have no need for statistics at this early stage of their social science education; and thus may set aside Part IV for later. For those who read on, however, what follows is intended to serve as a step-by-step introduction to the most commonly applied statistics in the social sciences, a mere jumping-off point for further study and other works. Ultimately, understanding the techniques and topics introduced below is a prerequisite to successfully designing quantitative research in the social sciences. Thus, everybody from undergraduates taking advanced level courses that assign academic articles to graduate students desiring a refresher before engaging a dedicated applied statistics course will benefit from this quick and mostly painless introduction.

In what follows we divide the subject into seven categories: data management (Chapter 17), univariate statistics (Chapter 18), probabilities (Chapter 19), statistical inference (Chapter 20), bivariate statistics (Chapter 21), regression (Chapter 22), and causal inference (Chapter 23). We begin by noting a simple distinction between statistics and statistical inference. Statistics allow us to parsimoniously describe a host of information or data. Statistical inference is the process of inferring from a sample to a population, which has arisen out of the need to address limitations in data collection and research design, in particular time and cost. In the sections that follow we introduce the most common statistical inference tests with applications to the social sciences. Before doing so, we turn to a discussion of data management, as well as the basic statistics, concepts, and properties necessary to understand the more sophisticated analyses and hypothesis tests that follow. That is, the chapters progress by building upon each other; introducing the foundational technique necessary to conduct the more complex analyses in the later chapters. Indeed most of the statistics introduced in this book are built on a few standard mathematical tools, especially indicators of central tendency and variance. Once you have mastered these basic tools, other more complex procedures will be fairly simple to conduct and easier to understand.

17 Data Management

The first step toward analysis is to get your data into a format that is convenient for the sort of analysis you wish to pursue. This, in turn, probably depends on the sort of data you are collecting. Here, we shall distinguish among four data types: **qualitative**, **medium-*N***, **large-*N***, and **textual** (though one can have data that is a combination of these; e.g., medium-*N* and qualitative, large-*N* and textual ... etc.). Methods for handling these data types are continually invented and reinvented, so the reader may wish to consult other sources to obtain the most up-to-date information on these subjects. Our intent is to provide a useful overview, in any case, not to delve into the details.

Qualitative Data

Suppose that you are trying to integrate data drawn from a limited number of units with a diversity of evidence. The evidence may have been gathered with any of the techniques (or combination of techniques) discussed in Chapter 13. It might include text, photos, maps, and other media. The nature of the material might be variegated – a combination of what informants said and did, the researcher's own observations and theories, multi-media artifacts, locations, relevant articles from academic journals and/or popular media, feedback from colleagues, and so forth. Some of the evidence may apply across all studied units in the sample while some is specific to certain units. But one doesn't have systematic observations for a limited set of variables across all units in the sample. It may not even be clear what the variables are, what the sample is, or what the population of the study is. There may be – at least initially – no specific hypothesis but rather a general research question that awaits further refinement. In other words, the investigation may be more exploratory (to discover a theory or hypothesis) than confirmatory (to test a theory or hypothesis).

This setting exemplifies a good deal of work often described as qualitative, so we shall refer to it as **qualitative data** (as defined in Chapter 4). Because of its unstructured nature, data of this sort presents the researcher with a problem: how to collect and organize all of the material in a fruitful way, a way that allows the researcher to think through possible connections – to theorize – and also to enlist supporting evidence for the construction of a systematic argument when the process of writing has begun.

For this purpose, a number of qualitative data analysis (QDA) programs have been developed. These include ATLAS-ti, MAXqda2, and NVivo (the successor to NUD*IST). A common feature of QDA programs is the ability to cross-reference entries so that data can be assembled and reassembled in many different ways, e.g., by time, by location, by informant, by informant's social group, or by some designated theme (which must be coded by the researcher). Programs are also fully searchable, allowing the user to pull text (or some audio-visual output) from one setting to another, e.g., from the database into the text of a paper.

Like all software programs, they require some investment of time on the part of the user before they can provide efficiency gains. However, if you plan to collect a great deal of qualitative data and you have no easy way of organizing it (using simple text files), you might consider making the investment.

Sometimes, initial work with qualitative data leads to a reduction of that data – perhaps by successive recoding – into a single table or a standard dataset, as discussed in later sections of this chapter. Qualitative data can often be reconfigured as quantitative data. Of course, there is always some loss of information in any data reduction process. However, in some circumstances the advantages – the opportunity to systematically measure and test a relationship across a large number of units – outweigh the disadvantages. In any case, such a reconfiguration does not mean that the entire project shifts from a "qualitative" mode to a "quantitative" mode. The shift may be applicable to one portion of a project, as happens typically in a multi-method research design (see Chapter 10).

