

Check Your Understanding

1. State the three conditions for a causal claim. Which one does random assignment address?
2. Explain the difference between random assignment and random sampling. Which did the resume study use?
3. A study finds that people who own dogs report better health. Name two possible confounders.
4. Define internal validity and external validity. Rate the resume study on each and give one reason for each rating.

Apply It

5. In the resume study, 235 of 2,435 resumes with white-sounding names received a callback, and 157 of 2,435 with Black-sounding names did. Calculate each rate, the difference in percentage points, and the ratio of the two rates. Show each step.
6. Use the file `ch03-resume-audit.csv`. Calculate callback rates by gender and type of name. Describe the pattern in two sentences.
7. A city wants to know whether a text-message reminder raises attendance at housing court. Design an experiment. Name the independent variable, the dependent variable, the unit that is assigned, and how you would assign it.

Think Deeper

8. The treatment in the resume study was a name, and the conclusion concerns race. What does a reader have to assume to move from one to the other? How could a new study test that assumption?
9. Employers in the study did not consent. Argue for and against the waiver of consent. What would you require of the researchers?

10. A company uses software to screen resumes. Describe an audit study that would test whether the software treats applicants differently by the type of name. What would the study show, and what would it leave open?

CHAPTER 4

Measurement and Scales

Where this chapter fits. Step 4 of the research process is to design the method and the forms. This chapter covers the first half of that step, which is deciding how each concept will be measured.

4.1 Six Questions About Food

Each December, the Census Bureau adds a set of questions about food to its monthly survey of households. The US Department of Agriculture uses the answers to estimate how many households had trouble getting enough food during the year. In 2024, the estimate was 13.7 percent of households. The department calls these households **food insecure**.

No survey question asks, "Are you food insecure?" The term is a construct, and most respondents would not know what it means. The department builds the measure from a series of specific questions about what happened in the household. A short form of the measure uses six questions. Table 4.1 lists them.

TABLE 4.1. *The six-item short form of the US household food security module. Source: USDA Economic Research Service.*

	Question, about the last 12 months	Answer counted as a yes
1	``The food that we bought just didn't last, and we didn't have money to get more.''	Often or sometimes true
2	``We couldn't afford to eat balanced meals.''	Often or sometimes true
3	Did you or other adults in the household ever cut the size of meals or skip meals because there wasn't enough money for food?	Yes
4	If yes to question 3, how often did this happen?	In three or more months
5	Did you ever eat less than you felt you should because there wasn't enough money for food?	Yes

	Question, about the last 12 months	Answer counted as a yes
6	Were you ever hungry but didn't eat because there wasn't enough money for food?	Yes

Each yes adds one point. A household with a score of 0 or 1 is classified as food secure. A score of 2 to 4 is **low food security**. A score of 5 or 6 is **very low food security**. Households in the last two groups are counted as food insecure.

The department has reported this measure every year since 1995. The report on 2024, released in December 2025, was the last in the series. The department has ended the annual report.

This chapter uses the six questions to explain how a measure is built from several items. It then covers the two tests that any measure has to pass.

4.2 Why Several Questions

A single question is often a weak measure of a construct. There are three reasons.

First, a construct usually has more than one part. Food insecurity includes worry about running out, lower quality of meals, and eating less. One question can cover only one of these.

Secondly, any single answer contains some random error. A respondent may misread a word or recall the wrong month. When several answers are added together, the random errors tend to cancel out.

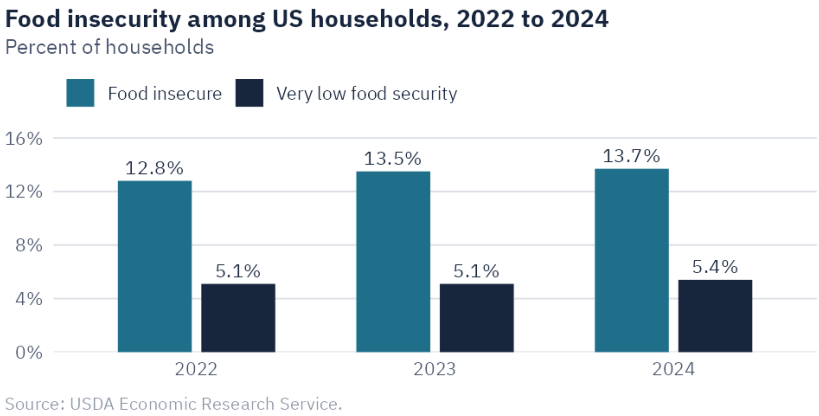


FIGURE 4.1. *Share of US households classified as food insecure and as having very low food security, 2022 to 2024. The department reported that the changes across these three years were not statistically significant. Source: USDA Economic Research Service, Household Food Security in the United States in 2024.*

Thirdly, a single yes-or-no question separates respondents into only two groups. Six questions give seven possible scores. The measure can then separate mild cases from severe ones.

A set of questions that are combined into one score is a **scale**. Each question is an **item**. The six food questions form a scale.

Many scales use **Likert items**. A Likert item presents a statement and asks how much the respondent agrees, usually with five ordered choices from "strongly disagree" to "strongly agree." The choices are scored 1 to 5 and the item scores are added. Rensis Likert introduced the format in 1932, and it is now the most common format in survey research.

Some items in a scale may be worded in the opposite direction. On a job satisfaction scale, "I look forward to work" and "I often think about quitting" point in opposite directions. The second item has to be **reverse scored** before the items are added. On a five-point item, a 5 becomes a 1 and a 4 becomes a 2. A researcher who forgets this step will produce a score with little meaning.

4.3 How the Items Line Up

The six food questions were not chosen at random. They differ in how severe a situation they describe.

Many households with low incomes run out of food before the next paycheck. Far fewer go hungry for lack of money. A household that reports the severe experience has almost always had the milder ones as well.

Figure 4.2 shows this pattern in a simulated dataset. The file contains answers from 600 simulated households. It was mathematically modeled to reflect the ordering of the items that the department reports. It does not describe real households.

Because the items are ordered by severity, the total score has a clear meaning. A score of 1 usually means that food ran short at some point. A score of 6 means that an adult in the household went hungry. The cut points at 2 and 5 divide the range into three levels.

A **cut point** turns a score into a category. Cut points are useful for reporting, and they are also choices. Two households with scores of 1 and 2 may differ very little. The measure still places them in different groups. A report should state the cut points it used.

4.4 Reliability

A measure is **reliable** if it gives consistent results. A bathroom scale that shows a different weight each time you step on it is unreliable. Reliability is about consistency and not about accuracy. A scale that always reads five pounds too heavy is reliable.

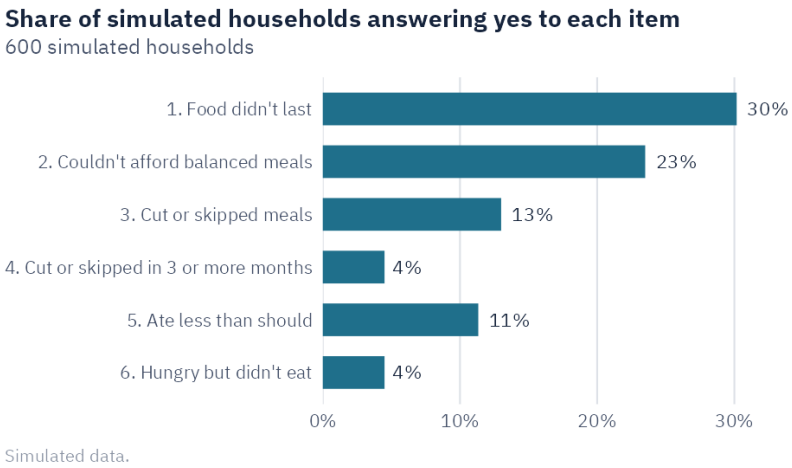


FIGURE 4.2. *Share of 600 simulated households that answered yes to each of the six items. The milder items receive more yes answers than the severe ones. The data are simulated for illustration and are not an estimate for any population.*

Researchers check reliability in three main ways.

Test-retest reliability. The same people answer the same questions twice, a few weeks apart. If the construct has not changed, the two sets of scores should be highly correlated.

Inter-rater reliability. Two people score the same material independently. If the scoring rules are clear, the two sets of scores should agree. This check applies when a person has to make a judgment, such as rating the quality of an essay.

Internal consistency. The items in a scale are compared with one another. If they measure the same construct, respondents who say yes to one should tend to say yes to the others.

The usual statistic for internal consistency is **Cronbach's alpha**. Its symbol is α. It compares the variance of the individual items with the variance of the total score.

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum s_i^2}{s_T^2}\right)$$

In the formula, k is the number of items, s_i^2 is the variance of item i, and s_T^2 is the variance of the total score. When the items rise and fall together, the total score varies more than the items do separately, and alpha is high.

Here is the calculation for the simulated file.

1. Count the items. The scale has $k = 6$ items.

2. Calculate the variance of each item and add the six variances. The sum is 0.691.
3. Calculate the total score for each household and find its variance. The variance is 1.986.
4. Divide the first number by the second. The result is 0.348.
5. Subtract from 1. The result is 0.652.
6. Multiply by $k/(k-1)$, which is $6/5 = 1.2$. Alpha is 0.78.

Alpha ranges from 0 to 1 in most cases. By convention, a value of 0.70 or higher is treated as acceptable for research on groups. A value of 0.90 or higher is expected when a score will be used to decide something about one person. These are conventions and not strict limits.

4.5 Work With the Data: Scale Scorer

The steps above computed alpha by hand for the six food items. The tool does the same work in two parts. Part 1 scores one household. Part 2 computes alpha for the simulated file of 600 households.

The tool runs in a web browser at quant.marginoferrormedia.com/tools/scale-scorer.html. It needs no account and no software.

1. In Part 1, answer yes to one item at a time. Note the score at which the category changes to low food security, and then to very low food security.
2. In Part 2, read alpha for all six items. It matches the value computed above.
3. Remove one item and record alpha. Put the item back and remove a different one. Note which item changes alpha the most.
4. Remove items one at a time until two remain. Record alpha at each step.

Alpha usually falls as a scale gets shorter, even when the remaining items are no worse than before. The note that follows explains why this limits what alpha can show.

The simulated answers are in the file `ch04-simulated-items.csv` at quant.marginoferror

What Alpha Does Not Show

Alpha rises when items are added, even weak ones. A scale with 30 items can reach a high alpha with items that are only loosely related. Alpha also does not show that the items measure one construct. Two groups of items that measure two different things can produce a high alpha together. Report alpha, and also examine how each item relates to the others.

4.6 Validity

A measure is **valid** if it measures the construct it is meant to measure. Reliability is necessary for validity, and it is not enough. The bathroom scale that reads five pounds too heavy is reliable and not valid.

Validity cannot be shown with one number. Researchers build a case for it from several kinds of evidence.

Content validity. The items should cover all parts of the construct and nothing outside it. Experts in the field review the items to check this. The food security items cover worry, quality of meals, and quantity of food. They do not cover the nutritional value of what was eaten, and the department states that the measure does not assess nutrition.

Criterion validity. Scores should relate to an outside standard that is already accepted. Households classified as food insecure should, on average, spend less on food. The department reports this check each year. In 2024, the typical food-secure household spent about 11 percent more on food than the typical food-insecure household of the same size and makeup.

