TWO

The Use of Statistical Methods in Social Research

Chapter Contents

What Statistics Is Not Good At

Describing Unique Phenomena in Great Detail

Statistics is a tool for discovering meaningful information from a large amount of numeric data. It is most useful for obtaining concise and precise information about a large number of cases. Cases may come in many different forms: groups of human beings, buffalos, crops, microchips, accidents, web pages, and so forth. When it is more important to know the characteristics of these cases as a whole than to learn about each particular unit, statistics starts to shine. It is simply too hard for human brains to detect any meaningful patterns in a large matrix of numbers. Statistics comes to the rescue with a few numbers and equations that summarize the patterns.

Conversely, this statistics is a very clumsy tool when the interest is in the details of one or very few unique cases the idiosyncrasies of which can be represented in many different aspects. For example, anthropologists try to understand the uniqueness of a very small number of cases in a particular context. They routinely carry out this kind of work by staying in a unique community for years, taking extremely detailed field notes, and finally writing up what Clifford Geertz calls 'thick description' (1973: 5–10). In sociology, beginning with Max Weber, there has been a long tradition of understanding meanings, interpretations, values and contexts. Comparative studies on a small number of cases – the so-called 'small-N' studies – have been an attractive research method to many students of social, historical and political sciences.

For example, why have some former communist nations been successful in transforming their economies while others failed? There are not many former communist nations in the world, and not all of them have tried to transform their economies. Moreover, each nation's economic development is unique because it is a product of the combined effects of many factors, including history, ideology, leadership, international relations, cultural values, its citizens' educational levels, political structure, and so forth. Here, we have many complicated variables the values of which are hard to define and very few cases within which to show any pattern, this is a situation in which statistics is not of much use.

Representing Relations and Networks

Obviously, social relations are of considerable importance when trying to understanding social phenomena. In the mid-1940s, Harvard University established the interdisciplinary Department of Social Relations under the leadership of Talcott Parsons. Nevertheless, it was not until the late 1970s that social scientists started to be able to study social relations directly and rigorously. Beforehand, social relations were discussed in abstract terms or through analogies. Following development over the past three decades or so, social network analysis (SNA) has become a very effective and popular method (Scott, 2000; Carrington et al., 2005).

The mathematical foundation of SNA is graph theory, which is different from the foundations of statistics: probability and matrix algebra. A graph consists of vertexes and edges (or arcs) that connect the vertexes. When used in social sciences, the vertexes can represent individual human beings or larger groups, and the edges can represent a variety of social relations. With statistics, we study the attributes of the vertexes, not the relations among the vertexes. That is, statistics cannot help us represent and analyse the overall structure of social relations among a group of individual units. To social scientists, this is a serious limitation, because in many situations it is structural positions, not attributes, that account for people's actions. To study social structure directly, you need to employ SNA, not statistics.

However, this does not mean that statistics has no part to play in the studying social networks. First of all, statistics can do a good job of measuring the quality of social relations. Even after years of development, many measures used in SNA remain crude, such as naming friends, having lunch, playing for the same sports club, sitting in the same board of directors, etc., which are barely able to capture the subtleness and complexity of the quality of social relations. In my opinion, SNA researchers should make use of statistical tools such as latent variables or indexes to obtain more refined measurements of social relations and incorporate these measurements in their analyses. Recent developments in SNA show that some statistical methods are highly useful for advanced analyses of social network data (Koehly and Pattison, 2005; Snijders, 2005).

Identifying Causal Relations

We shall come back to the difficult subject of causal inference in Chapter 11. Here we discuss it in light of the limitations of statistical methods. Broadly speaking, making a causal argument involves not one but several tasks: identifying causal relations; modelling complex causal relations; verifying the causal relations; and measuring the size of causal effect. The statistics most serious limitation lies in its inability of identifying causal connections, but it can be of great value for measuring and modelling.

It is important to recognize that the establishment of causal relations is a cognitive process. The idea of the existence of a certain causal relationship comes to our mind in different ways: the observation of regular connections among several phenomena, the discovery of unusual phenomena, the deduction of axioms, and so forth. In these activities, we need heuristics, not statistics (Abbott, 2004). This is so because the function of statistics is to numerically describe what has already happened, not to speculate, hypothesize, or theorize some unobservable processes that may be at work. Statistics can help us to evaluate how well a piece of numeric evidence would support a perceived causal process, but it is not able to offer an explanation for that process. What we should observe and what we can say about unobservable processes are beyond the capabilities of statistics.

