CONCLUSIONS

Arguments take shape with concepts, and concepts take empirical shape through indicators. This is the succession of topics we have followed through Chapters 2 and 3. It is worth remembering that these topics are interwoven. Concepts are built for use in arguments; they don't always make sense outside of that particular context. Indicators are inextricably linked to the concept they are intended to measure. They have no intrinsic meaning.

Nonetheless, in order to understand each component of social science methodology we need to take these components apart. That is what the foregoing chapters have attempted to do. In the next chapter, we look at the task of empirical analysis, i.e., how arguments are tested.

KEY TERMS

- ●perationalization
- Resonance
- Internal coherence
- External differentiation
- Theoretical utility
- Scope-condition
- Dependent variable
- Independent variable
- Consistency
- Minimal definition
- Maximal definition
- Indicator
- Level/ladder of abstraction
- Conceptualization
- Measurement
- Scale (categorical, numeric, nominal, ordinal, interval, ratio)
- Binary
- Aggregation
- Boolean
- Necessary and sufficient conditions
- Additive
- Multiplicative
- Factor-analytic

4 Analyses

We began (in Chapter 2) with arguments and proceeded (in Chapter 3) to conceptualization and measurement. In this chapter, we turn to the problem of how to analyze an argument empirically. This may be referred to variously as *appraisal*, *assessment*, *corroboration*, *demonstration*, *empirics*, *evaluation*, *methods*, *proof*, or *testing*. Pursued in a self-conscious fashion, this stage of research involves a **research design**, i.e., an explicit method of selecting and analyzing data.

We begin by introducing a set of key terms that are necessary to understand the construction of a research design. We proceed to a discussion of the general issues that all analyses encounter. This includes precision and validity, internal and external validity, sample representativeness, sample size, probability and non-probability sampling, and missing-ness. The terms and topics introduced in this chapter will enter the narrative in later chapters repeatedly. This chapter therefore plays a foundational role in the book.

Definitions

A standard empirical analysis involves a number of components, which must be clarified before we continue. Much of this vocabulary is borrowed from survey research. Nonetheless, the concepts are helpful in all styles of research, whether quantitative or qualitative, and are illustrated in Figure 4.1.

The most basic unit in any analysis is an **observation**. ●bservations are the pieces of evidence deemed relevant for demonstrating an argument. In a standard matrix (rectangular) dataset, an observation is usually represented as a row. Each row in Figure 4.1 represents a single observation.

Each observation should record values for all relevant **variables**. In causal analysis, this includes X (the causal factor of theoretical interest) and Y (the outcome of interest), along with any other variables deemed essential for the analysis. In a rectangular dataset, variables are usually represented with vertical lines. There are three variables in Figure 4.1: X, Z, and Y.

An observation is drawn from a **unit** or **case** – bounded entities such as individuals, organizations, communities, or nation-states, which may be observed spatially and/or temporally (through time). The terms *unit* and *case* are more or less equivalent. (While a unit is bounded spatially, a case may also have implicit or explicit temporal boundaries.)