Construct validity. Scores should relate to other measures in the ways that theory predicts. Two kinds of evidence support it. **Convergent validity** means that the measure correlates with measures of related constructs, such as income. **Discriminant validity** means that the measure does not correlate strongly with measures of unrelated constructs.

Table 4.2 shows how reliability and validity combine.

TABLE 4.2. *Combinations of reliability and validity.*

	Valid	Not valid
Reliable	The measure is consistent and measures the right construct. This is the goal.	The measure is consistent and measures the wrong thing. This case is hard to detect, because the results look stable.
Not reliable	Not possible. An inconsistent measure cannot be valid.	The measure is inconsistent and its results cannot be interpreted.

4.7 A Name Is Part of the Measure

Until 2006, the department's most severe category was labeled "food insecurity with hunger." A panel of the National Academies reviewed the measure and reported its findings that year. The panel concluded that the questions measure a household's access to food. Hunger is a physical condition of a person, and the panel found that the questions do not measure it directly. The department changed the label to "very low food security."

The questions did not change. The scores did not change. The label changed because it claimed more than the items could support. This is a validity question about the name of a measure. A reader of any report should check whether the label on a number matches the questions behind it.

4.8 Borrow Before You Build

A new scale takes months to develop and test. For most constructs, a tested scale already exists. A researcher who uses an existing scale gains two things. The scale has known reliability and validity. The results can also be compared with those of other studies that used the same scale.

If you use an existing scale, use it as written. Changing the wording, the order, or the answer choices can change how it performs, and the published evidence no longer applies. If you have to change it, report the change and check reliability again with your own data.

Ethics Moment: Asking About Hardship

The food questions ask people to report that they could not feed their household. For some respondents, that is a painful thing to say. A researcher has three duties here. Ask only what the study needs. Tell respondents that they may skip any question. Protect the answers so that no one can be identified. Short, tested scales help with the first duty, because they collect the needed information with fewer questions. Chapter 5 covers how the wording and the order of questions affect a respondent's comfort.

AI Reality Check: A Scale That Reads Well Is Not a Tested Scale

A language model can write ten Likert items on any topic in seconds. The items will be grammatical and will look like a scale. They have not been tested. No one knows their reliability, and no one has checked that they measure the construct. A model may also reproduce the items of a published scale with small changes in wording. The published evidence then no longer applies, and the copy may breach a copyright. Use a model to find candidate scales and to draft items for testing. Then test them with real respondents before you rely on the scores.

4.9 Putting It All Together

A scale combines several items into one score. Several items cover more of a construct than one item can, and their random errors tend to cancel out. The

six-item food security scale orders its items by severity, so that a higher score means a more severe situation. Cut points turn the score into categories.

A measure has to be reliable and valid. Reliability is consistency, and Cronbach's alpha is the usual statistic for the internal consistency of a scale. Validity is whether the measure captures the intended construct. The case for validity rests on content, criterion, and construct evidence.

The label on a measure is part of the measure. The department removed the word "hunger" from its categories because the questions did not measure hunger directly.

4.10 Looking Ahead

This chapter treated the wording of the six questions as given. Each word in them was tested before the questions were used. Chapter 5 covers how questions are written and how small changes in wording can change the answers. It also covers how to test a questionnaire before it is used.

4.11 Further Reading and References

- Rabbitt, M. P., Reed-Jones, M., Hales, L. J., Suttles, S., and Burke, M. P. (2025). *Household Food Security in the United States in 2024* (ERR-358). US Department of Agriculture, Economic Research Service. The source for the 2022 to 2024 figures.
- Bickel, G., Nord, M., Price, C., Hamilton, W., and Cook, J. (2000). *Guide to Measuring Household Food Security.* US Department of Agriculture, Food and Nutrition Service. The guide describes the items and the scoring.
- National Research Council (2006). *Food Insecurity and Hunger in the United States: An Assessment of the Measure.* National Academies Press. The panel report that led to the change in labels.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297 to 334.
- DeVellis, R. F., and Thorpe, C. T. (2021). *Scale Development: Theory and Applications* (5th ed.). Sage. A practical guide to building and testing a scale.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22*(140), 1 to 55.

4.12 Key Terms

- **Construct validity.** Evidence that scores relate to other measures in the ways theory predicts.
- **Content validity.** Evidence that the items cover all parts of the construct.

- **Convergent validity.** Evidence that a measure correlates with measures of related constructs.
- **Criterion validity.** Evidence that scores relate to an accepted outside standard.
- **Cronbach's alpha.** A statistic for the internal consistency of a scale.
- **Cut point.** A score that divides respondents into categories.
- **Discriminant validity.** Evidence that a measure does not correlate strongly with unrelated constructs.
- **Internal consistency.** The degree to which the items in a scale go together.
- **Inter-rater reliability.** Agreement between two people who score the same material.
- **Item.** One question in a scale.
- **Likert item.** A statement with ordered choices from strong disagreement to strong agreement.
- **Reliability.** The consistency of a measure.
- **Reverse scoring.** Flipping the scores of an item worded in the opposite direction.
- **Scale.** A set of items combined into one score.
- **Test-retest reliability.** Agreement between scores from the same people at two times.
- **Validity.** Whether a measure captures the construct it is meant to measure.

4.13 Exercises

Check Your Understanding

1. Give two reasons a scale of several items is usually better than a single question.
2. A measure gives the same wrong answer every time. Is it reliable? Is it valid?
3. Match each check to a kind of validity. (a) Experts review the items for coverage. (b) Scores are compared with household food spending. (c) Scores are compared with a measure of income.
4. Why did the department remove the word "hunger" from its category labels in 2006?

Apply It

5. A household answers yes to items 1, 2, and 5 and no to the rest. Calculate the score and name the category.
6. A four-item scale has item variances of 0.9, 1.1, 1.0, and 1.2. The variance of the total score is 9.8. Calculate Cronbach's alpha. Show each step.

7. Use the file `ch04-simulated-items.csv`. Calculate the total score for each household. What share of the simulated households falls in each of the three categories?

Think Deeper

8. A household with a score of 1 is "food secure," and a household with a score of 2 is "food insecure." What is gained and what is lost by reporting categories in place of scores?
9. The department has ended its annual report. What would a researcher lose if the same six questions were no longer asked each year? What could replace them, and with what loss in comparability?
10. Choose a construct that interests you, such as loneliness or financial stress. Find a published scale that measures it. Report the number of items, a sample item, and the reliability the authors report.

PART III

Instruments and Samples

CHAPTER 5

Writing Questions

Where this chapter fits. Step 4 of the research process continues. This chapter covers the wording of the questions that go on the form.

5.1 A Question That Took Years to Write

Each year, state health departments in the United States interview more than 400,000 adults by telephone. The survey is the Behavioral Risk Factor Surveillance System. The Centers for Disease Control and Prevention has coordinated it since 1984. One of its questions reads as follows.

> 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?

The question looks simple. Each part of it reflects a decision. "Physical health" is defined inside the question, so that respondents do not have to guess what it covers. The time period is stated as 30 days. The answer is a number of days, which is easier to report than a rating. The wording has been kept the same for decades, so that results can be compared across years and across states.

Suppose a county health department wants to survey its own residents. The staff could write new questions. A better plan is to borrow tested questions such as this one and add new questions only where no tested question exists. This chapter covers both tasks. It explains how respondents answer a question, how wording can change the answers, and how to test a questionnaire before using it.

5.2 How a Respondent Answers

Tourangeau, Rips, and Rasinski (2000) describe four steps that a respondent goes through to answer a survey question.

1. **Comprehension.** The respondent works out what the question is asking.
2. **Retrieval.** The respondent searches memory for the relevant information.

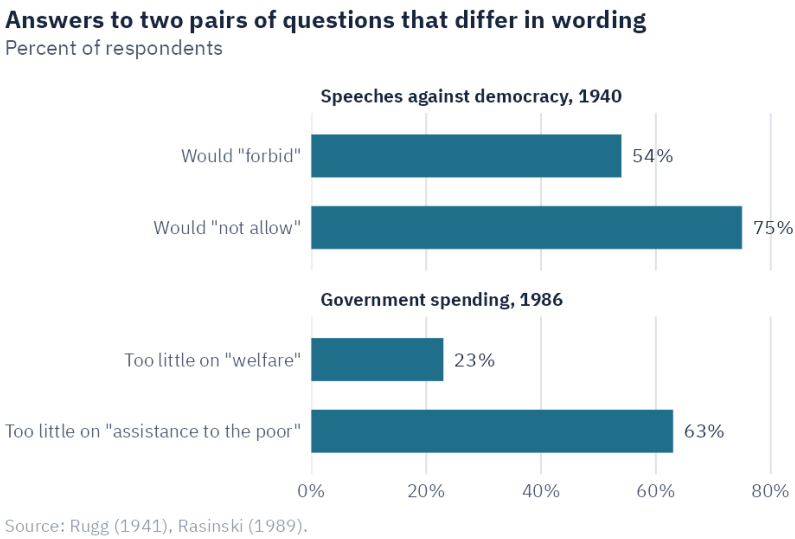


FIGURE 5.1. *Results of two split-ballot experiments on question wording. In each pair, two random halves of a national sample received different wording. Sources: Rugg (1941) and Rasinski (1989).*

3. **Judgment.** The respondent combines what was recalled into an answer.
4. **Response.** The respondent fits that answer to the choices offered and decides whether to report it.

An error can enter at each step. A respondent may misread a term. The respondent may not remember, may estimate carelessly, or may choose an answer that sounds better than the truth. Good question writing makes each step easier.

The health question above supports each step. It defines its term, which helps comprehension. It limits the period to 30 days, which helps retrieval. It asks for a count of days, which makes the judgment concrete. It accepts any number from 0 to 30, so the respondent does not have to fit the answer to categories.

5.3 Wording Changes the Answers

Small changes in wording can produce large changes in results. Researchers have shown this with **split-ballot experiments**. In a split-ballot experiment, a random half of the sample receives one version of a question and the other half receives a second version. Because the halves were formed at random, a difference in the answers can be attributed to the wording.

Figure 5.1 shows two well-known results.

In 1940, one half of a national sample was asked whether the United States should *forbid* public speeches against democracy. Fifty-four percent said yes. The other half was asked whether the United States should *allow* such speeches. Seventy-five percent said no. To forbid and to not allow mean the same thing. The share that chose that position differed by 21 points.

In the 1980s, a national survey asked about government spending. Half the sample was asked about spending on "welfare." The other half was asked about spending on "assistance to the poor." In 1986, about 23 percent said too little was spent on welfare. About 63 percent said too little was spent on assistance to the poor.

Neither result means that respondents were careless. Words carry associations, and respondents react to them. A researcher cannot write a question that has no wording. The researcher can choose neutral words, keep the wording fixed over time, and report the exact question with every result.

5.4 Common Problems in Questions

Most weak questions have one of a small number of problems. Table 5.1 lists them with an example and a revision.