This also explains why statisticians have been extremely cautious of moving from association to causation, especially for non-experimental studies. Most recently, techniques drawing on the idea of counterfactuals have made some progress in confirming the causal effect of candidate causal factors (Morgan and Winship, 2007). No matter how sophisticated these methods are, however, the candidate causal factors arise from theories and current knowledge; they cannot come from statistical methods. What statistics can do is to measure and model the proposed causal connections when conditions (research design, quality of data, and so on) permit. Deriving causal statements completely based on statistical models is to manipulate statistics beyond its capabilities.

Modelling Nonlinear Dynamics

There are at least three things about nonlinear dynamic processes that statistical methods will find very hard to cope with. First, the data underpinning statistical analysis do not contain sufficient information about the dynamic entity, be it a group of human beings, a social movement, or any other social phenomenon. The information underpinning statistical analysis is *the attributes* of the entity, not the records of *how the entity's behaviours have evolved over time*. In certain circumstances, the temporal changes of these attributes may represent the changes of the entities themselves, but there is always a gap between the two, and in some situations the attributes may give a distorted reflection of the dynamic process. Therefore, it is at best clumsy and at worst inadequate to take the attributes as proxies of the process. The objective of statistical methods is to study the varied distribution of the attributes across cases, not the changes that the targeted entity has gone through. This inability to study dynamics derives from the restricted structure of the statistical analysis data. Statistics is good at representing the variation of variables across cases rather than any actual processes.

In relation to that, there is a limit to which statistical methods can incorporate time, an essential element of dynamic processes. Data collected from standalone cross-sectional surveys are inherently static. Although the whole data collection process may take a long period of time, such as a year, it is assumed that the data refer to the situation at one particular time point. Data collected in repeated, cross-sectional social surveys do reveal situations at different time points. Nevertheless, because the respondents differ from one time point to another, that is, the entities under study have changed, the data simply cannot show any dynamic processes of the same thing unless it can be established that all the samples represent the same population. Data collected with longitudinal designs are best suited for statistical analyses of temporal processes, but even here the analysis of dynamics is seriously limited due to the restricted structure of the data matrix. As we shall see later in Chapter 8, longitudinal models tell us the probability that a variable's effect changes across time points rather than models the dynamic directly. Such studies are useful for exploring the patterns within the data, but they constrain our theoretical imagination because our theories of a particular dynamic process may not be properly represented by the attributes. Statistical methods may be of great use for measuring the actual change of some important quantities, but they cannot supply a specific functional relationship that links the interested entities at different time points. The specific function that models the dynamic should be backed up by a strong theory, not derived from fitness to a particular set of data.

This brings us to another limitation of statistics: it is very likely that functions that model dynamic processes and are supported by a theory involve nonlinear relationships, because changes are rarely constant, universal or stable. As they are inconstant, specific and unstable, dynamic processes usually pass different phases. To represent these phases with a single mathematical function is very hard, and it cannot be done without a specific theory. Statistical methods, designed to explore and represent the regular patterns in the data collected in a specific context, have to rely heavily on linear models and their transformations because that is the most convenient form for specifying relationships. Although efforts are made to ensure that the linearity assumption is reasonable, the modelling process rarely begins with a nonlinear idea.

What Statistics Is Good At

To many social scientists, causal explanation is the most important objective (Stinchcombe, 1968; Abbott, 2004; Cartwright, 2007). There should be no doubt that social researchers should make every effort to achieve causal explanation whenever they can. Nevertheless, I would take a softer line than that: explanation is *the* purpose of social science research. I would not even claim that it is *the most important* purpose. There are many other things social researchers should aspire to achieve, and statistical methods can make valuable contributions to these seemingly less honourable undertakings. The uses listed and discussed below are for illustration only, they are not meant to be exhaustive and mutually exclusive.