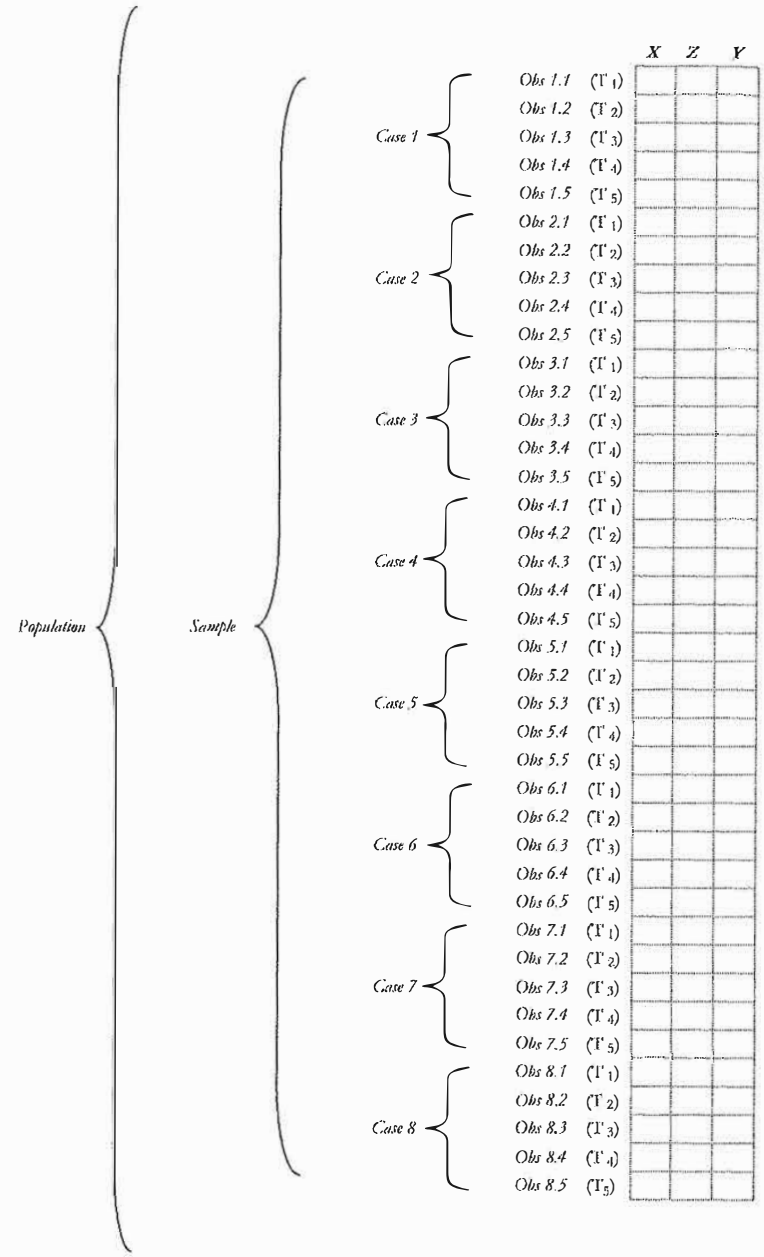

Figure 4.1 Time-series cross-section dataset
Population = Indeterminate. Cases/Units = 8. Sample/Observations (N) = 40. Cells = 120. Time-periods (T) = 5. Variables (X, Z, Y) = 3.

Collectively, the observations in a study comprise a study's **sample**. The size of a sample is the total number of observations, often denoted as "N." (N may also refer to the number of units or cases, which may be considerably less than the number of observations. This should be clear from context.)

A **population** is the universe of phenomena that a hypothesis seeks to describe or explain. It usually remains unstudied, or is studied only in a very informal manner, e.g., through the secondary literature.

The sample (the observations that are actually studied) is drawn from the population, and is usually much smaller. Hence, the notion of *sampling* from a population. Note, however, that the term sample, as used here, does not imply that the studied observations have been chosen randomly from a population. This ideal is rarely followed in practice, as discussed below.

Occasionally, a set of observations includes the entire population of interest. This is known as a **census**. A population census includes all persons residing within a country (though of course coverage is never entirely comprehensive). Likewise, a census study of nation-states would include all nation-states. Since census studies (where $N_{sample} = N_{population}$) are rare, we leave them aside in the following discussion.

These interrelated concepts are illustrated in Figure 4.1, where we can see a fairly typical time-series cross-section research design in a rectangular dataset format. Recall, observations are represented as rows, variables as columns, and cells as their intersection. Note that cells are nested within observations, observations are nested within units (aka cases), units are nested within the sample, and the sample is nested within the population.

Hypothetically, let us imagine that the population of the inference includes all worker-training programs in the United States and the sample consists of eight programs, observed annually for five years, yielding a sample of forty observations (N = 40). The **units of analysis** (the type of phenomena treated as observations in an analysis) in this hypothetical example are program-years.

All these terms are slippery insofar as they depend for their meaning on a particular proposition and a corresponding research design. Any changes in that proposition may affect the sorts of phenomena that are classified as observations and units, not to mention the composition of the sample and the population. Thus, an investigation of worker-training programs might begin by identifying *programs* as the principal unit of analysis but then shift to a lower **level of analysis** (e.g., *participants*) or a higher level of analysis (e.g., *states*) at different points in the study. Sometimes, different levels of analysis are combined in a single study. This is common in *case study* work (see Chapter 9) and is the defining feature of *hierarchical* (*multi-level*) statistical models.

Before leaving this discussion of basic terms we must address the important distinction between *quantitative* and *qualitative* analysis. This contrast is ubiquitous, and will no doubt be familiar to the reader. But it is also ambiguous since these terms can be defined in many different ways, and sometimes they are not defined at all. In this text, we adopt the following definitions.