Table 5.1. *Eight common problems in survey questions, with a revision for each.*

Problem	Example	Revision
Double-barreled. Two questions in one.	``How satisfied are you with the cost and quality of your health care?''	Ask about cost and quality in two questions.
Leading. The wording suggests an answer.	``Do you agree that the county should do more to protect children's health?''	``Should the county spend more, less, or about the same on children's health programs?''
Loaded. A word carries a strong association.	``Do you support handouts for people who do not work?''	``Do you favor or oppose cash assistance for adults who are unemployed?''
Vague term. A word that people read differently.	``Do you exercise regularly?''	``In the past 7 days, on how many days did you exercise for at least 30 minutes?''
Double negative. Two negatives in one sentence.	``Do you disagree that smoking should not be allowed in parks?''	``Should smoking be allowed in public parks?''
Jargon. A term the respondent may not know.	``Have you used telehealth modalities?''	``In the past 12 months, have you had a medical visit by phone or video?''
Long recall period. The period is too long to remember.	``How many times did you see a doctor in the past five years?''	``In the past 12 months, how many times did you see a doctor?''
Assumed knowledge. The question assumes a fact.	``How would you rate the county's new clinic?''	First ask whether the respondent has visited the clinic.

A useful habit is to read each question aloud and ask what a respondent would need to know to answer it. If the answer includes a definition you did not supply, add it to the question.

5.5 Work With the Data: Question Reviewer

Table 5.1 lists eight problems. The tool checks a draft question for several of them. It uses fixed rules, so it can miss a problem and can flag wording that is fine.

The tool runs in a web browser at quant.marginoferrormedia.com/tools/question-reviewer.html. It needs no account and no software.

1. Press "Review the question" for the sample question. The tool marks two problems. Find each one in Table 5.1.
2. Rewrite the sample as two neutral questions. Review each one.
3. Press "Example 2" and "Example 3." Predict the problems before you read what the tool marks.
4. Type a question of your own and review it. Revise until the rules find nothing.

A question with no marks has passed a first check. It has not yet been tested. Only people from the population can show whether they understand the question as you intended. A later section of this chapter covers how to test with them.

5.6 The Answer Choices

A question is not complete without its answer choices. An **open-ended question** lets the respondent answer in any words. A **closed-ended question** offers a fixed set of choices.

Open-ended questions collect answers the researcher did not expect. They take longer to answer and have to be coded before they can be counted. Closed-ended questions are quick to answer and easy to count. They limit the respondent to the choices the researcher wrote.

Fixed choices have to meet two requirements. They should be **mutually exclusive**, so that no answer fits two choices. They should be **exhaustive**, so that every respondent can find a choice that fits. Age groups of "18 to 30" and "30 to 45" fail the first test for a respondent who is 30. A list of health insurance types without "none" fails the second.

Rating scales raise three further choices.

Number of points. Five to seven points work well for most rating scales. Fewer points lose detail. More points ask respondents for finer judgments than most can make.

Balance. A scale should offer the same number of positive and negative choices. "Excellent, very good, good, fair, poor" has three positive choices and two that are not. The health survey has used this scale for decades, and its results are well understood. A new scale should be balanced.

Labels. Label every point with words where possible. A scale with labels only at the two ends leaves respondents to decide what the middle points mean.

A final choice is whether to offer "don't know." Without it, respondents who have no opinion will often pick one anyway. With it, some respondents who do have an opinion will choose it to save effort. Offer "don't know" for questions of fact that a respondent might not know. Consider leaving it out for questions about the respondent's own experience.

5.7 Bias in the Answers

Some patterns in answers come from the respondent and not from the wording.

Social desirability bias is the tendency to give the answer that makes a good impression. Respondents tend to report more voting and exercise than they do, and less drinking and smoking. The bias is stronger when an interviewer is present and weaker on a self-completed form. It can be reduced with neutral wording and a credible promise that answers are confidential.

Acquiescence is the tendency to agree with statements regardless of content. A scale made only of agree-disagree items will inflate agreement. One remedy is to mix items worded in both directions, as described in Chapter 4. A second is to ask a direct question. "How would you rate your health?" is better than "My health is good. Agree or disagree?"

Satisficing is the tendency to give an answer that is good enough with little thought. A respondent who is tired or bored may pick the first reasonable choice or give the same rating to every item in a grid. Short questionnaires and varied formats reduce it. Chapter 7 covers how to detect it in the data.

5.8 The Order of Questions

A question can be affected by the questions that come before it. This is a **question-order effect**.

In 1948, Hyman and Sheatsley asked Americans about Communist reporters from other countries. Should they be allowed into the United States to report the news as they saw it? When this question came first, 36 percent said yes. A second version first asked whether American reporters should be allowed into Russia. When the question about Communist reporters followed that one, 73 percent said yes. Respondents who had just endorsed access for American reporters applied the same standard to the other side.

Order effects cannot be removed. They can be managed with a few rules.

1. Begin with questions that are easy and that relate to the stated topic.
2. Move from general questions to specific ones within a topic.
3. Group questions on the same topic and use a short sentence to introduce each group.
4. Place sensitive questions after the respondent has had time to build trust.
5. Place demographic questions such as age and income at the end.
6. Keep the order fixed across years if results will be compared.

A **filter question** routes respondents past questions that do not apply to them. A respondent who has not visited the county clinic should skip the questions that rate it. Online survey software applies this routing automatically, and the routing has to be tested like any other part of the questionnaire.

5.9 Test Before You Field

No researcher can predict every way a question will be read. A questionnaire has to be tested with people before it is used. Three methods are standard.

Expert review. People who know the topic and people who know survey design read the draft. They catch missing content and known wording problems.

Cognitive interviewing. A small number of people from the target population answer the draft questions and explain their thinking aloud. The interviewer asks follow-up questions such as "What does 'physical health' mean to you in this question?" Five to fifteen interviews usually reveal the main problems. This method tests comprehension directly, and it is the most useful of the three.

Pilot test. The full questionnaire is given to a small sample under the conditions of the real survey. The pilot shows how long the survey takes and where respondents quit. It also shows which questions are skipped and whether the routing works.

Table 5.2 shows how the county in the example could put its questionnaire together.

TABLE 5.2. *A plan for an example county health survey. Borrowed questions need less testing than new ones.*

Section	Source of the questions	Testing needed
General health and healthy days	National health survey, used as written	None for wording. Check that they fit the mode.
Access to care	National health survey, used as written	None for wording.
Use of the county clinic	New questions	Expert review, cognitive interviews, pilot.
Housing and food	Six-item food security scale from Chapter 4	None for wording.
Demographics	Census Bureau wording	None for wording.

Mode Matters

The same question can perform differently by phone, on paper, and on a screen. A long list of choices is hard to follow when read aloud. A grid of ratings is easy to fill in carelessly on a phone screen. When you borrow a question, check the mode in which it was tested. If your mode differs, include the question in your cognitive interviews.

Ethics Moment: Wording Can Be Used to Get a Result

The experiments in this chapter show that wording can move answers by 20 points or more. A sponsor who wants a particular result can choose the wording that produces it. A survey written this way is sometimes called a push poll or an advocacy survey. A researcher has two duties. The first is to write neutral questions even when a client would prefer otherwise. The second is to publish the exact wording of every question with the results, so that readers can judge it for themselves.

AI Reality Check: A Draft and a Reviewer, Not a Respondent

A language model can draft questions quickly and can find many of the problems in Table 5.1. Both uses save time. Two limits apply. A model tends to write questions that are long and formal, and they need editing for the reading level of your respondents. A model also cannot replace a cognitive interview. It can predict how a typical reader might interpret a question. It cannot tell you how a retired farmer or a recent immigrant in your county will read it. Use a model for drafting and for a first review, and then test the questions with people from the population you will survey.

5.10 Putting It All Together

A respondent answers a question in four steps. The steps are comprehension, retrieval, judgment, and response. Good wording makes each step easier. Split-ballot experiments show that small changes in wording and order can change results by 20 points or more.

Most weak questions are double-barreled, leading, loaded, vague, or too demanding of memory. Answer choices have to be mutually exclusive and exhaustive, and rating scales should be balanced and labeled. Social desirability, acquiescence, and satisficing affect answers regardless of the wording.

Borrow tested questions where they exist and use them as written. Test new questions with expert review, cognitive interviews, and a pilot.

5.11 Looking Ahead

A well-written questionnaire still has to reach the right people. Chapter 6 covers how a sample is drawn and what happens when part of the population has no chance of being included. The case is the 2020 Census, which counted the country accurately in total and missed several groups at high rates.

5.12 Further Reading and References

- Tourangeau, R., Rips, L. J., and Rasinski, K. (2000). *The Psychology of Survey Response.* Cambridge University Press. The source of the four-step model.
- Schuman, H., and Presser, S. (1981). *Questions and Answers in Attitude Surveys.* Academic Press. The standard collection of experiments on wording and order.
- Rugg, D. (1941). Experiments in wording questions: II. *Public Opinion Quarterly, 5*(1), 91 to 92.
- Rasinski, K. A. (1989). The effect of question wording on public support for government spending. *Public Opinion Quarterly, 53*(3), 388 to 394.
- Hyman, H. H., and Sheatsley, P. B. (1950). The current status of American public opinion. In J. C. Payne (Ed.), *The Teaching of Contemporary Affairs.* National Council for the Social Studies.
- Dillman, D. A., Smyth, J. D., and Christian, L. M. (2014). *Internet, Phone, Mail, and Mixed-Mode Surveys* (4th ed.). Wiley. A practical guide to questionnaire design.
- Willis, G. B. (2005). *Cognitive Interviewing: A Tool for Improving Questionnaire Design.* Sage.
- Centers for Disease Control and Prevention. *Behavioral Risk Factor Surveillance System Questionnaires.* The questionnaires for every year are public.

5.13 Key Terms

- **Acquiescence.** The tendency to agree with statements regardless of content.
- **Closed-ended question.** A question with a fixed set of answer choices.
- **Cognitive interviewing.** A test in which respondents explain their thinking as they answer draft questions.
- **Double-barreled question.** A question that asks two things at once.
- **Exhaustive.** A set of choices in which every respondent can find one that fits.
- **Filter question.** A question that routes respondents past questions that do not apply.
- **Leading question.** A question whose wording suggests an answer.

- **Mutually exclusive.** A set of choices in which no answer fits more than one.
- **Open-ended question.** A question the respondent answers in any words.
- **Pilot test.** A trial of the full questionnaire with a small sample.
- **Question-order effect.** A change in answers caused by the questions that came before.
- **Satisficing.** Giving an answer that is good enough with little thought.
- **Social desirability bias.** The tendency to give the answer that makes a good impression.
- **Split-ballot experiment.** A test in which random halves of a sample receive different versions of a question.

5.14 Exercises

Check Your Understanding

1. Name the four steps a respondent goes through to answer a question. Give one error that can enter at each step.
2. Identify the problem in each question. (a) "How often do you eat fruits and vegetables?" (b) "Don't you agree that clinics should stay open later?" (c) "Are you satisfied with your doctor and your pharmacy?"
3. Explain why a split-ballot experiment supports a causal claim about wording. Use the terms from Chapter 3.
4. The age groups on a form are "under 25, 25 to 40, 40 to 65, over 65." Which requirement do they fail?

Apply It

5. Rewrite each question from Exercise 2 so that the problem is removed.
6. Write a question that measures how many days in the past week a respondent ate breakfast. Include the answer choices. State which of the four steps your wording is meant to help.
7. Use the file `ch05-wording-experiments.csv`. For each of the three experiments, calculate the difference in percentage points between the two versions.

Think Deeper

8. The general health question uses an unbalanced scale and has been asked the same way for decades. Should the survey switch to a balanced scale? What would be gained and what would be lost?
9. A client asks you to word a question so that support for a policy looks as high as possible. Describe what you would say and what you would offer to do in place of the request.