Establishing the Target Phenomenon

While planning his prefatory paper for the 13th issue of *Annual Review of Sociology*, Robert K. Merton (1987) gave the editors a list of 45 topics on which he could write focused discussions. The space allocated to him, however, could accommodate only three. The three selected, being each ranked 1 of 15, indicate the importance of those topics. Of the three selected, 'Establishing the Phenomenon' came first. The reason cannot be clearer: social scientists would waste a huge amount of resources if they reached the end of their research only to find that the phenomenon they have planned to explain actually does not exist or exists in a different form. As Merton shrewdly reminds us:

> In the abstract, it needs hardly be said that before one proceeds to explain or to interpret a phenomenon, it is advisable to establish that the phenomenon actually exists, that it is enough of a regularity to require and to allow explanation. Yet, sometimes in science as often in everyday life, explanations are provided of matters that are not and never were. (1987: 2)

He then illustrated such careless practice in both natural and social sciences.

How could such embarrassing incidents occur in the first place? Why do people rush to explain something even before checking out the existence of what they are trying to explain? Although it is likely that some do it on purpose, let us believe that the majority of researchers maintain a high level of intellectual integrity. And alternatively, assume that they simply forget to verify the existence of their phenomena before explaining them, how can they forget? This is not a simple matter of bad memory.

Here is a more plausible explanation. The researcher is in an exciting process of justifying or illustrating the power of a theory with a piece of information obtained from a particular source. The main worry is that information will slow down the whole research process, and worse, if the information turns out to be incorrect, then the whole project will crumble. Consequently, the researcher wishfully assumes the establishment of the phenomenon. Another possibility is that the researcher may find little value in fact-finding activities because the main interest is in theorizing, explaining, or interpreting. Fact-finding is something of low academic value and can be easily done by others.

To avoid these embarrassing and wasteful incidents, it would be useful to require that all research questions be preceded with a statement about the factual existence of the phenomenon under study. Pertinent to the subject of this book, although not all facts are numeric, reliable statistics can supply the evidence for establishing the phenomenon. Using statistics is not the only way of establishing the phenomenon, of course, and it can be a highly controversial process: the definition of the phenomenon, the theoretical as well as empirical boundaries, the trustworthiness of the evidence, and so on. How to resolve these disputes is beyond statistical analysis, but regardless how they are resolved, producing statistics is perhaps the best way of establishing the phenomenon.

Detecting Patterns Among a Huge Amount of Information

Most statistical methods are data reduction tools, most useful when the amount of numeric information tends to overwhelm our cognitive capacity. With a few numbers, equations, or graphs, statistical analyses can reveal patterns hidden in a data matrix of thousands of cases by hundreds of variables, or even several such matrices. With statistical procedures properly followed, not only can researchers establish the phenomenon in question, but they can discover patterns that they may have never thought about before. The phrase 'data mining' best expresses such exploratory process: we may have a vague idea of what we shall find based on our common sense – and it is a commendable practice to keep common sense in our mind – but we can never be completely certain about exactly what we shall find until we see the results. I personally find it the most exciting experience of using statistics to challenge the status quo with numeric results, an excellent opportunity to show the value of statistical work. A large amount of money has been invested in social surveys, which then have generated a huge amount of numeric data. It is an academic sin not to maximize the use of this freely available source of information.

Comparing Groups (Broadly Defined)

Human beings differ along many dimensions: gender, age, race and ethnicity, wealth, cultural values and so on. To describe, understand and explain these differences constitutes a major research activity for statistical analysis (Harkness et al., 2002; Liao, 2002). Statistical methods cannot help us explain *why* people are so different, but they are good at describing *how* they are different. Again, it is presumptuous to think that all 'the why questions' are necessarily answerable, worth answering and more important than 'the how questions'. In social reality, many 'how questions' with regards to human groups are much more important than 'why questions' because their answers provide the starting point for actions, *regardless of whether we know why people are so different*. Clearly, it is one of statistics' specialties to measure the magnitude, the scale and the scope of group differences, and further to connect these differences with other interested factors. One particular contribution that statistics can make is to assess the comparability of measurements used across different groups, such as communities, organizations or nations. For example, results from different social surveys may not be comparable even though the wordings of the original survey questions are exactly the same. Some other factors, such as people's different understandings of the questions, may render the results incomparable. By carrying out a few statistical tests, we may make our comparisons more reliable.