Quantitative analysis is a formal analysis of matrix-based observations. A matrix observation is the conventional sort, represented as a row in a

rectangular dataset (illustrated in Figure 4.1). Each observation is coded along a number of dimensions understood as columns in the matrix and as variables in an analysis. All observations are regarded as examples of the same general phenomenon and are presumed to have been drawn from the same population. Each is regarded as comparable to all the others (with some degree of error) with respect to whatever analysis is undertaken. The analysis is "formal" insofar as it rests on an explicit framework of inference such as logic/set theory, Bayesian statistics, frequentist statistics, or randomization inference.

By contrast, **qualitative** analysis refers to an informal analysis of non-comparable observations. Non-comparable observations cannot be arrayed in a matrix format because they are examples of different things, drawn from different populations. The analysis is "informal" insofar as it is articulated with natural language and is unconnected to an explicit and general framework of inference. When applied in the context of causal inference this sort of evidence may be referred to as *causal-process observations*, *clues*, *colligation*, *congruence*, *genetic explanation*, *narrative analysis*, or *process tracing*.[60]

There is a strong elective affinity between quantitative analysis and large samples, as well as between qualitative analysis and small samples. One would be hard-pressed to apply informal styles of analysis to a sample of 1,000. Likewise, one would be hard-pressed to apply a formal analysis to a sample of two. The size of a sample thus influences the style of analysis. However, it does not determine it. This is apparent in the middle range. A sample of 20 may be analyzed formally or informally (or both). Thus, when we use the terms quantitative and qualitative the reader should understand that the former usually (but not always) corresponds to large samples and the latter usually (but not always) corresponds to small samples. The qual/quant distinction is not solely a matter of *N*.

So defined, there is no epistemological gulf separating quantitative and qualitative analysis. Indeed, any qualitative analysis can be quantified – with the cost of reducing complex, multifaceted data into matrix observations. Many quantitative datasets began life, one might say, as a series of qualitative observations. These were then coded in a systematic fashion to generate a set of observations that could be arrayed in a dataset. Often, this is useful. But not always; nor is it always possible. That is why qualitative analysis continues to play an important role in social science research – and especially in case study research, as discussed in Chapter 9.

That said, quantitative analysis is dominant in many fields and also has a more developed methodology. Consequently, when we use the term "observation" or "analysis" in this book the reader can assume that we are talking about the quantitative variety, unless stated otherwise.[61]

Precision and Validity

Social science analyses strive for *precision* (aka *reliability*) and *validity* (aka the absence of *bias*). **Precision** refers to the closeness of repeated estimates of the

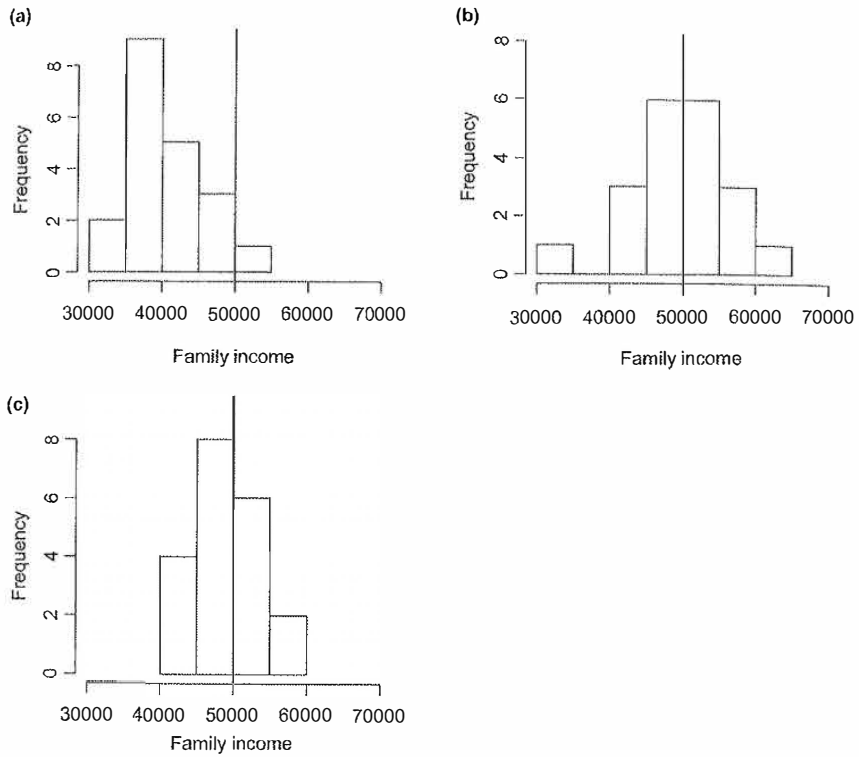

Figure 4.2 Precision and validity
Methods of sampling: (a) hardline telephones, (b) hardline phones and cell phones, (c) door-to-door canvassing. Vertical line: true value, according to US census.

phenomenon of interest when using the same measurement instrument or causal model. **Validity** refers to the closeness of an estimate to the true (often unknown) value. To explore these concepts we employ a hypothetical example.