10. Draft three questions on a topic you care about. Ask two people to answer them and to explain what each question meant to them. Report what you learned and how you would revise the questions.

CHAPTER 6

Sampling and Coverage

Where this chapter fits. Step 5 of the research process is to plan the sample. This chapter covers whom to ask and how to select them.

6.1 An Accurate Total and an Uneven Count

The 2020 Census counted 331.4 million people. Two years later, the Census Bureau reported how accurate the count had been. For the nation as a whole, the result was good. The estimated net error for people living in households was 0.24 percent, an undercount of about 782,000 people. That figure was within the range that chance alone could produce.

The national figure combined errors in both directions. The bureau estimated that the census missed 18.8 million people. It also counted millions of people twice or in the wrong place, and it filled in records for households that did not respond. The misses and the extra counts nearly canceled out in the total.

They did not cancel out for every group. Figure 6.1 shows the net error for several groups.

The census undercounted Hispanic residents by 4.99 percent and Black residents by 3.30 percent. It undercounted American Indians and Alaska Natives living on reservations by 5.64 percent. It undercounted children under age 5 by 2.79 percent and renters by 1.48 percent. It overcounted non-Hispanic white residents by 1.64 percent, Asian residents by 2.62 percent, and homeowners by 0.43 percent.

Census counts decide how seats in Congress are divided among the states. They also guide the distribution of federal funds. A group that is undercounted receives less of both.

The census tries to count everyone and still misses people at different rates. A survey, which contacts only a sample, faces the same problem at every step. This chapter covers how samples are drawn, how much a sample result can vary, and how a sample can leave people out.

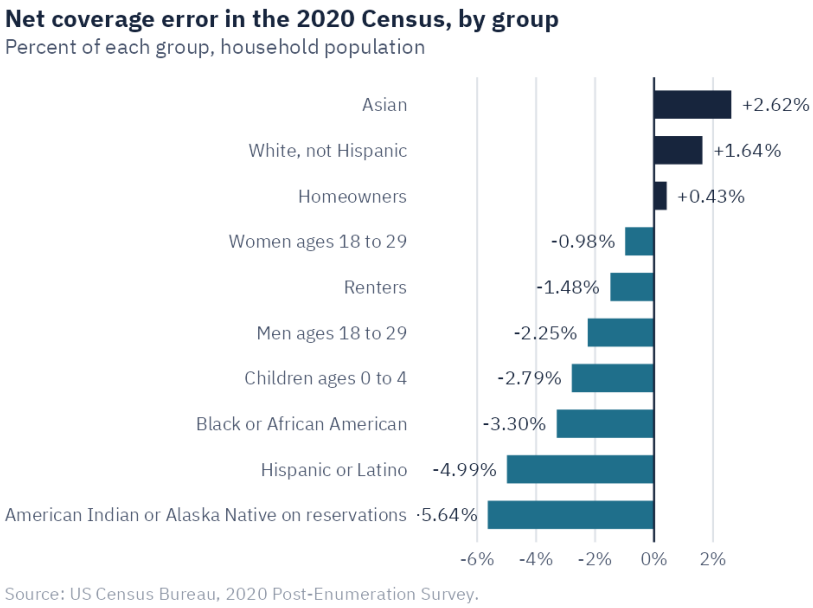


FIGURE 6.1. *Estimated net coverage error in the 2020 Census for selected groups. Negative values are undercounts and positive values are overcounts. Source: US Census Bureau, 2020 Post-Enumeration Survey.*

6.2 Population, Frame, and Sample

Three terms from Chapter 2 organize this chapter.

The **target population** is the group the study aims to describe. The **sampling frame** is the list from which the sample is drawn. The **sample** is the set of units selected from the frame.

A **census** attempts to collect data from every unit in the population. A **sample survey** collects data from part of it. A census is not always more accurate than a sample. A census of a large population needs a very large field staff, and errors in the field are hard to control at that scale. A well-run sample can spend more effort on each case. The Census Bureau measures the accuracy of the census with a sample survey.

6.3 Probability Sampling

In a **probability sample**, every unit in the frame has a known chance of selection that is greater than zero. Chance decides which units are selected. This property allows the researcher to calculate how much the result could vary from sample to sample.

Four designs are common.

Simple random sampling. Every unit has the same chance of selection, and every possible sample of a given size is equally likely. It is the simplest design to analyze. It requires a complete list of the population, which is often not available.

Systematic sampling. The researcher picks a random starting point and then selects every kth unit on the list. To draw 100 from a list of 5,000, select every 50th name. The method is easy to carry out. It works well unless the list has a repeating pattern that matches the interval.

Stratified sampling. The researcher divides the population into groups, called strata, and draws a random sample within each one. A county survey could stratify by district so that every district is represented. A researcher can also sample a small group at a higher rate so that it has enough cases to analyze. Weights then restore each group to its true share.

Cluster sampling. The researcher selects groups of units and then surveys units within the selected groups. A national survey may select counties, then blocks within those counties, then households within those blocks. This is **multistage sampling**. It lowers the cost of travel and does not need a list of every person in the country. People in the same cluster tend to be alike. A cluster sample is therefore less precise than a simple random sample of the same size.

Table 6.1 compares the four.

Table 6.1. *Four probability sampling designs.*

Design	How units are selected	Main advantage	Main limit
Simple random	Every unit has an equal chance	Simple to analyze	Needs a complete list
Systematic	Every kth unit after a random start	Easy to carry out	Patterns in the list can bias it
Stratified	Random samples within groups	Each group is represented	Needs group membership before sampling
Cluster	Random groups, then units within them	Lower cost, no full list needed	Less precise for the same sample size

6.4 Samples Without Known Chances

In a **nonprobability sample**, the chance that a unit is selected is unknown. Three kinds are common.

A **convenience sample** takes whoever is easy to reach, such as shoppers at one store or students in one class. A **quota sample** recruits until it has a set number of people in each category, such as 50 men and 50 women. A **snowball sample** asks each participant to refer others. It is used to reach groups that have no list, such as people who use a particular drug.

Most online surveys today use **opt-in panels**. People sign up to take surveys for payment, and researchers draw from the panel. The panel is a nonprobability sample, because people chose to join. Panel companies use quotas and weights to make the sample resemble the population on age, sex, race, and education. The sample can match the population on those variables and still differ on the variable being studied.

Nonprobability samples have uses. They suit pilot tests, experiments with random assignment, and studies of groups that cannot be listed. They do not support a margin of error, because the theory behind it assumes that chance decided who was selected.

The best-known failure of a large nonprobability sample is from 1936. The *Literary Digest* magazine mailed about 10 million ballots before the presidential election and received about 2.4 million back. It drew its mailing list largely from telephone directories and automobile registrations. The poll predicted that Alf Landon would win with 57 percent of the vote. Franklin Roosevelt won with about 61 percent. The frame left out many voters with lower incomes, and the people who returned ballots differed from those who did not. A sample of 2.4 million did not correct for either problem.

6.5 How Much a Sample Can Vary

A probability sample supports a statement about **sampling error**. For a proportion from a simple random sample, the **margin of error** at the 95 percent level is calculated as follows.

$$\text{margin of error} = 1.96 \times \sqrt{\frac{p(1-p)}{n}}$$

In the formula, p is the sample proportion and n is the sample size. Suppose a survey of 400 adults finds that 30 percent rent their homes.

1. Multiply p by $1 - p$. The result is $0.30 \times 0.70 = 0.21$.
2. Divide by n. The result is $0.21/400 = 0.000525$.
3. Take the square root. The result is 0.0229.
4. Multiply by 1.96. The margin of error is 0.045, or 4.5 percentage points.

The survey's estimate is 30 percent, plus or minus 4.5 points. If the survey were repeated many times, about 95 percent of such intervals would contain the true proportion.

Figure 6.2 shows two facts about sample size.

First, the gain from each added case shrinks. Going from 100 to 400 cases cuts the margin from 9.8 points to 4.9. Going from 1,000 to 2,500 cuts it from 3.1 to 2.0. To cut the margin in half, the sample has to be four times as large.

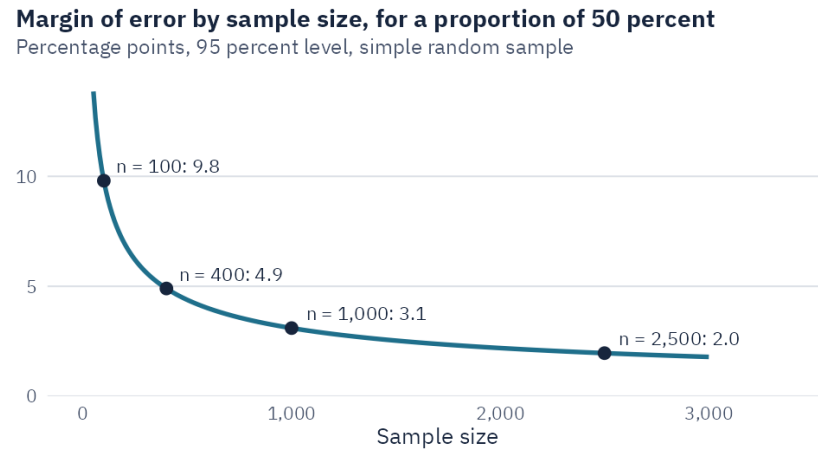


FIGURE 6.2. *Margin of error for a proportion of 50 percent, by sample size, for a simple random sample. The margin falls quickly up to about 1,000 cases and slowly after that.*

Secondly, the size of the population does not appear in the formula. A sample of 1,000 gives about the same margin for a city of 200,000 as for a country of 300 million. A sample is more precise when it has more units. It does not depend on what share of the population they represent, unless that share is large.

Subgroups Have Their Own Sample Sizes

A survey of 1,000 people has a margin of about 3 points for the whole sample. If 120 of them are renters under age 30, the margin for that subgroup is about 9 points. Plan the sample size for the smallest group you intend to report.

6.6 Coverage and Nonresponse

The margin of error describes sampling error only. Two other sources often matter more.

Coverage. A frame can leave out part of the population. The census builds its frame from a master list of addresses. Some homes are harder to list. Examples are a household in a converted garage, a family doubled up in one apartment, and a rural home with no street address. The undercount of renters, young children, and residents of reservations follows from where the frame is weakest.

Nonresponse. A unit can be on the frame, be selected, and still not respond. The **response rate** is the share of selected units that complete the survey. A low response rate does not by itself bias the result. Bias arises when the people who respond differ from those who do not on the variable being measured. In

the census, households that did not respond were visited by field staff. When that failed, the bureau used information from neighbors or from administrative records, or it filled in the record with a statistical estimate.

Researchers adjust for both problems with **weights**. If young adults are 20 percent of the population and 10 percent of the respondents, each young respondent is counted twice. Weighting corrects the sample on the variables used to build the weights. It does not correct a difference on a variable that the weights did not include.

6.7 Work With the Data: Sampling Explorer

This chapter has described two different problems. A sample varies from one draw to the next, and a frame can leave people out. The tool shows both in one simulated population of 10,000 households. The true share with a housing cost problem is 23.8 percent.

The tool runs in a web browser at quant.marginoferrormedia.com/tools/sampling-explorer.html. It needs no account and no software.