Measuring the Unique Effects of Risk (or Contributing) Factors

Often researchers have several factors in mind that they believe have at least some partial causal effects on the target phenomenon. The causal effect is established not through experiments but through what John Goldthorpe (2000) calls 'robust association' in observational studies; that is, although we cannot make causal arguments purely based on observed associations, if the association exists in almost all the situations that we have examined, then there is at least no clear evidence against the proposed causal relationship. We may not have sufficient evidence to show that a few factors are really the causes and how they work, but if their effects persist, we must pay careful attention to the size of these effects. This is the logic followed by medical doctors who are able to explain why it is more likely for a particular type of people, say females above the age of 60, to get a particular disease, but they would be well prepared to take these 'risk factors' (gender and age) very seriously.

Evaluating and Assessing the Impact of Policies, Actions or Events

Statistical methods can help us evaluate the impact of policies, actions or events (Rossi and Wright, 1984; Freudenburg, 1986). Will the introduction of speed cameras make drivers slow down? How many peoples' financial and marital lives will be affected, desirably and undesirably, by the building of a casino centre? Questions such as these abound in policy-related issues, and statistics can offer valuable help for answering them. Perhaps this is why Andrew Abbott claims that the only place where statistics can find itself of some use is in social policy while condemning the practice of employing statistics to make causal arguments (2004: 40). The contribution by statistics is widely deemed to be legitimate and helpful because policy-makers and researchers are usually content with the discovery of regularities as a reasonably good basis for taking actions. 'The point was to decide whether to take some action, not to understand mechanisms' (Abbott, 2004: 38).

Types of Statistical Methods

Univariate, Bivariate and Multivariate

The first distinction refers to the number of variables involved. An analysis is 'univariate' if it is focused on only one variable. An analysis of two variables is 'bivariate' and one of multiple variables 'multivariate'. It is very rare for any serious analysis to completely focus on only a single variable, but knowledge of one variable is necessary for understanding more complicated situations with multiple variables.

Numeric Versus Graphic

Numeric methods produce exact and meaningful numbers, while graphic methods visualize the data with charts or plots to show patterns, unusual cases, clustering and other features in the data. It is too big an exaggeration to say that 'a graph is worth more than a thousand words'. If it is true, all statistical studies should be presented with graphs rather than numbers. Graphs can reveal things that are hard to see in numbers, but they are much less accurate and precise than numbers, and there are situations in which graphs can hardly struggle to show anything meaningful that is difficult to show with numbers. For example, a table can be much more effective than a graph because the reader does not have to go back and forth between the graph and the key. Besides, graphs produced with advanced statistics may be hard to understand. The bottom line is to use both types of methods in a complementary manner in order to achieve clarity and accessibility.

Cross-sectional Versus. Longitudinal

This is a research design issue. Most studies collect data on the same group of respondents at only one time point. They are 'cross-sectional' because their aim is to cover the situations in different sections of the population. Obviously, these studies are static by design. If the main objective is to study changing situations temporally, then the same group of respondents must be followed and contacted at each one of several time points. This type of study is longitudinal, which is technically more difficult and financially more costly.

The distinction between 'cross-sectional' and 'longitudinal' is, however, not as clear-cut as it appears. Some cross-sectional social surveys are repeated at multiple time points. The participating respondents are different from one time point to another, but the operation and the contents of the study (sampling schemes, instruments used, administration) remain the same. This type of study is not as good as completely longitudinal studies for the purpose of studying change because it is hard to tell whether the observed changes are effects of the interested factors or of the differences among the respondents. In contrast, sometimes longitudinal studies may have to incorporate elements of cross-sectional design. For example, attrition, the loss of some respondents from one wave to another, is a constant problem for longitudinal surveys. To make up the loss, researchers usually recruit new respondents in a particular wave in order to maintain a desired sample size. As a consequence, the group of respondents do not remain completely the same. Different statistical methods will be needed depending on how the data were collected. The general principle is to incorporate features of the design into data analysis.

Descriptive Versus Inferential

To make the last distinction, we need first to make the distinction between a population and a sample. The population is the ultimate target of our investigation, which must be defined before any empirical work is conducted. For most large-scale sample surveys, the target population is usually all the adults living in private residences at a particular point of time, but one must think the specific population carefully for a particular research purpose.