In recent years, survey researchers have employed a number of techniques to obtain representative samples of the general public. These include (a) hardline telephones, (b) hardline phones and cell phones, and (c) door-to-door canvassing.

In order to evaluate their precision and validity, we shall imagine employing each of these recruitment techniques to conduct 20 surveys of the general public in the United States. Each survey includes 2,000 participants, who are chosen in whatever fashion the pollster believes will result in the most representative sample. There is only one question on the survey: What is your family income? For each sample, we calculate the **median** (that value for which there are equal numbers of values above and below). These results are plotted on the graphs in Figure 4.2.

We shall assume that census results represent the true value of household income in the country. In 2006, the United States Census Bureau reported that

the median annual household income was roughly $50,000. This value is represented by a vertical line in Figure 4.2, and is the quantity of theoretical interest.

Results for the first recruitment technique – hardline telephones – are presented in panel (a) of Figure 4.2. They suggest that this method is fairly precise, as the estimates cluster tightly around the sample median. However, the sample median falls far from the true value (the median value for the population, as revealed by census data), suggesting that the method is not valid.

Results for the second recruitment technique – hardline phones and cell phones – are presented in panel (b). They suggest that the method is imprecise, as estimates vary widely around the sample median. However, the sample median falls close to the true value, so the technique may be regarded as valid (in repeated sampling).

Results for the third recruitment technique – door-to-door canvassing – presented in panel (c), is both precise (tightly clustered) and valid (close to the population value of interest). Thus, on the basis of this set of tests, door-to-door canvassing is superior to the other sampling techniques. Of course, these results are entirely hypothetical. You should also note that many other factors – including cost – may figure into a pollster's decision to employ a method of sampling. Nonetheless, the illustration is effective in demonstrating a crucial distinction between precision and validity.

Let us now elaborate a bit on these concepts and their employment in social science research.

Precision, or reliability, refers to level of stochastic (random) error, or **noise**, encountered in measurement or some other feature of estimation. Precision in measurement is often assessable through **reliability tests**, where the same technique of data gathering is employed multiple times in order to ascertain its reliability. Reliability tests might focus on a survey technique, as in our example, on experts who code data from a primary or secondary source, or on any aspect of data collection. If multiple applications of the measurement instrument reveal a high level of consistency one may regard the chosen instrument or model as reliable (precise). This is typically calculated as the inverse of the **variance** (i.e., dispersion around the mean). Greater variance means less reliability. The same logic applies to precision in causal inference, i.e., when one is comparing models employed to assess a causal relationship between X and Y.

If the opportunity to test multiple iterations of an indicator or model is not present then the issue of reliability remains at the level of an assumption. But it is nonetheless crucial. A high probability of random error may doom even the simplest generalization about the world.

Validity, by contrast, refers to *systematic* measurement error, error that – by definition – introduces bias into the resulting analysis. One often finds, for example, that the level of accuracy with which an indicator is measured varies directly with some factor of theoretical interest. For example, in constructing international indicators for human development (e.g., life expectancy, literacy) and economic performance (e.g., inflation, GDP) we rely on surveys conducted in countries throughout the world. Sometimes, these surveys present a more

favorable picture of the quality of life or the strength of the economy than is warranted. And there is some reason to imagine that autocratic governments engage in this practice to a greater extent than democratic governments. If so, these indicators suffer from systematic bias. However, because we cannot be sure of this bias, and have no estimate of its extent, we have no easy means to correct it.