1. Leave the sample size at 400 and the share of renters missing at 0 percent. Press "Draw 300 samples." Read the middle 95 percent of the estimates.
2. Raise the sample size and draw again. The estimates move closer together. Their average stays near 23.8 percent.
3. Set the share of renters missing from the frame to 40 percent. Draw again. Read the average and the bias.
4. With renters still missing, raise the sample size once more and draw.

In the last step the estimates move closer together around the wrong value. With 40 percent of renters missing, the average is about 21 percent and the bias is close to 3 points at any sample size. A larger sample reduces variable error. It does not repair a frame that leaves people out.

6.8 How the Census Checks Itself

The Census Bureau estimates its own error with a method called **dual-system estimation**. Biologists use the same idea to estimate the size of a wildlife population, and they call it capture and recapture.

After the 2020 Census, the bureau carried out a separate survey of about 10,000 census blocks. Interviewers listed every household in those blocks independently of the census. The bureau then matched the survey records to the census records, person by person.

The logic uses three numbers. Suppose that in one area the census counted 900 people. The independent survey found 800 people. Of those 800, 720 could be matched to a census record.

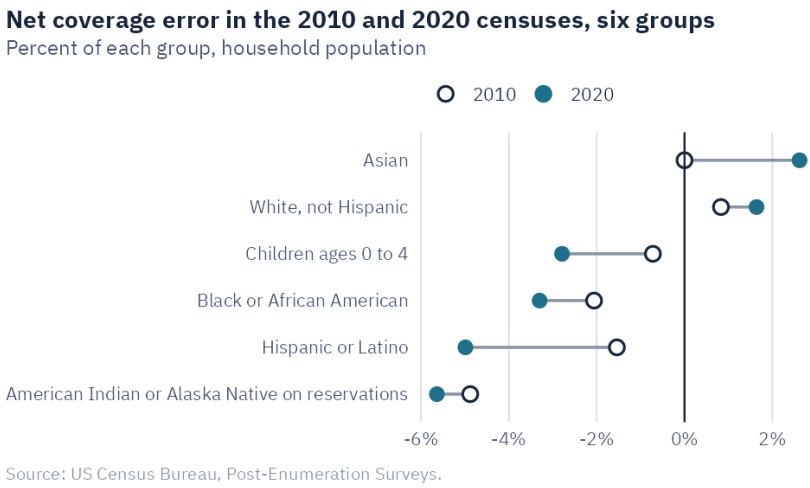


FIGURE 6.3. *Estimated net coverage error in the 2010 and 2020 censuses for six groups. The undercount of Hispanic residents and of young children was larger in 2020 than in 2010. Source: US Census Bureau, 2010 and 2020 Post-Enumeration Surveys.*

1. Calculate the share of survey people the census found. The share is $720/800 = 0.90$.
2. Treat that share as the census coverage rate for the area. The census found about 90 percent of the people.
3. Divide the census count by the coverage rate. The estimate of the true population is $900/0.90 = 1{,}000$.
4. Compare with the census count. The census missed about 100 people, an undercount of 10 percent.

The bureau applies this logic within groups defined by age, race, and housing tenure. The results are the estimates in Figure 6.1. The survey is itself a sample, and so each estimate has its own sampling error.

Figure 6.3 compares 2020 with 2010. The undercount of Hispanic residents rose from 1.54 percent to 4.99 percent. The undercount of children under 5 rose from 0.72 percent to 2.79 percent. The 2020 Census was conducted during a pandemic, which disrupted field work. The national total remained accurate, and the errors for groups grew.

Ethics Moment: Who Bears the Cost of a Weak Frame

Coverage error does not fall evenly. The people a frame misses are more often renters, young children, people of color, and people in rural areas. When a survey guides how money or services are distributed, those groups

receive less. A researcher can reduce the harm in three ways. Spend more of the budget on the groups that are hardest to reach. Report results for those groups with their own margins of error. State plainly who the frame could not reach.

AI Reality Check: Synthetic Respondents Are Not a Sample

Some companies now offer "synthetic respondents." A language model is asked to answer survey questions as if it were a person with a given age, income, and location. The answers are fast and cheap. They are not a sample of any population. The model produces the answers that its training data make likely for that description. It cannot report an opinion that formed last month, and it tends to produce answers that vary less than those of real people. Synthetic answers can help test whether a questionnaire works. They should not be reported as the views of a population.

6.9 Putting It All Together

A probability sample gives every unit in the frame a known chance of selection. Simple random, systematic, stratified, and cluster designs are the main types. A probability sample supports a margin of error, which depends on the sample size and not on the size of the population. Nonprobability samples, including opt-in online panels, do not support a margin of error.

The margin of error covers sampling error only. Coverage error arises when the frame leaves out part of the population. Nonresponse error arises when those who respond differ from those who do not. Weights can correct for differences on the variables they include.

The 2020 Census shows how errors can cancel in a total and remain large for groups. The national net error was 0.24 percent. The undercount was 4.99 percent for Hispanic residents and 5.64 percent for American Indians and Alaska Natives living on reservations.

6.10 Looking Ahead

Once a sample is drawn and a questionnaire is ready, data collection begins. Chapter 7 covers what happens between the first response and a clean data file. The case is an online panel in which some of the responses did not come from attentive people.

6.11 Further Reading and References

- US Census Bureau (2022). *Census Bureau Releases Estimates of Undercount and Overcount in the 2020 Census* (Release CB22-CN.02). The source for the coverage estimates in this chapter.
- Khubba, S., Heim, K., and Hong, J. (2022). *National Census Coverage Estimates for People in the United States by Demographic Characteristics.* US Census Bureau. The technical report on the Post-Enumeration Survey.
- Cohn, D., and Passel, J. S. (2022). *Key Facts About the Quality of the 2020 Census.* Pew Research Center. A summary of omissions, duplicates, and filled-in records.
- Lohr, S. L. (2021). *Sampling: Design and Analysis* (3rd ed.). CRC Press. A full textbook on sampling designs.
- Squire, P. (1988). Why the 1936 Literary Digest poll failed. *Public Opinion Quarterly, 52*(1), 125 to 133.
- Mercer, A., Lau, A., and Kennedy, C. (2018). *For Weighting Online Opt-In Samples, What Matters Most?* Pew Research Center.

6.12 Key Terms

- **Census.** An attempt to collect data from every unit in a population.
- **Cluster sampling.** Selecting groups of units and then surveying units within them.
- **Convenience sample.** A sample of whoever is easy to reach.
- **Dual-system estimation.** Estimating coverage by matching two independent counts of the same area.
- **Margin of error.** The amount a sample result is expected to vary because of sampling, at a stated level of confidence.
- **Multistage sampling.** Cluster sampling carried out in more than one step.
- **Nonprobability sample.** A sample in which the chance of selection is unknown.
- **Opt-in panel.** A group of people who signed up to take surveys.
- **Probability sample.** A sample in which every unit in the frame has a known chance of selection above zero.
- **Quota sample.** A sample recruited until each category has a set number of people.
- **Response rate.** The share of selected units that complete the survey.
- **Simple random sampling.** A design in which every unit has an equal chance of selection.
- **Snowball sample.** A sample built by asking participants to refer others.
- **Stratified sampling.** Drawing random samples within groups of the population.

- **Systematic sampling.** Selecting every *k*th unit after a random start.
- **Weights.** Numbers that make some respondents count more or less so the sample matches the population.

6.13 Exercises

Check Your Understanding

1. State the difference between a probability sample and a nonprobability sample. Which one supports a margin of error?
2. The 2020 Census had a net error of 0.24 percent and missed 18.8 million people. Explain how both statements can be true.
3. Why does a sample of 1,000 give about the same margin of error for a city as for a country?
4. A survey has a response rate of 9 percent. Under what condition would its results still be unbiased?

Apply It

5. A survey of 900 adults finds that 42 percent have a regular doctor. Calculate the margin of error at the 95 percent level. Show each step.
6. A census counts 4,500 people in an area. An independent survey finds 1,000 people there, and 920 of them match a census record. Estimate the true population and the percent undercount. Show each step.
7. A county has 60,000 adults in the city and 20,000 in rural areas. You can afford 800 interviews and want a margin of about 5 points for each area. Propose a stratified design and explain how you would weight the results.

Think Deeper

8. The census overcounted homeowners and undercounted renters. List three features of renting that could explain the difference. Which source of error from Chapter 2 does each one involve?
9. An online panel is weighted to match the population on age, sex, race, and education. A client says it is "representative." What would you want to know before you agreed?
10. Use the file `ch06-census-coverage.csv`. A state receives $2,000 per resident each year from programs tied to census counts. Estimate what a 4.99 percent undercount costs a community of 50,000 Hispanic residents over ten years. State what your estimate assumes.

PART IV

Data, Analysis, and Reporting

CHAPTER 7

Collecting and Cleaning Data

Where this chapter fits. Step 6 of the research process is to collect the data, and step 7 begins with preparing the file. This chapter covers both.

7.1 A Thousand Completed Surveys

A housing nonprofit wants to know how many renters in its region spend more than they can afford on rent. It buys 1,000 completed surveys from an online panel company. The survey takes about seven minutes. Within two days, the file arrives with 1,000 rows.

The first tabulation shows that 38.9 percent of respondents report a rent burden. Before the staff reports that figure, an analyst looks at the file more closely. Some respondents finished in under a minute. Some gave the same answer to every item in a six-item grid, including an item that was worded in reverse. Several typed the same sentence into the open-ended box, and one of those sentences begins "As an AI language model."

The panel in this chapter is simulated. It contains 880 attentive respondents and 120 planted cases. Seventy of the planted cases are careless people who rushed through. Fifty are automated programs, or bots. The file was mathematically modeled to reflect patterns that researchers have reported in real opt-in panels. Because the planted cases are known, a reader can test each screening rule and see what it catches and what it misses.

The problem is real. The Pew Research Center tested several opt-in panels in 2019. It estimated that 4 to 7 percent of respondents gave bogus answers (Kennedy et al., 2020). A later Pew study asked opt-in respondents whether they were licensed to operate a nuclear submarine. Among those under age 30, 12 percent said yes (Mercer, Kennedy, and Keeter, 2024). The true share of adults with that license is close to zero.

This chapter covers the step between a questionnaire and a dataset. It begins with how data are collected and ends with a clean, documented file.

7.2 Modes of Collection

The way a survey reaches people is its **mode**. Table 7.1 compares the four main modes.

TABLE 7.1. *The four main modes of survey data collection.*

Mode	Strengths	Limits
In person	Highest response rates. An interviewer can explain questions and show materials.	Highest cost. Respondents may shade answers on sensitive topics.
Telephone	An interviewer can clarify questions. Faster than in person.	Response rates have fallen sharply. Long lists of choices are hard to follow by ear.
Mail	Reaches almost every household through address lists. Private for sensitive questions.	Slow. No one can clarify a question. Routing has to be simple.
Web	Lowest cost and fastest. Routing is automatic. Private.	Leaves out people without internet access. Opt-in panels are not probability samples.

Many surveys now combine modes. A common design mails an invitation to a random sample of addresses and asks people to respond online. Those who do not respond receive a paper form. This design uses a probability sample and still collects most responses at web cost.

Mode affects answers as well as cost. Respondents report more sensitive behavior on a self-completed form than to an interviewer. A researcher who changes mode between two years of a survey should expect some of the change in results to come from the mode.

7.3 Consent and Confidentiality

Before a respondent answers, the researcher has to obtain **informed consent**. The respondent is told the purpose of the study, how long it will take, any risks, and that taking part is voluntary. The respondent may skip any question and may stop at any time.