For various reasons, either because we cannot study the whole population directly or because there is no need to do so, we draw a sample from the population in the hope that the information drawn from the sample would still allow us to say something about the population. If we make conclusions about the population based on our analysis of the sample data, then we are doing inferential statistics; that is, we are inferring from the sample to the population. In contrast, if our analysis focuses on the sample or the population alone, then we are doing descriptive statistics. A bad practice commonly seen in social research is not clearly stating whether the study targets a particular population, and if there is an interested population, how it is defined. Alternatively, little attention is paid to the connection between the population and its sample. This is a very important issue because it determines the kind of statistics – descriptive or inferential – used.

Ten Rules of Using Statistics

During the past years of learning, using and teaching statistics, I have built up a set of rules of using statistics in social research. I have found it very important and useful to keep them in mind.[1]

Rule No. 1 – Understand the Subject Matter

The raw material for statistical analysis is the data matrix. On appearance, a matrix is just a set of numbers, and statistics is mostly concerned with the procedures that process the numbers. The numbers do carry meanings, of course, and they are produced in a particular context. It should be common sense now that data must be understood and interpreted in relation to a particular subject matter and its specific context (van Belle, 2002: 4). But this point is still worth repeating, especially for novice users, because the subject matter and its context are not directly attached to the data, so researchers tend to forget them when their attention is focused on the data. Without intensive training, our human brain is simply not good at considering things that are out of sight. Researchers must realize that the process of turning every piece of information into a number has further separated the data from its subject matter.

In contrast, for 'qualitative data' such as documents, conversations and photos, the data and their meanings are tightly intertwined, thus forcing us to go back to the data's context and meaning. Researchers using statistics need to avoid becoming buried so deeply in statistical procedures that they lose sight of what they plan to argue substantively. An effective method is to repeatedly ask yourself: 'What can I say based on the data and my analysis?'

Rule No. 2 – Learn How the Data Were Collected and Examine Their Quality

It is safe to say that the researchers no longer collects their own large numeric datasets. Today, many researchers analyse secondary data, that is, data that were collected by others. Even when they are involved in collecting primary data, this is not a one person job, and consequently they have to accept that at least a part of the

[1] I may not be able to identify the original source of all these rules, so I apologize if I couldn't give credit to the person who firstly suggested it.

data collection process that has been carried out by their co-researchers. This rule is thus valid not only for researchers using secondary data but for all researchers.

To follow this rule, try to answer the following questions. An answer may not be always available for every question listed here, but researchers should try as hard as they can to find them. First, there are a series of questions about the nature of the study from which the data were collected, and researchers should keep in mind the implications when conducting subsequent analysis. Is it an experimental, quasi-experimental, or an observational study? Is time an essential element in the research design? Answers to these questions may indicate how far the researcher can go in making a causal argument.

Next, are there a population and a sample? How is the population defined? If there is a sample, how is it drawn? The answers will determine whether statistical inference can be appropriately made, and the answers will affect weighting and statistical estimates as well.

We also need to consider the method of data collection: How was the data collection administered, by telephone, personal interviews, post or email? What were the sample size and the response rate? Information about these issues will highlight the quality of the data. Even for a study with a high response rate, there are usually many missing values. But the data matrix usually appears to be very clean when one is analysing a secondary dataset. How was it cleaned up? This is a highly difficult issue, but we should try at least to answer the following three questions when reporting the results so that our reader can have a good sense of the data quality:

(1) Why some data were missing – was it failure of contact, refusal, lack of knowledge of the subject matter?
(2) How have the missing data been handled – ignoring them, imputing from available data with an average, or nearby values, or other method?
(3) What are the effects of the previous strategy on the final results and conclusions?

The final issue relates to variables: How many variables are there to estimate the same phenomenon or concept? What was the original rationale of creating those items? What were the options from which respondents chose? How were they worded? Answers to these questions will affect the choice of statistical procedures and interpretations of statistical results.