While we can usually assess reliability we rarely have a fix on the problem of validity. Note that our hypothetical example is highly unusual in one important respect: we know the true value of the measure of interest – $50,000. In the world of social science it is rare to possess a definitive measure of anything of great substantive importance. Indeed, our hypothetical example might be challenged on the ground that population censuses are never entirely comprehensive; some citizens escape the prying eyes of government surveyors. In the 2000 US Census, for example, despite elaborate advertising and outreach, the overall mail response rate was only 64%.[62] Worryingly, these non-respondents may be quite different than respondents, leading to systematic bias, a problem the Census Bureau is aware of and attempts to evaluate.[63] For example, homeless people are less likely to be contacted in a census and since homeless people have much lower family income than people with stable addresses it is reasonable to suppose that all census-based data (unless adjusted to correct for this deficiency) is biased to some extent. Various sociodemographic groups and geographic areas have been shown to be underrepresented, especially Blacks and Hispanics.[64]

All of this is to say that validity, unlike reliability, is very difficult to test in a definitive fashion. Even our toy example is open to dispute. And with more complex concepts such as democracy or GDP the points of potential dispute are multiplied. For causal inferences, which build on descriptive inferences, the problem is magnified even further.

Internal and External Validity

Typically, researchers examine only a small number of instances of a phenomenon of interest. If one is trying to ascertain the median income within a large population, as in the example explored above, one might sample only 2,000 respondents. On this basis one would hope to infer median income across 320 million people. Even more striking is the attempt to learn something about human nature from experiments conducted on a sample of college students drawn from a single classroom, a common practice in psychology. Here, a sample of several hundred may be the basis for generalizations about the entire human race.

In either case, social science must grapple with a crucial question: how to relate findings drawn from a sample to a larger population of interest.

This problem engages researchers in a two-level game. The first part of the game concerns reaching conclusions about the sample. The second part of the game is about extrapolating those conclusions to a larger population, sometimes referred to as a problem of **inference**.

To distinguish these two spheres of truth social scientists invoke a distinction between **internal** and **external validity**. The first refers to the validity of a hypothesis with respect to the studied sample. In our previous example, a problem of internal validity may arise if respondents lie about their family income, perhaps in response to perceived norms. If so, a calculation of mean family income for a single sample – or a group of samples – may be too high or too low.

The second issue arises when we try to extend sample-based results to a larger population. Naturally, if there are problems of internal validity there are likely to be problems of external validity. But even if our analyses of the sample are correct they may not be correct across a larger population. This is the problem of external validity, toward which the rest of the chapter is directed.

With respect to external validity, two characteristics of a sample are especially relevant: *sample representativeness* and *sample size*.

Sample Representativeness

The external validity of a study is grounded largely on the **representativeness** of a chosen sample. Is the sample similar to the population with respect to the hypothesis that is being tested? Are we entitled to generalize from a given sample to a larger universe of cases?

In the case of research into income one must consider whether the sample exhibits the same income distribution as the general population. In the case of research into cognitive properties of human nature one must consider whether college students are similar in these respects to other humans (and how far back in time a result might be generalizable). And with respect to studies of worker-training programs one must consider whether the chosen program sites are representative of a larger population of programs that one wishes to understand.

Note that questions about representativeness are also questions about how to *define* the population. Consider the study of worker-training programs that focuses on programs in the state of New York. It could be that results from this study are generalizable (a) to that state (only), (b) to the United States, (c) to advanced industrial societies, or (d) to all societies in the contemporary era. Likewise, it could be that the results are generalizable (a) to native-born unemployed persons between the ages of 20 and 50 without disabilities, (b) to unemployed people between the ages of 20 and 50 without disabilities, (c) to people between the ages of 20 and 50 without disabilities, (d) to people without disabilities, or (e) to all people.

The point is that any consideration of external validity forces one to be very specific about the population and the scope-conditions of an inference. What is the population, exactly? What is it, exactly, that is generalizable to that population?

Unfortunately, these questions are often difficult to answer in a definitive fashion, for reasons already discussed (see *Boundedness* in Chapter 2). However, they must be addressed, even if only in a speculative fashion.

Sample Size (*N*)

The second characteristic that impacts a study's external validity is the size of the sample upon which the study is based. More observations are better than fewer because they provide a more precise estimate of the quantity of interest.

Suppose one is trying to figure out the effect of a worker-training program on employment prospects or earnings but one has available information for only one program. Under the circumstances, it will probably be difficult to reach any firm conclusions about the matter. Of course, one case is a lot better than none. Indeed, it is a quantum leap. Yet, empirical research with only one case is also highly indeterminate, and apt to be consistent with a wide variety of competing hypotheses. Conclusions about a broader population are hazardous when one considers the many opportunities for error and the highly stochastic nature of most social phenomena.