Two terms describe how answers are protected, and they are often confused.

A survey is **anonymous** if no one, including the researcher, can link answers to a person. A survey is **confidential** if the researcher can make the link and promises not to reveal it. A panel survey that pays respondents is confidential and not anonymous, because the panel company knows who answered. A researcher should state which protection applies and should not promise anonymity that the design cannot deliver.

Most universities and many agencies require that a review board approve a study with human participants before data collection begins. The board checks that risks are small, that consent is informed, and that the data will be protected.

7.4 Monitoring the Field

Data collection needs attention while it is under way. Three numbers are worth checking daily.

The **response rate** is the share of sampled units that complete the survey. It applies to probability samples. For an opt-in panel, no response rate can be calculated, because no defined sample was invited. The **break-off rate** is the share of respondents who start and do not finish. A high break-off rate at one question often means that the question is confusing or intrusive. The **completion time** shows whether the survey is as long as planned and whether some respondents are moving too fast to have read it.

Incentives raise response. A small payment sent in advance, with no condition attached, generally raises response more than a larger payment promised for completion. In opt-in panels, payment is the reason people take part. It is also the reason that some people and some programs complete surveys as fast as possible.

7.5 Four Signs of a Low-Quality Response

Researchers use several checks to find responses that should not be trusted. Each one is a rule applied to the data.

Speeding. A respondent who finishes far faster than the typical respondent has probably not read the questions. A common rule flags anyone who finishes in less than 40 percent of the median time.

Straightlining. A respondent gives the same answer to every item in a grid. When the grid includes an item worded in reverse, identical answers contradict each other.

Attention checks. The survey includes a question with one correct answer, such as "Select 'somewhat disagree' for this item." A respondent who fails it was not reading.

Open-ended answers. A text box collects an answer in the respondent's own words. Empty answers, one-word answers, off-topic sentences, and identical sentences across respondents are all signs of a problem.

Figure 7.1 shows the completion times in the simulated file.

No single check finds every bad case. Table 7.2 shows what each rule catches in the simulated file.

Table 7.2. *Cases flagged by each screening rule in the simulated panel of 1,000. The panel contains 70 careless respondents, 50 bots, and 880 attentive respondents. Simulated data.*

Rule	Careless caught (of 70)	Bots caught (of 50)	Attentive flagged (of 880)
Speeding	59	23	2
Straightlining	45	0	3
Failed attention check	40	21	25

TABLE 7.2. *Cases flagged by each screening rule in the simulated panel of 1,000. The panel contains 70 careless respondents, 50 bots, and 880 attentive respondents. Simulated data.*

Rule	Careless caught (of 70)	Bots caught (of 50)	Attentive flagged (of 880)
Open-ended answer	44	30	0
Any one flag	68	44	30
Two or more flags	62	23	0

The table shows a trade-off. A rule that removes anyone with a single flag catches most of the planted cases. It also removes some attentive respondents, such as a fast reader or someone who misread the attention check. A rule that requires two flags removes fewer attentive respondents and lets more bad cases through.

A rule that removes attentive respondents has a cost. If fast readers are younger, a strict speed rule removes younger people at a higher rate, and the sample changes. Many researchers therefore require at least two flags before removing a case. They report how many cases each rule removed.

7.6 What Screening Does to the Result

The planted cases did not answer the rent question the way attentive respondents did. Bots that answer at random choose "yes" about half the time, and in this file they were set to say yes more often. Careless respondents are close to half and half.

Figure 7.2 shows the effect. With all 1,000 cases, the estimate is 38.9 percent. Among the 880 attentive respondents, it is 35.1 percent. After the two-flag rule, the estimate is 36.5 percent. That is closer to the attentive figure and still above it, because some planted cases passed the rule.

Twelve percent of the file was bad, and the estimate moved by about 4 points. Bad cases do the most damage when the true share is far from 50 percent. Random answers pull every estimate toward the middle. A rare condition will appear more common than it is, which is the pattern in the Pew submarine result.

7.7 Work With the Data: Response Quality Screener

The section above applied one screening rule to the simulated panel. The tool lets you try other rules on the same 1,000 cases. Because the file is simulated, the tool can show which cases each rule removed.

The tool runs in a web browser at quant.marginoferrormedia.com/tools/quality-screener.html. It needs no account and no software.

1. Start with all four flags selected and a threshold of two flags. The tool removes 85 cases. Read the table to see how many were attentive, careless, or bots.

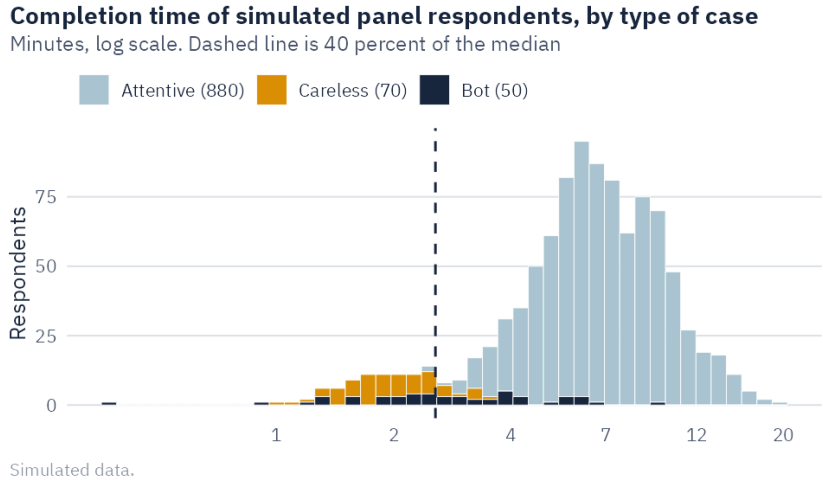


FIGURE 7.1. *Completion time for 1,000 simulated panel respondents, by the type of case. The dashed line marks 40 percent of the median time. Most careless respondents and about half of the bots fall below it. A few attentive respondents do too. The data are simulated.*

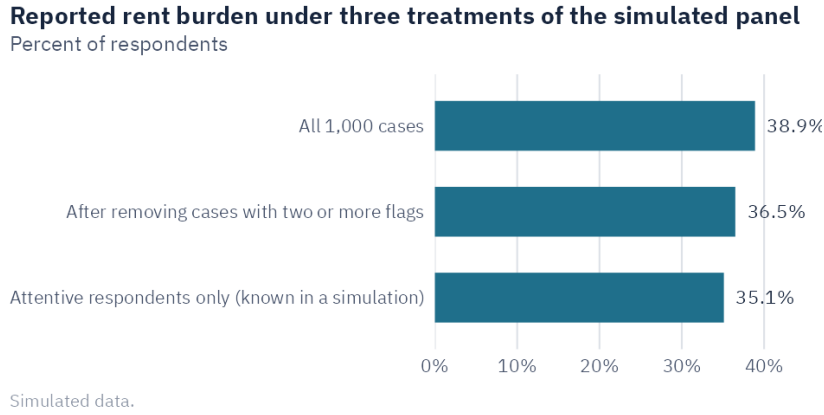


FIGURE 7.2. *Share of respondents reporting a rent burden in the simulated panel under three treatments of the data. The estimate from the attentive respondents alone is the correct value for this simulated file. Simulated data.*

2. Change the threshold to one flag. The tool now removes 142 cases, and 30 of them are attentive respondents.
3. Change the threshold to three flags. Fewer attentive respondents are removed, and more planted cases stay in the file.
4. Return to two flags. Turn off one flag at a time and note which planted cases each flag was catching.

Each rule is a trade. A strict rule removes more bad cases and more good ones. A lenient rule keeps the good cases and leaves bad ones in. Compare the rent estimate after screening with the estimate for attentive respondents only. No rule in the tool closes the gap completely.

The file is `ch07-simulated-panel.csv` at quant.marginoferrormedia.com/data.

7.8 Cleaning the File

The screen for low-quality responses is one part of **data cleaning**. The full process has several steps, and the order matters.

1. **Keep the raw file.** Save the file exactly as it arrived and never edit it. All cleaning is done on a copy, with a script or a written log.
2. **Write a codebook.** A **codebook** lists every variable, its question wording, its allowed values, and the codes for missing data.
3. **Check ranges.** Each variable should contain only allowed values. An age of 212 or a rating of 7 on a five-point scale is an entry error.
4. **Check consistency.** Answers should agree with one another. A respondent who reports no children should not report a child's age.
5. **Find duplicates.** The same person may appear twice. Check respondent identifiers and, in web surveys, repeated device or network addresses.
6. **Apply the screening rules.** Use the rules for speeding, straightlining, attention, and open-ended answers.
7. **Examine missing data.** Count the missing values for each variable and check whether they are concentrated in one group.
8. **Record every decision.** A **cleaning log** states each rule, the number of cases it affected, and the reason for it.

Missing data needs care. When a respondent skips a question, the simplest approach is to leave that respondent out of any analysis that uses the question. This is acceptable when few values are missing and they are not concentrated in one group. If one in four respondents skipped the income question, the people who answered it may not represent the rest. A researcher should report the share missing for every key variable.

An **outlier** is a value far from the others. An outlier can be an error, such as a rent of $85,000 a month that was meant to be $850. It can also be a real and

unusual case. Check outliers against other answers before changing them. Do not remove a value only because it is unusual.

Decide the Rules Before You See the Results

Set the screening rules before you look at the main results. A researcher who tries several rules and keeps the one that gives the preferred answer has chosen the result. Write the rules down in advance, apply them once, and report them.

Ethics Moment: Removing People From the Data

Every screening rule removes some real people who answered in good faith. Speed rules remove fast readers. Attention checks are failed more often by people reading in a second language. Rules about open-ended answers penalize people who write little. A researcher has a duty to check who is being removed. Compare the removed cases with the kept cases on age, education, and language. If one group is removed at a much higher rate, the rule is changing the sample, and the report should say so.

AI Reality Check: Bots That Pass the Checks

The checks in this chapter were designed for careless people and simple bots. Programs built on language models are harder to detect. They can read a question, give consistent answers, pass an attention check, and write a plausible open-ended answer. Westwood (2025) built such a program and reported that it passed nearly all of the standard checks. No screening rule is a full defense. The stronger protection is in the sample. A panel recruited from a random sample of addresses gives a fraudulent respondent no easy way to join. When a study has to use an opt-in panel, use several checks, expect some bad cases to remain, and report the limits.

7.9 Putting It All Together

Mode affects cost, coverage, and the answers themselves. Before collecting data, a researcher obtains informed consent and states whether answers are anonymous or confidential.

Online opt-in panels contain some responses from careless people and automated programs. Four checks help find them. They are speeding, straightlining, attention checks, and a review of open-ended answers. No single check is enough, and every check removes some attentive respondents. In the simulated file, 12 percent of the cases were planted, and they moved the main estimate by about 4 percentage points.

Data cleaning starts with an untouched raw file and ends with a codebook and a log of every decision. Screening rules should be set before the results are known.

7.10 Looking Ahead

The file is now clean. Chapter 8 covers the last two steps, which are the analysis plan and the report. The case is a file of 534 workers. It supports several different estimates of the pay gap between women and men, depending on choices the analyst makes.