Medium-*N* Data

Suppose one is dealing with a small or medium number of cases and information for those cases that can be represented as variables (dimensions that are equivalent, and thus potentially measurable, for each case in the sample). One may not have sufficient information to fully code all the variables, or one can do so only preliminarily and there are question marks. The data may not be entirely numeric; it may take the form of qualitative judgments – strong/weak, present/largely present/largely absent/entirely absent, and so forth. Moreover, the precise boundaries of the population (and hence of an appropriate sample) may be open to question. In this setting, a reasonable approach is to try to represent the data one has collected in a single, unified table – sometimes called a **truth-table**.

An example, comprising 20 cases and three variables, is provided in Table 17.1. Note that this table incorporates various types of data – binary (*X*), textual (*Z*), and interval (*Y*). For some purposes, it may make sense to recode all variables in a binary fashion, as demonstrated in Table 17.2. And for other purposes, it may make sense to reduce this information so that cases with similar or identical codings are listed together, as part of the same "primitive" case type, as shown in Table 17.3. In this fashion we are able to collapse 20 cases into eight rows, making potential relationships easier to visualize. Because we keep track of the number of cases (*N*) falling into each primitive type, no information is lost.

Table 17.1 Truth-table with "raw" coding

	Attributes		
Cases	*X*	*Z*	*Y*
1	1	Small-medium	0
2	0	Large	5
3	1	Large	18
4	0	Small	3
5	1	Medium-large	44
6	1	Large	4
7	0	Small	6
8	1	Large	77
9	1	Large	98
10	0	Small-medium	46
11	1	Medium-large	33
12	0	Large	46
13	1	Small	68
14	0	Large	12
15	1	Medium-large	25
16	1	Small	37
17	0	Small	52
18	1	Small	51
19	1	Small-medium	11
20	0	Large	2

There are many uses for information presented in a tabular format. One use of such a table is to reveal possible causal relationships. For example, one might regard *Y* as the outcome and *X* and *Z* as possible causes – either independently or in combination. Truth-tables are especially useful for highlighting set-theoretic relationships, where a single variable or combination of variables has a necessary or sufficient relationship to an outcome. This is the logic of **qualitative comparative analysis** (QCA), a rather complex algorithm (actually set of related algorithms) for analyzing set-theoretic relationships in small- to medium-sized samples. Sometimes, a tabular format leads eventually to a standard dataset, as discussed in the next section.

Large-*N* Data

For evidence that can be represented as numeric data across a large number of cases a standard dataset matrix is appropriate. Here, each observation occupies a separate row (similar rows are not combined into "primitives"). There may also be short text ("string") variables – e.g., proper nouns representing persons or places

Table 17.2 Truth-table with binary coding

	Attributes		
Cases	*X*	*Z*	*Y*
1	1	0	0
2	0	1	0
3	1	1	0
4	0	0	0
5	1	1	0
6	1	1	0
7	0	0	0
8	1	1	1
9	1	1	1
10	0	0	0
11	1	1	0
12	0	1	0
13	1	0	1
14	0	1	0
15	1	1	0
16	1	0	0
17	0	0	1
18	1	0	1
19	1	0	0
20	0	1	0

Table 17.3 Reduced truth-table with primitive case types

		Attributes		
Case types	*N*	*X*	*Z*	*Y*
A	2	1	1	1
B	5	1	1	0
C	2	1	0	1
D	3	1	0	0
E	1	0	1	1
F	3	0	1	0
G	1	0	0	1
H	3	0	0	0
	20			

or events stored in the dataset. However, each string variable must also be represented by an accompanying numeric variable if it is to play a part in the resulting analysis.

This sort of data can generally be stored in a rectangular dataset – a two-dimensional matrix – that can be read with various software, such as R, Stata, Microsoft Excel, or simple text editors. Excel is user-friendly and able to store both numeric and string data and calculate univariate (descriptive) statistics (see Chapter 18). It is less useful for the more sophisticated statistical analyses addressed in the subsequent chapters. For this you will need a statistical package such as SPSS, Stata, SAS, or R. If you are dealing primarily with numeric data and you know you will be doing some statistical analysis (beyond univariate analysis) you may choose to enter your data directly into the statistical package. You may also start with Excel and convert at a later time (once data collection is complete). If the data is more complex, involving relationships among more than two dimensions, relational databases – for example, those based on the SQL programming language – may be required. However, this is beyond the level of complexity most researchers face in their work.