One sort of problem stems from problems of sampling error encountered when the sample is very small. Note in order to make accurate inferences about a larger population one must have a sample that is similar to that population in relevant respects. The chances of finding such a sample when the sample is small are considerably reduced. One's chances of achieving a representative sample increase with sample size – if the sample is chosen randomly from the population. We provide a more detailed look at the importance of sample size in probability sampling in Chapter 21.

Uncertainty associated with sampling variability is captured in a statistic known as a **confidence interval**. A confidence interval indicates the bounds within which a true value is likely to fall at a chosen level of probability. A 95% confidence interval means that our confidence interval captures the true value 95% of the time. A 90% confidence interval means that our confidence interval captures the true value 90% of the time. We describe how to calculate confidence intervals and provide a more precise treatment of their interpretation in Chapter 21.

A larger sample is advisable if everything else is equal. Of course, this is sometimes not the case. For example, sometimes increasing the size of a sample decreases its representativeness. Consider a sample that is representative. Now add cases non-randomly. Chances are, the larger sample is less representative than the smaller sample. Likewise, sometimes one can gain greater leverage on a question with a carefully chosen small sample than with a large sample; this is the justification for purposive case selection procedures, discussed in the context of case study research (Chapter 9). In particular, an empirical study whose sole purpose is to disprove an invariant causal or descriptive law can achieve this purpose with a single observation – so long as it contradicts the hypothesis.

The point remains, obtaining a large sample is a noble objective so long as it doesn't interfere with other goals. When it does, the researcher needs to decide which goals to prioritize – or, alternatively, adopt a multi-method research design that incorporates large- and small-*N* components.

How large is large enough? What is an appropriate size for a sample? There is no easy answer to this question. It is not the case that sample size should be proportional to the size of the population. Sampling error rests primarily on the size of the sample not the sample/population ratio. Following precedent, i.e., what other scholars have done, may be appropriate – but only if the goals of your analysis are similar to theirs. Occasionally, one may calculate an appropriate sample size by stipulating the goals of the analysis and an acceptable confidence interval for the variable of theoretical interest. However, this only works if there is a single hypothesis (not a multitude of hypotheses) and if it is possible to identify a benchmark confidence interval (which is not always possible).

Requisite sample size depends, in general, on the relative strength of the "signal" (the variable of theoretical interest) relative to background "noise" (factors that might muffle the signal). Let us say that instead of seeking to estimate the height of individuals within a population we were interested in estimating the impact of a day-long worker-training program on subsequent earnings over the succeeding two years. As one might imagine, the impact of a single day's training on subsequent earnings is likely to be fairly minimal, and many factors other than training affect earnings. In this setting, one would presume that the ratio of signal to noise is pretty low, requiring a great many observations in order to discern a causal effect (if indeed there is one).

Probability Sampling

The preferred approach to sampling is to choose cases *randomly* from the population. Because cases are chosen randomly, one knows the probability that any given case will be chosen as a member of the sample. This approach to sampling is therefore referred to as **probability sampling**.

In **simple random sampling**, each case within the population has an equal chance of being selected for the sample. (This is sometimes referred to as an *equal probability sample*.) The mechanism of selection might be drawing balls from an urn, as in raffles. More commonly, random selection is achieved with a random-number generator from a computer program. (These may be found online or as part of a software package.) The statistics we introduce in subsequent discussion assume random sampling.

In **systematic sampling**, members of a population are chosen at equal intervals, e.g., every 10th or every 1,000th. This assures equal probability sampling *if* the chosen interval is not associated with any particular feature of the population, a matter that may be difficult to discern.

In **cluster sampling**, members of a population are divided into clusters (groups), clusters are chosen randomly (using some random-selection mechanism), and then each member of the cluster is automatically included in the sample. This approach is usually taken for logistical reasons, i.e., when it is easy to include all members of a naturally occurring cluster such as a school, neighborhood, census tract, or family.

In **stratified sampling**, each member of the population is assigned to a stratum and cases are chosen randomly from within each stratum. If the number chosen from each stratum does not reflect the proportional size of that stratum within the population, cases will need to be re-weighted so that the resulting sample is representative of the population. For example, in a sample of 2,000 individuals drawn from the United States it may be important for theoretical reasons to identify strata composed of various minority groups. While some minorities like Hispanics and African-Americans are quite large, others such as Jewish-Americans are quite small (roughly 3% of the general population). A stratum composed of 3% of 2,000 yields a sub-sample of only 60 individuals – too small to allow for precise estimates of the actual population of Jews in the United States. Under the circumstances, it probably makes sense to *over-sample* among Jews, granting Jews a greater probability of entering the sample than members of other social groups. Let us say that the researcher decides to select twice as many Jews as their share of the population would allow, raising their share of the sample to 6%. Representativeness can then be restored to the sample by down-weighting Jews in the sample – in this case, granting half the weight to Jewish respondents as to other respondents. This approach is generally cheaper to implement than the alternative approach – doubling the size of the entire sample (e.g., from 2,000 to 4,000) – and achieves the same results. Naturally, it requires that one identify those strata which are of theoretical interest prior to drawing the sample.