7.11 Further Reading and References

- Kennedy, C., Hatley, N., Lau, A., Mercer, A., Keeter, S., Ferno, J., and Asare-Marfo, D. (2020). *Assessing the Risks to Online Polls From Bogus Respondents.* Pew Research Center.
- Mercer, A., Kennedy, C., and Keeter, S. (2024). *Online Opt-In Polls Can Produce Misleading Results, Especially for Young Adults and Hispanic Adults.* Pew Research Center. The source of the submarine result.
- Westwood, S. J. (2025). The potential existential threat of large language models to online survey research. *Proceedings of the National Academy of Sciences, 122.*
- Meade, A. W., and Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437 to 455. A comparison of methods for finding careless respondents.
- Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*(3), 213 to 236. The source of the term satisficing.
- Dillman, D. A., Smyth, J. D., and Christian, L. M. (2014). *Internet, Phone, Mail, and Mixed-Mode Surveys* (4th ed.). Wiley.

7.12 Key Terms

- **Anonymous.** No one, including the researcher, can link answers to a person.
- **Attention check.** A question with one correct answer, used to test whether a respondent is reading.
- **Break-off rate.** The share of respondents who start a survey and do not finish.
- **Cleaning log.** A record of each cleaning rule, the cases it affected, and the reason for it.

- **Codebook.** A list of every variable with its wording, allowed values, and missing-data codes.
- **Confidential.** The researcher can link answers to a person and promises not to reveal the link.
- **Data cleaning.** The process of checking and correcting a data file before analysis.
- **Informed consent.** Agreement to take part after being told the purpose, risks, and voluntary nature of a study.
- **Missing data.** Values that are absent because a respondent skipped a question or broke off.
- **Mode.** The way a survey reaches respondents, such as mail, telephone, or web.
- **Outlier.** A value far from the other values of a variable.
- **Speeding.** Completing a survey far faster than the typical respondent.
- **Straightlining.** Giving the same answer to every item in a grid.

7.13 Exercises

Check Your Understanding

1. Explain the difference between an anonymous survey and a confidential one. Which applies to a paid online panel?
2. Why is a straightlined grid more informative when one item is worded in reverse?
3. Random answers pull an estimate toward 50 percent. Explain why this makes a rare condition look more common.
4. Why should the raw data file never be edited?

Apply It

5. A survey has a median completion time of 9 minutes. At what time would the 40 percent rule flag a respondent for speeding? Show the calculation.
6. Use the file `ch07-simulated-panel.csv`. Apply the speeding rule and the attention check. How many cases does each one flag, and how many are flagged by both?
7. A file has 1,200 cases. A screening rule removes 96 of them. The estimate of interest is 22 percent before the rule and 19 percent after. Write two sentences for the methods section of a report that state what was done.

Think Deeper

8. A screening rule removes 6 percent of respondents with a college degree and 15 percent of respondents without one. What would you check next, and what would you report?

9. A colleague proposes trying five screening rules and keeping the one that gives the cleanest result. Explain the problem with this plan and propose an alternative.
10. Programs built on language models can now pass most attention checks. Describe two changes to the design of a study that would reduce the risk without relying on a screening rule.

CHAPTER 8

The Analysis Plan and the Report

Where this chapter fits. Step 7 of the research process ends with analysis and interpretation, and step 8 is the report. This chapter covers how to plan the analysis before seeing the data and what a report has to contain.

8.1 One File, Several Answers

In May 1985, the Census Bureau's monthly household survey asked workers about their pay. Economists have used a sample of 534 of those workers as a teaching dataset for decades. The file records each worker's pay, sex, years of education, years of work experience, occupation, union membership, and a few other characteristics. Pay is expressed as an annual figure, calculated from the hourly wage at 2,000 hours a year.

The file contains 245 women and 289 men. Suppose an analyst is asked one question. How large was the pay gap between women and men?

The file supports several answers.

- Women's mean pay was $15,758 and men's was $19,990. By this measure, women earned 79 cents for each dollar men earned.
- Women's median pay was $13,600 and men's was $17,860. By this measure, the figure is 76 cents.
- Among workers with the same education and experience, women earned about 22.7 percent less than men.
- Among workers with the same education, experience, occupation, union status, and region, women earned about 19.3 percent less.

Each figure is calculated correctly. Each one answers a slightly different question. An analyst who runs all of them and reports one has made a choice that the reader cannot see.

This chapter covers the last two steps of a study. The first is the analysis plan, which is written before the data are examined. The second is the report, which tells readers what was done.

The file is from 1985. It is used here to show how analysis choices affect a result. It does not describe pay today.

8.2 Write the Plan First

An **analysis plan** states how the data will be analyzed. It is written before the analyst looks at the results. A plan answers five questions.

1. **What is the outcome?** Name the variable and how it is measured. Here it is annual pay.
2. **What is the comparison?** Name the groups or the relationship of interest. Here it is women compared with men.
3. **What statistic will summarize it?** State whether the plan uses a mean, a median, a ratio, or a model.
4. **What will be held constant?** List any variables the analysis will adjust for, with the reason for each.
5. **Which cases are included?** State any rules for excluding cases, as described in Chapter 7.

The plan connects the analysis to the research question from Chapter 1. A different question leads to a different plan. "How much less do women earn than men?" calls for a comparison with no adjustment. "Are women paid less than men for the same work?" calls for an adjustment for the kind of work. Neither question is the correct one. They are different questions, and a report should say which one it answers.

8.3 Choosing a Summary

The first choice is the statistic.

Figure 8.1 shows that pay is skewed. Most workers are grouped at the lower end, and a few earn far more. The highest value in the file is $89,000, which is more than five times the median.

In a skewed distribution, the mean is pulled toward the long tail. The median is not. For pay, income, and wealth, the median usually describes the typical case better than the mean. A plan should state which one it will use and should not switch after the results are known.

The second choice is how to express the gap. The same two medians can be reported in three ways.

- As a **difference**, women earned $4,260 less.
- As a **ratio**, women earned 76 cents for each dollar.
- As a **percent difference**, women earned 24 percent less.

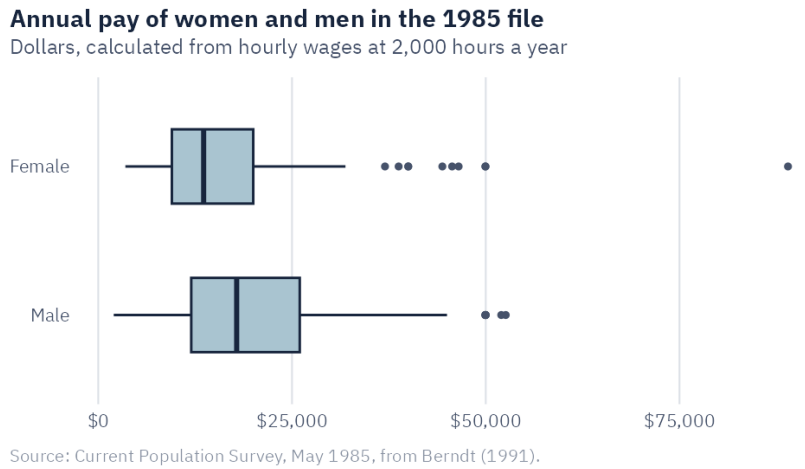


FIGURE 8.1. *Distribution of annual pay for 245 women and 289 men in the 1985 file. The box spans the middle half of each group and the line inside it marks the median. Both distributions have a long upper tail. Source: Current Population Survey, May 1985, from Berndt (1991).*

The three statements describe one fact. The ratio and the percent difference can be compared across years, because they do not depend on the price level. The dollar difference is easier for a general reader to picture. A report can give more than one form, and it should label each one.

8.4 What "Adjusted" Means

Women and men in the file differ in more than pay. They work in different occupations, and men are more likely to belong to a union. An analyst may want to compare women and men who are alike in these respects.

A **regression model** does this. It estimates how the outcome changes with one variable while the other variables in the model are held constant. Chapter 11 of *Margin of Error* covers the method. For this chapter, one idea is enough. A variable that the model holds constant is a **control variable**, and the gap that remains is the **adjusted gap**.

Figure 8.2 shows the gap under four sets of control variables.

With no adjustment, women earned 20.6 percent less. When the model holds education and experience constant, the gap is 22.7 percent. The gap grows because the women in this file had the same average education as the men and slightly more experience. Those two variables do not account for any of the gap.

When the model also holds occupation, union membership, and region constant, the gap falls to 19.3 percent. Part of the raw gap is associated with the

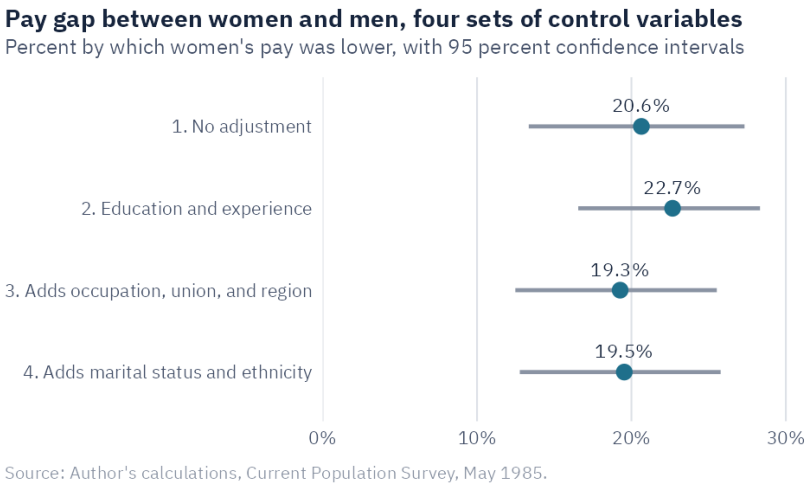


FIGURE 8.2. *Estimated pay gap between women and men in the 1985 file under four sets of control variables. The points are estimates and the lines are 95 percent confidence intervals. The estimate ranges from about 19 to about 23 percent. Source: author's calculations from the Current Population Survey, May 1985.*

kinds of jobs women and men held. Adding marital status and ethnicity changes the estimate very little.

8.5 A Control Variable Is a Choice About the Question

It can seem that more control variables always give a better answer. They give an answer to a narrower question.

Consider occupation. In the file, 76 of the 245 women worked in office jobs, and 126 of the 289 men worked in manual jobs. Holding occupation constant compares women and men within the same type of job. That comparison is useful if the question is whether an employer pays women and men differently for the same work.

Suppose, though, that women were steered away from higher-paying occupations. Then occupation is one of the routes by which sex affects pay. A model that holds occupation constant removes that route from the estimate. The adjusted gap will be smaller, and the part that was removed is a result of the process being studied.

A variable that lies on the path between a cause and an outcome is a **mediator**. Adjusting for a mediator answers the question "What remains of the gap after this path is set aside?" That can be the right question. The analyst has to choose it on purpose and say so.

For this reason, a careful report gives both figures. The unadjusted gap describes the overall difference in pay. The adjusted gap describes the difference among workers who are alike on the listed variables. A reader needs both, with the list of control variables, to understand either.

8.6 Work With the Data: Pay Gap Explorer

The last two sections showed that the adjusted gap depends on which variables are held constant. The tool lets you make that choice yourself for the 534 workers in the 1985 file.

The tool runs in a web browser at quant.marginoferrormedia.com/tools/gap-explorer.html. It needs no account and no software.