A simple two-dimensional matrix format is illustrated in Table 17.4. In the first row are the variable names. These may have to be shortened to suit the restrictions of a software program, and may need to be represented as a single word, e.g., *Country_code* rather than *Country code*. For each string variable (Country and City) there is an accompanying numeric variable (Country code and City code). The units are nested within each other – city within country.

Next, we find a time variable – measuring the year to which the data applies. Time variables might also include months, days, hours, seconds, and so forth. It depends of course upon the temporal units in which the data is collected.

If all the data refers to the same time-period, then the analysis must be **cross-sectional**, and no time variable is needed (it would have the same value for all observations). Things are slightly more complicated with **lagged** variables. In the eighth column you will see that Income/capita is lagged by one period. That is, the value entered in the first row – corresponding with the year 2005 – is actually the value for the previous year (2004). You can create complex time-dependent relationships in time-series and cross-sectional datasets simply by lagging variables at different intervals. (Sometimes, the data software will do this for you.) Note that a variable can be forward-lagged or backward-lagged.

After the time variable (Year), Table 17.4 contains a series of variables that represent the factors of theoretical interest – in this case, *Population*, *Income/capita*, and *Area*. This dataset will allow you to examine relationships among these three variables through time (at annual intervals) and across two levels (city and country).

The final column is labeled *Notes*, and is intended to store information pertaining to each observation. This might be about sources for that observation, special problems of interpretation or reliability, or whatever details one may wish to keep track of. Unfortunately, many statistical packages limit the number of characters

Table 17.4 Large-*N* dataset structure

1	2	3	4	5	6	7	8	9	10
Country	Country code	City	City code	Year	Population	Income / Cap	Income / Cap$_{t-1}$	Area	Notes
Austria	01	Vienna	01	2005	1632569	37900	37700	8428.1	Eurostat
Austria	01	Vienna	01	2006	1652449	39500	37900	8428.1	
Austria	01	Vienna	01	2007	1661246	40600	39500	8428.1	
Austria	01	Graz	02	2005	241298	33400	32800	3414.1	
Austria	01	Graz	02	2006	244997	34800	33400	3414.1	
Austria	01	Graz	02	2007	247624	35500	34800	3414.1	
Belgium	02	Antwerp	01	2005	457749	34900	33300	954	
Belgium	02	Antwerp	01	2006	461496	35700	34900	954	
Belgium	02	Antwerp	01	2007	466203	36800	35700	954	
Belgium	02	Ghent	02	2005	230951	30500	30200	1266	
Belgium	02	Ghent	02	2006	233120	31600	30500	1266	
Belgium	02	Ghent	02	2007	235143	33200	31600	1266	
Bulgaria	03	Sofia	01	2005	1148429	15200	13900	10679.5	Area 2010
Bulgaria	03	Sofia	01	2006	1154010	17900	15200	10679.5	Area 2010
Bulgaria	03	Sofia	01	2007	1156796	21200	17900	10679.5	Area 2010
Bulgaria	03	Plovdiv	02	2005	341873	6600	6200	5802.5	Area 2010
Bulgaria	03	Plovdiv	02	2006	343662	7100	6600	5802.5	Area 2010
Bulgaria	03	Plovdiv	02	2007	345249	7400	7100	5802.5	Area 2010

in a string variable, and thus limit the sort of notes you can take about an individual observation. Excel and Stata are more permissive. For example, if you wish to insert a note about a specific data *cell*, Excel will allow this but other statistical packages generally will not.[177]

Any piece of clarifying information that can't be inserted into a database will need to be kept somewhere – traditionally in a text file called a **codebook**, which explains what each variable means and the sources from which it is gathered. Eventually, one hopes that statistical packages will become more accommodating of meta-data, descriptions of data including its provenance, as it is quite complicated to have data in one location and explanations of that data in another.

Excel is a superb tool for data storage and management, e.g., moving variables and observations around to different locations within a spreadsheet or in related sheets, or creating new variables or observations based on those you have. However, you should be aware that this ease of manipulation also introduces a risk. Specifically, it is easy to uncouple the variables that describe a specific observation. For example, if you block all the data from column 4, and paste it down one row, it will be out of sync with the rest of the dataset. A chance error like this will likely destroy all the (real) relationships in your dataset; in their place you will find spurious (illusory) relationships. By contrast, most other statistical packages make it harder to separate variables connected with a single observation, which makes them harder to manipulate but also erects a barrier to data management errors of this sort.