Rule No. 3 – When Studying a Single Variable, Analyse and Report Statistics of Both its Centre and its Spread

Most people are satisfied with statistics that describe the centre of a variable, such as the mode (the most frequent value), the median (the value in the middle) or the mean (the arithmetic or the weighted average), as they offer a quick and clear summary of the variable's values.

For at least two reasons, any of these alone is not sufficient for depicting the whole picture of a variable. First, our real interest should be *the overall distribution* of a variable's values, of which the centre is only a part. There are several other things we must look at, such as the number of modes, cyclical patterns, is there a skew to one side, and so forth. At the minimum, we should know the degree of variation of the values. 'A measure of centre alone can be misleading', as McCabe and Moore have reminded us, 'The simplest useful numerical description of a distribution consists of both a measure of centre and a measure of spread' (2006: 44). The measure of spread includes index of variation, interquartile range and variance. We shall learn the exact meaning of these terms in Chapter 5, but the reader should keep this general rule in mind from now on.

Another reason that we should not be satisfied with a measure of centre alone is that it varies from sample to sample. One can easily see this in the results published by different polling agents. Which one should we trust? None of them, because they are all from a sample, and sample results are always different from the true but unknown value in the population. Loosely speaking, we call the uncertainty of sampling results 'sampling errors'. There are other types of errors, but the sample error alone should be sufficient to make the point that one should not be concerned with the specific statistics itself. What we should be concerned with are the procedures that have been adopted to produce the statistics and the possible range of variations from sample to sample.

Rule No. 4 – Use Both Numeric and Graphic Methods as Supplementary Tools

Words and graphs serve different functions and thus have different utilities. Graphs are intuitively appealing and effective of showing relative positions in a single picture. However, these advantages come with some conditions. For example, it is unnecessary and perhaps a waste of space to produce a graph when there are very few pieces of information. Van Belle listed two situations in which it is not useful to graph data: 'when there are few data points, or when there are too many relationships to be investigated. In the latter case a table may be more effective' (2002: 159). Therefore, he suggests that we use 'sentence structure for displaying 2 to 5 numbers, tables for displaying more numerical information, and graphs for complex relationships' (2002: 154). A table may be clearer and more informative than a graph when all variables are categorical and we want to compare the groups defined by these variables. No matter how we will present the data, it will be a trail-and-error process, not a one-off activity.

Rule No. 5 – Refrain from Using a Pie Chart or a Bar Chart (Especially a Stacked or Three-dimensional Bar Chart)

The reader may be perplexed by this rule, as pie charts and bar charts are widely used in all sorts of publications and are standard functions in statistical software, why should we not use them at all? Actually, van Belle suggests that we not use pie and bar charts at all (2002: 160–7). I think it is a bit too radical to ban pie and bar charts once and for all; therefore, I suggest to refrain from using these charts.

The best way of making sense of this rule is to do the following: put a pie chart (or a bar chart) and the data table based on which the chart is produced side by side, examine them and ask yourself 'What is the added value of the chart?' The answer would be 'very little, none, or even negative'. When there are a limited number of categories in the data table, there is no need to put them in a chart – in this situation our brain is capable of recognizing the information in the original data. If we use a pie or bar chart, we will have to use a key, thereby forcing our eyes to go back and forth between the chart and the key. We may also have to put the value for each category in its corresponding slice in order to know the exact value – we are not that good at recognizing which slice is larger when two have very similar values, which repeats the information in the table.

In addition, when there are a large number of categories in the data table, the chart simply cannot handle so much information – it will look very crowded, especially so when some slices are too tiny to be clearly recognized. Similarly, in a bar chart, bars for categories whose values are very small will be suppressed by categories of much bigger values, making it very hard for us to figure out exactly how small they are. Some would say that an advantage of a bar chart is to sort the categories in order so that we can detect a descending or ascending trend. This, however, can be easily and more effectively achieved by a table with all categories being sorted in order.

That said, I think that in one situation pie and bar charts can be very useful, that is, when we try to compare several groups with a pie or bar chart representing each of them, and each pie or bar chart can show the information clearly. A single pie or bar chart is of little value.

Rule No. 6 – When Creating a Table or a Graph, Draw it by Hand on a Piece of Paper before Producing it on a Computer.