1. Read the gap with no adjustment. Women's hourly pay was lower by 20.6 percent.
2. Select education and experience. Read the new gap and its confidence interval.
3. Add occupation. The gap falls. Decide whether occupation is a fair variable to hold constant, or a route by which pay comes to differ.
4. Look at the small marks. Each is the gap under one of the 64 possible sets of control variables. Find the smallest and the largest.
5. Select all six variables and read the gap.

With all six variables held constant, the gap is 19.5 percent. Every one of the 64 choices is defensible, and they do not agree. An analyst who chooses the set after seeing the results can report almost any figure in that range. This is the reason to write the plan first.

The file is `ch08-wage-gap.csv` at quant.marginoferrormedia.com/data.

Three Checks on a Control Variable

Before you add a control variable, ask three questions. Could it affect both the group and the outcome? If so, it is a confounder, and adjusting for it is appropriate. Could the group affect it? If so, it may be a mediator, and adjusting for it changes the question. Is it measured well? A poorly measured control adjusts only partly.

8.7 Many Reasonable Analyses

The four estimates in Figure 8.2 do not exhaust the choices. An analyst could also use the mean or the median, or include or exclude part-time workers. The analyst could group occupations in a different way or drop the highest earners. Each choice is defensible. Together they produce dozens of possible analyses.

Gelman and Loken (2014) call this the **garden of forking paths**. Few analysts set out to mislead. The more common problem is that an analyst makes each choice after seeing how it affects the result. The analyst then settles on the path that gives the expected answer.

A well-known study measured how much the path matters. Silberzahn and colleagues (2018) gave one dataset and one research question to 29 teams of analysts. The question was whether soccer referees gave more red cards to players with darker skin. The teams used 21 different sets of control variables. Twenty teams found a statistically significant effect, and nine did not. The estimated effects ranged from slightly negative to nearly three times the odds.

Two practices reduce the problem.

Preregistration. The analyst records the analysis plan in a public, time-stamped registry before analyzing the data. Readers can then compare the report with the plan. Analyses that were not in the plan can still be reported, and they are labeled as exploratory.

Reporting several specifications. The analyst runs every reasonable version of the analysis and reports all of the results, as in Figure 8.2. If the estimates agree, the finding does not depend on one choice. If they disagree, the reader learns that it does.

8.8 Stating Uncertainty

The 534 workers are a sample. A different sample would give somewhat different estimates. The lines in Figure 8.2 are **confidence intervals**. Each one shows a range of values that are consistent with the data at the 95 percent level.

For the model with education and experience, the interval runs from about 17 to about 28 percent. A report should give the interval with the estimate. "Women earned about 23 percent less, with a 95 percent confidence interval of 17 to 28 percent" is a complete statement. "Women earned 22.7 percent less" suggests more precision than a sample of 534 can support.

The interval covers sampling error only. As Chapter 2 explained, it does not cover errors in how pay was reported or in who responded to the survey.

Two rules of presentation follow.

1. **Round to the precision the data support.** With an interval that is about 12 points wide, a decimal place adds nothing. Report "about 23 percent."
2. **Do not treat overlapping estimates as different.** The four intervals in Figure 8.2 overlap. The data do not show that the four specifications give different answers.

8.9 What a Report Has to Contain

A report allows a reader to judge the result and another researcher to repeat the study. The American Association for Public Opinion Research lists the items that a survey report should disclose. Table 8.1 adapts that list and links each item to a chapter of this book.

Table 8.1. *Items that a research report should disclose, adapted from the disclosure standards of the American Association for Public Opinion Research.*

Item to report	What to state	Chapter
Research question	The population, the concept, and the time and place	1
Definitions	How each construct was defined and measured	1 and 4
Design	Whether the study was a survey, an experiment, or another design	3
Question wording	The exact wording and order of the questions reported	5
Sample	The frame, the sampling method, and the sample size	6
Dates and mode	When and how the data were collected	7
Response and exclusions	The response rate, or a statement that none applies, and every exclusion rule with counts	6 and 7
Weights	Whether weights were used and what they were based on	6
Analysis	The statistic, the control variables, and whether the plan was registered	8
Uncertainty	The margin of error or confidence interval, and what it leaves out	2 and 8
Limitations	Who was left out and what the design cannot show	All
Sponsor	Who paid for the study and who conducted it	5

A **limitations** section is not an apology. It tells the reader where the result applies. For the 1985 file, three limits matter. The data are four decades old. The sample is 534 workers from one month. The file has no measure of hours worked per week, job tenure with the current employer, or the size of the firm.

The order of a report also matters. Most readers want the finding first. A common structure is a one-paragraph summary, the main findings with figures, the methods, and the limitations. Technical detail goes in an appendix. *Make It Visible*, the visualization book in this series, covers how to design the figures.

Ethics Moment: The Number That Gets Repeated

A report on a pay gap will be quoted by people on different sides of a debate. One side may quote the unadjusted gap as evidence of unequal pay for equal work, which it does not measure. The other side may quote the adjusted gap as evidence that no problem exists, which it does not show either. An analyst cannot control how a number is used. The analyst can make misuse harder. Put both figures in the same sentence. State what each one measures. Do not publish either figure without the other.

AI Reality Check: A Model Will Run Any Analysis You Ask For

A language model with access to a data file can run all four specifications in this chapter in under a minute. It can run forty more. This makes the forking paths problem worse, because each added analysis costs nothing. A model will also write a confident paragraph about whichever result you ask it to describe. Use the tool for what it does well, which is running a plan you have already written. Write the plan first. Ask the model to carry it out and to report every specification. Then check a sample of its numbers against your own calculation.

8.10 Putting It All Together

An analysis plan is written before the results are seen. It names the outcome, the comparison, the statistic, the control variables, and the cases included. Each of these choices can change the answer.

In the 1985 file, women earned between about 19 and 23 percent less than men, depending on the control variables. The unadjusted gap describes the overall difference. The adjusted gap describes the difference among workers who are alike on the listed variables. A control variable that lies on the path from cause to outcome changes the question being answered.

Many reasonable analyses of one dataset can give different results. Preregistration and reporting several specifications make the choices visible. A report states the estimate with its uncertainty, lists every item a reader needs to judge it, and describes its limits.

8.11 Looking Back and Looking Ahead

This book has followed one path. A question became a definition and then a measure. The measure was built into a tested questionnaire. A sample was drawn, data were collected and cleaned, and the analysis was planned and reported. At each step, a choice affected the final number.

Numbers answer questions of how many and how much. They are less able to answer why people act as they do or what an experience was like for them. Those questions call for a different kind of data. *Beyond the Numbers*, the companion to this book, covers research based on interviews, focus groups, and written text.

8.12 Further Reading and References

- Berndt, E. R. (1991). *The Practice of Econometrics: Classic and Contemporary.* Addison-Wesley. The source of the 534-worker file from the May 1985 Current Population Survey.
- Gelman, A., and Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*(6), 460 to 465. The source of the phrase "garden of forking paths."
- Silberzahn, R., Uhlmann, E. L., Martin, D. P., and others (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science, 1*(3), 337 to 356.
- Blau, F. D., and Kahn, L. M. (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature, 55*(3), 789 to 865. A review of research on adjusted and unadjusted gaps.
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., and Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*(11), 2600 to 2606.
- American Association for Public Opinion Research. *Code of Professional Ethics and Practices* and *Transparency Initiative Disclosure Elements.* The disclosure standards adapted in this chapter.

8.13 Key Terms

- **Adjusted gap.** A difference between groups that remains after a model holds other variables constant.
- **Analysis plan.** A statement, written before the results are seen, of how the data will be analyzed.
- **Confidence interval.** A range of values consistent with the data at a stated level of confidence.
- **Control variable.** A variable that a model holds constant.
- **Garden of forking paths.** The many reasonable analyses that one dataset allows.
- **Limitations.** A statement of who was left out and what the design cannot show.
- **Mediator.** A variable on the path between a cause and an outcome.

- **Preregistration.** Recording an analysis plan in a public registry before analyzing the data.
- **Regression model.** A method that estimates how an outcome changes with one variable while others are held constant.

8.14 Exercises

Check Your Understanding

1. List the five questions an analysis plan answers.
2. Why does the median usually describe typical pay better than the mean?
3. Explain the difference between a confounder and a mediator. Why does it matter which one a control variable is?
4. What does a 95 percent confidence interval cover, and what does it leave out?

Apply It

5. In a company, women's median pay is $58,000 and men's is $66,000. Express the gap as a difference, a ratio, and a percent difference. Show each calculation.
6. Use the file `ch08-wage-gap.csv`. Calculate the median pay of women and men who are union members and of those who are not. Describe the result in two sentences.
7. Write an analysis plan for this question. "Do renters in the county report worse health than homeowners?" Answer all five questions and name two control variables with a reason for each.

Think Deeper

8. A report states, "After controlling for occupation, hours, and industry, the pay gap is 5 percent." What question does that figure answer? What question does it leave open?
9. In the red card study, 29 teams reached different results from one dataset. Does this mean the question cannot be answered? What should a reader conclude from the full set of results?
10. Choose a published survey report. Check it against Table 8.1. Which items does it disclose, and which are missing? What does the missing information prevent you from judging?

Index of Tools and Data

Tools

Each tool appears inside its chapter, in the section called Work With the Data. This page lists all eight in one place. Each one runs in a web browser and needs no account. Nothing you type is sent anywhere or stored.

Tool	Chapter	What it does
One Night or One Year	1	Compares a count on one night with a count over one year for the same shelter.
Survey Error Tree	2	Presents short cases and asks which of the seven sources of error applies.
Random Assignment Lab	3	Shuffles the real callback totals to show how large a gap chance produces.
Scale Scorer	4	Scores the six-item food security scale and shows how alpha changes when items are removed.
Question Reviewer	5	Checks a draft survey question for common wording problems.
Sampling Explorer	6	Draws repeated samples and shows variable error and coverage bias.
Response Quality Screener	7	Applies screening rules to the simulated panel and shows what each rule removes.
Pay Gap Explorer	8	Shows the pay gap under any set of control variables you choose.

Data

Every data file is a plain CSV file. The files are at quant.marginoferrormedia.com/data.

File	Chapter	Contents	Real or simulated
ch01-pit-counts.csv	1	Point-in-time counts of homelessness, selected years	Real. HUD
ch02-unauthorized-estimates.csv	2	Seven published estimates of the unauthorized immigrant population	Real. DHS, Pew, CMS, MPI
ch03-resume-audit.csv	3	4,870 resumes from the audit study	Real. Bertrand and Mullainathan (2004)
ch04-food-security.csv	4	Household food insecurity, 2022 to 2024	Real. USDA
ch04-simulated-items.csv	4	Answers of 600 households to six items	Simulated
ch05-wording-experiments.csv	5	Results of three wording and order experiments	Real. Published studies
ch06-census-coverage.csv	6	Net coverage error in the 2010 and 2020 censuses	Real. Census Bureau
ch07-simulated-panel.csv	7	1,000 online panel responses with 120 planted cases	Simulated
ch08-wage-gap.csv	8	Pay and characteristics of 534 workers, May 1985	Real. Current Population Survey

The two simulated files were mathematically modeled to reflect patterns reported in published research. They do not describe real people. The scripts that produced them are in the book's repository.

The Series

Margin of Error covers statistics. *Make It Visible* covers charts and data storytelling. *Before the Numbers* covers quantitative research design. *Beyond the Numbers* covers qualitative research. All four are from Margin of Error Media.