Textual Data

Suppose the data of theoretical interest is textual in nature. That is, you wish to analyze a number of texts, e.g., articles drawn from newspapers, high school textbooks, novels, websites, political speeches, party platforms, constitutions, transcripts from hearings or meetings, and so forth.

If there is a modest number of texts, or just a few key texts, you may enlist the traditional approach of reading, marking up texts, and taking notes in a separate text file. The oldest continuous tradition of textual analysis, biblical exegesis, rests on the close analysis of key texts – in this case, religious texts. The tradition of in-depth analysis of key texts lives on in history and other humanities disciplines and, more selectively, in the social sciences. A close reading of texts is clearly justified in the study of constitutions or founding documents, key court decisions, key legislation, influential speeches, or influential theorists.

Suppose, however, that the number of texts that you wish to analyze is very large, e.g., thousands of speeches, newspaper articles, tweets, blogs, books, or other texts. In order to reduce this plenitude of information so that it can reveal a coherent story or pattern you will need a mechanized system of storage and retrieval, known generically as **content analysis**, or **textual analysis**. Sometimes, distinctions are drawn among these terms. However, for our purposes it is helpful to consider them as part of the same overall project: to reduce and analyze meanings contained in large-*N* textual data.

Early versions of text analysis relied on hand-coding. The unit of analysis might be the word, line, sentence, or paragraph, and the objective would be to code each unit along some set of parameters. Some years ago, Gerring wrote a book about party ideologies in the United States that relied, in part, on this sort of coding to differentiate party positions of the major American parties from the early nineteenth century to the present. Thus, Gerring coded the parties' positions on specific issues like tariffs and more general philosophical matters like the proper role of government in American society (Gerring 1998). Other work in this tradition has looked at the content of presidential speech (Ceaser et al. 1981), the development of American national identity (Merritt 1966), and cross-national party ideologies (Budge 2000).

Qualitative data management software such as ATLAS-ti, MAXqda2, or NVivo (reviewed above) may prove useful in assisting in the process of hand-coding, speeding up the process and accuracy with which texts can be coded and those codings stored and analyzed, and offering a handy way of retrieving the original texts for in-depth analysis or direct quotations.

The old tradition of hand-coding is now complemented by a slew of techniques that process words in an automated or semi-automated fashion. This has the advantage of incorporating a much larger (in principle, infinite) number of texts, and thus may span longer ranges of time and more contexts. It is also sometimes easier to obtain a sample that is representative of an identifiable population. And it distances (though never entirely removes) the coder from the process of coding, mitigating one source of researcher bias.

Some texts, such as those stored in text-readable format on the Internet, are already in a format suitable for automated analysis. Google has developed several online tools for this purpose. Google Trends tracks the frequency of online searches. This has been shown to be useful in predicting outcomes such as elections, flu outbreaks, and consumer behavior, as well as for measuring the salience of various issues (Granka 2013; Mellon 2013). Google Ngram Viewer counts the number of times a given word or phrase appears in the Google Books repository, a historical library of digitized books. This may be used to construct a timeline of the frequency of mentions of a keyword (e.g., "democracy"), and this in turn may be interpreted for clues about the development and salience of a concept. One may also track Twitter feeds, Facebook posts, and other social media data. Increasingly a number of political questions are being explored with data from various Web platforms.

The automated analysis of most texts requires first converting those texts into a machine-readable format, i.e., a text file. If the original is a hard copy, it may be scanned and then converted to text with optical-character-recognition (OCR) software. If the original is a PDF, it may be directly convertible into text. If it exists as part of a database like Lexis Nexis or ProQuest, it may be possible to download the texts of interest as a single batch. If the texts must be extracted from somewhere on the Web, various Web crawlers (or scrapers) may be employed. These sorts of tools generally involve some programming skill and are thus not in the purview of most social scientists, though this is likely to change as demand increases and packaged programs for data scraping become available.[178] Once texts are in a machine-readable format, one may proceed to employ various analytic techniques.[179]

Dictionary techniques compute "the rate at which key words appear in a text to classify documents into categories or to measure the extent to which documents belong to particular categories" (Grimmer and Stewart 2013). For example, one might wish to measure the "positive" or "negative" tone of various documents. Using a pre-set dictionary of words, each of which is classified as one or another, one may arrive at a summary measure, which can then be compared across texts. Much depends, evidently, on the choice of a dictionary by which texts will be analyzed.