I realized the value of this rule from my own experience: there are so many times that I produced a table or a graph directly on a computer, only to find that I needed one more row or column, some parts were not necessary, the direction should have been reversed, or there was a better way of presenting the same information. We all know that a good practice of time-saving is to prepare a shopping list before shopping in a supermarket, because we know much better what we exactly want when at home. We should follow a similar practice in creating a table or a graph. The increasing power of computers should not preclude manual alternatives. Before making a table or a graph, we had better ask ourselves: (1) What do I really want to show in this table or graph? (2) What are the expected values or patterns? (3) What will it make me able to say about the substantive issue?

Rule No. 7 – Do Not Cut and Paste All Outputs Generated by Computer Software

In order to make computer software popular, computer programmers want to satisfy as many different demands as possible, so that all sorts of people can find what they want. An advantage for the software's marketability, it may not be desirable for a particular user. As a user, you should know exactly what *you* want from the software and what the software has produced for you. You must be able to know which pieces of information in the outputs are exactly what you want, which are relevant and which can be discarded. Learn as much as you can about what a particular piece of computer software can do for you, so that you know as much as you can about what you can delegated to the computer to do and what you have kept back for your independent thinking.

Rule No. 8 – Edit the Outputs That You Want to Include in Your Report

Similar to deleting redundant words or breaking up a long sentence into easily understandable short ones, editing is no less important for presenting statistical work. For example, one of my statistics teachers at Columbia University, Andrew Gelman, used to ask us to keep only two decimal digits. Many statistical software programs, however, automatically produce numbers with four or even six decimal digits, which is both unnecessary and distractive for most purposes. More generally, decimal numbers may not make sense at all for some variables, such as sex, ethnic groups, or cities. For another example, after producing a table, you may find some cells empty, so you may think whether it is necessary and sensible to merge some of them with others. There are some issues of style as well: lines under the categories should be darker than others, and many academic journals require that no vertical lines be used in a table, some numbers must be flagged up with asteroids to signal statistical significance, and a note of resources may be needed under the table.

Rule No. 9 – No Cheating in Making Causal Statements.

One thing you must never forget after reading this book or taking any statistics class, no matter how elementary, is that correlation is not causation. Perhaps you have learnt it or have heard it somewhere, but it is still worth emphasizing. Statisticians are extremely cautious about making causal statements. There is even a legendary story about this. A well-known statistics professor was dying, but he could not close his eyes before checking one thing with his students. Summoning them to his deathbed, the professor prompted them by saying: 'Correlation is …' Fortunately, his students did not let him down.

Albeit common sense now, this rule has been manipulated by many researchers for the purpose of claiming a discovery of causal connections among some social phenomena. Indeed, it is very tempting *not to* make such claim when you see strong and statistically significant correlations in the results. Being aware of the embarrassment of confusing the two, many researchers would not explicitly use the term 'cause' or 'causal', but it is very clear that nothing but causation is in their mind. Such disguised causal statements come in a variety of forms, for example, 'A has happened, consequently, we have witnessed B', 'the statistical results indicate a strong and significant impact of A on B' or 'given the statistical significance of this coefficient, it is no wonder that we observe such a big magnitude of B'.

Rule No. 10 – Translate a Statistical Model into Words and Ask: Do Those Words Make Sense?

As a human device for representing a connection or a process in reality, models come in a variety of forms: verbal, graphical, numerical or visual. Very often, each can be transformed into another, and such transformations help us gain extra knowledge and insight that cannot be easily seen in an alternative form. Statistical methods usually present models with mathematical equations, which we can submit to independent procedures, not only making our analysis more rigorous but also producing some results that we can hardly expect to see beforehand.

Nevertheless, to us as social researchers, mathematical operations are not the ultimate concern. Rather, we want to say something substantively fresh and important. Presenting our model in a verbal form can help us ensure that we have not lost contact with the background, the meaning and the relevance of the statistics. Most social researchers know that they must address their substantive research question with statistical analysis. My observation is that most of their attention is given only to *the results*, but we must know what the model says before we estimate its parameters and interpret its results. A statistical model is not just an equation with symbols on each side; it says something about what the model builder believes. However, these beliefs, assumptions and understanding are usually implied in the model, hidden from direct observation, and we need to talk them out if we want to understand and use the model meaningfully.