Various approaches – simple, systematic, cluster, stratified – may be combined in *multi-stage sampling*. For example, one might begin with clusters, stratify within clusters, and then sample randomly within each cluster/stratum. The key point is that whatever system of probability sampling is employed, disproportionalities should be corrected (by weighting) so that the sample is representative of the population of interest.

A key advantage of probability sampling is that one can estimate sampling variability (from sample to sample), thus providing estimates of precision to accompany whatever inferences one wishes to draw. It is not enough to say that a sample is "large" and therefore "precise." One wishes, as well, to know *how* precise a given sample estimate is, that is, how close it is likely to be to the population parameter.

Non-Probability Sampling

A very different approach to sampling is to select cases *non-randomly* from a population. A small number of cases may be selected with specific purposes in mind, as in case study designs (see Chapter 9). Cases may be chosen in a **snowball** fashion. This approach is common in interview-based research, when one relies on each respondent to suggest other possible respondents – creating a snowball effect (the ball of respondents gets larger as each new snowflake joins the ball). Cases may also be chosen for logistical reasons, e.g., because they are accessible, cheap, or for some other reason easy to study. This is sometimes known as **convenience sampling**.

In all of these approaches the researcher has no way of assessing the probability, for each case in the population, of being selected into the sample. Accordingly, these approaches are sometimes referred to as *non-probability samples*. Such approaches produce samples of uncertain representativeness. We don't know how similar they are to the population of interest; likewise, we may be uncertain about what that population is. If the sample is small, as it is with case study designs, then the study also faces a problem of reliability (imprecision).

From the perspective of external validity there is little to be said in favor of non-probability sampling. However, external validity is not the only goal of social science research, and other goals sometimes require non-probability approaches to sampling.

At an early stage of investigation, when not much is known about the phenomenon of interest and before one has identified a specific hypothesis, it is common to focus on a small number of cases so that one may observe those cases in an intensive fashion. Likewise, if one already has a clear sense of a relationship but one does not know why it obtains, one might prefer to focus on a small number of cases, intensively observed. These are classic justifications for a case study research design.

While it may be feasible to select a single case, or several cases, randomly from a population, our earlier discussion – of sample size – shows how unreliable such tiny samples can be. For this reason, one is generally advised not to use a probability-based method for choosing a very small sample. The exception would be a situation in which cases found within the same stratum are all equally informative. Here, one may elect to choose cases within that stratum randomly. While this introduces an element of randomness, it applies only to the chosen stratum or strata. Presumably, not all strata would be included so that the resulting sample remains unrepresentative of the population.

Even where studies incorporate a large sample it still may be undesirable, or impossible, to implement probability-based sampling procedures. Worker-training programs cannot draw randomly from the universe of unemployed people because many people will refuse to participate. As such, the sample of subjects analyzed in such a study are not likely to be representative of the larger population of unemployed – though they might be considered representative of a smaller population vaguely defined as "those who are willing to participate in a worker-training program."

Wherever random sampling techniques are inapplicable researchers must struggle to define the representativeness of a sample, and hence the plausible generalizability of results based on that sample. This is true regardless of whether the sample is very small (i.e., a case study format) or very large.

Missing-ness

In the discussion so far we have assumed that all cases in the population can be accessed through probability sampling procedures and that the chosen cases can be included in the sample, i.e., they can be studied. Unfortunately, this does not always hold.

There may be slippage between the population and the **sampling frame**, those members of the population who are accessible to the probability sampling procedure. For example, if a survey is conducted by telephone the only persons who can become members of the sample are those who have telephones. Thus, if one has access to all telephones – through random-digit dialing – one can obtain a representative sample of telephone owners.