One interesting approach to this problem derives the dictionary from texts identified as playing an especially influential or paradigmatic role. These key texts provide the reference point by which all other texts are analyzed. For example, in order to estimate ideological location one might choose several reference texts that seem archetypically "liberal/left" or "conservative/right." The frequency distribution of words in these texts may then be compared with other political texts in order to identify the latent ideology of the authors of those texts (Laver, Benoit, and Garry 2003).

Supervised learning techniques begin with old-fashioned coding by humans (either the researcher or his or her accomplices). From repeated codings, a computer program learns how to replicate the coding process, which it can then replicate for other documents. **Unsupervised techniques** rely on algorithms to sort texts into categories and to measure distances separating texts and/or piles.

It should be apparent that the results of any automated textual analysis bear close scrutiny, both for purposes of interpretation (what do the discovered patterns mean?) and validation (do they mean what the author thinks they mean?). The key point for present purposes is that we are nowhere near a situation in which artificial intelligence can replace human intelligence. Analysis by humans thus remains central to the analysis of texts produced by humans.

Examples, Data, and Software

In the statistics chapters that follow we use the same examples as in the previous sections of the book: social capital, democracy and worker-training programs. In doing so we mostly rely on hypothetical data for two reasons: (1) to make the examples as easy as possible to follow; and (2) to allow readers to work through the analyses by hand. In addition we frequently make reference to data that would allow the reader to apply these techniques to related and real inquiries. It should go without saying then that substantive inferences should not be drawn from the examples below.

Many countries have one or two standard surveys that are repeated at regular intervals to give snapshots of political, social, or economic features of that society. In the United States, the longest-running and most popular nationally representative survey of voting age Americans is the American National Election Studies (ANES).[180] Like the election years before it, in 2012 the ANES collected a host of information, including demographics, political and social behavior, as well as preferences and attitudes on politics, society, and the economy. The data allows us to explore features of social capital, like group membership, representation, political engagement, as well as potentially related factors, like identification with political parties and feelings toward political institutions and figures. An excellent source for data on comparative democracy is the Quality of Government (QoG) Institute. The QoG offers a number of datasets, including the Expert Survey of public administrators in 107 countries, the Social Policy dataset, which includes a number of social issues and conditions particularly for the OECD countries, and the Regional dataset with regional information on corruption in the EU. Perhaps most notably, they house the Standard dataset, a cross-sectional time-series with global coverage from 1946 to 2012 on various government quality measures and correlates. They have grouped their Standard dataset variables by a heuristic, which includes their "What It Is" variables, like corruption, bureaucratic quality, and democracy, and their "How To Get It" variables, pertaining to electoral rules, forms of government, federalism, legal and colonial origin, religion and social fractionalization, as well as their "What You Get" variables, such as economic and human development, international and domestic peace, environmental sustainability, gender equality, and satisfied, trusting, and confident citizens.[181]

Those interested in the worker-training examples might utilize data from LaLonde (1986) and Dehejia and Wahba (1999), which we will refer to as the LaLonde data. The LaLonde data is used to evaluate worker-training programs on earnings and includes an indicator of assignment to training programs as well as demographic variables for 313 adults. It is a sub-sample of data that includes randomized experimental data from the National Supported Work Demonstration (NSW) and non-experimental comparison groups from the Population Survey of Income Dynamics (PSID) and the Current Population Survey (CPS).[182]

You are advised to practice the techniques introduced in these chapters by hand, as well as with a dataset and statistical software such as R, SAS, SPSS, or Stata. There is no clear champion of statistical software, with each providing relative advantages. SPSS is perhaps the most user-friendly for those with little programming experience, but has traditionally been the least versatile. Stata, followed by SAS and R have steeper learning curves but include a host of more advanced features. In terms of pricing, all are proprietary and costly, apart from R which is free and open-source to boot. The tables and graphs in the statistics chapters have all been produced with R. Because different software is used across the disciplines – and even within disciplines – we do not include play-by-play introductions to any one of them. Such introductions are readily available, however, either in hardcopy or on the Web. In short, the best way to learn is to do.

CONCLUSIONS

This chapter began by addressing issues of data management. In doing so, we distinguished four types of data: qualitative, medium-*N*, large-*N*, and textual. Each involves somewhat different techniques which must be mastered if one is to work with data of that sort.

KEY TERMS

- Qualitative data
- Medium-N data
- Large-N data
- Textual data
- Truth-table
- Qualitative comparative analysis
- Lagged variables
- Codebook
- Meta-data
- Content analysis
- Textual analysis
- Dictionary techniques
- Supervised learning techniques
- Unsupervised learning techniques