Another source of bias occurs when a chosen case cannot be studied, or can be studied only partially. This might be because a respondent refuses to participate in a study (*non-response*). It might be because the respondent completes only part of a survey or does not adhere to the protocol of an experiment (*non-compliance*). It might be because a chosen case is especially sensitive, for political or ethical reasons, and therefore cannot be included in the sample. In a historical study, it might be that a chosen case does not offer the archival records that would be required in order to conduct the study. Lots of things may intervene to thwart the goals of a sampling procedure.

We shall refer to these problems generically as **missing-ness**, that is *missing data*. What is meant by missing data is that a sample lacks observations for some units that should (by some probability-based principle of selection) be included.

If the pattern of missing data is random it causes little harm. Suppose that those who own phones and agree to conduct a survey are no different (in relevant respects) to those who do not. The survey researcher need only increase the number of calls in order to obtain the desired sample size, which will in any case be representative.

If, however, the pattern of missing-ness is systematic then the sample will be biased. For example, if telephone owners are different from those who don't own telephones in ways that are relevant to the analysis, estimates will be biased away from the (true) population parameters. (A good deal of research has gone into this question, with inconclusive results.)

A potential solution is to fill in missing data, creating a full sample that is larger and – one hopes – more representative than the truncated sample. If one has a good guess about the nature of the missing data one may develop a simple decision rule for filling in missing observations. For example, if one knows (from other sources) the mean income of persons without telephones, and their share of the general population, one might use this value for all such phantom respondents, thus rectifying the non-representativeness of the sampling frame.

Another approach is to employ a statistical model (an algorithm) to estimate missing values based on patterns in the data that have been gathered. This requires knowing something about the cases that have missing values. Let us say that we know their telephone prefix, and that we can safely assume people with the same prefixes share certain characteristics (because they live in the same area or were assigned their cell phone number at the same time). On this basis, we might estimate the income of these non-respondents based on information that we have already collected from respondents with the same prefix who answered their phones and completed the survey.[65]

CONCLUSIONS

In this chapter, we have introduced the core elements of empirical analysis in social science, applicable to both descriptive and causal analyses. After reviewing key terms, we discussed the twin goals of precision and validity. Next, we distinguished between internal and external validity. The final sections of the chapter dealt with the attempt to achieve external validity. This quest involves sample representativeness and sample size. Specific techniques include probability sampling, missing-ness, and **non-probability sampling**.

By way of conclusion, it is worth pointing out the obvious: all empirical knowledge is to some extent uncertain. This stems, arguably, from stochastic features of the world at a subatomic level – a matter debated by scholars of philosophy of science and physics. It stems, in any case, from our inability to attain perfect understanding of the complex world around us. Consequently, there is always a degree of uncertainty about any statement we might make about the world, even if the degree of uncertainty is judged to be quite small. Note that as the topic increases in importance the level of uncertainty usually rises in tandem. We can make mundane statements about the world with a high level of certainty, but we cannot pronounce upon the causes of democratization, or the causes of war and peace, with such assurance. It follows that the most relevant work in social science – in the sense of addressing issues that ordinary people care deeply about – is often accompanied by a high degree of uncertainty.

One of the features that distinguish science from other modes of apprehending the world – for which "journalism" is our convenient foil – is the attempt to represent uncertainty in a forthright and realistic manner (rather than sweeping it under the rug, so to speak). To that end, we must address a common misunderstanding.

Measures of precision, as discussed in this chapter, usually encompass only one source of uncertainty – that associated with sampling from a population. Other sources of uncertainty such as that associated with measurement (discussed in Chapter 3) or causal inference (discussed in Part II) are not generally incorporated into sampling-based statistics. Thus, when you encounter terms that purport to measure uncertainty – e.g., *variance*, *standard error*, *t statistic*, *confidence interval* – it is important to bear in mind that these statistics are probably taking account of only one threat to inference. Other threats to inference, although more common and more problematic, are harder to measure in an objective fashion, and hence go unreported or are dealt with in prose. This is another reminder that methodological adequacy is often not summarizable in handy statistical formats. You have to slog through the details of a research design to see its strengths and weaknesses.

KEY TERMS

- Research design
- Observation

- Variable
- Unit/case
- Sample
- Population
- Census
- Unit of analysis
- Level of analysis
- Quantitative
- Qualitative
- Precision
- Validity
- Median
- Noise
- Reliability tests
- Variance
- Inference
- Internal/external validity
- Representativeness
- Sample size
- Confidence interval
- Probability sampling
- Simple random sampling
- Systematic sampling
- Cluster sampling
- Stratified sampling
- Snowball sampling
- Convenience sampling
- Sampling frame
- Missing-ness
- Non-probability sampling