Chapter 1
Introduction

David J. Hand
Imperial College, United Kingdom

1.1. Why "Intelligent Data Analysis"?

It must be obvious to everyone - to everyone who is reading this book, at least - that progress in computer technology is radically altering human life. Some of the changes are subtle and concealed. The microprocessors that control traffic lights or dishwashers, are examples. But others are overt and striking. The very word processor on which I am creating this chapter could not have been imagined 50 years ago; speech recognition devices, such as are now available for attachment to PCs, could have been imagined, but no-one would have had any idea of how to build such a thing.

This book is about one of those overt and striking changes: the way in which computer technology is enabling us to answer questions which would have defied an answer, perhaps even have defied a formulation, only a few decades ago. In particular, this book is about a technology which rides on top of the progress in electronic and computer hardware: the technology of data analysis.

It is fair to say that modern data analysis is a very different kind of animal from anything which existed prior to about 1950. Indeed, it is no exaggeration to say that modern data is a very different kind of animal from anything which existed before. We will discuss in some detail exactly what is meant by data in the modern world in Section 1.3 but, to get the ball rolling, it seems more convenient to begin, in this section, by briefly examining the notion of "intelligent data analysis". Why analyse data? Why is this book concerned with "intelligent" data analysis? What is the alternative to "intelligent" data analysis? And so on. In between these two sections, in Section 1.2, we will look at the cause of all this change: the computer and its impact.

To get started, we will assume in this opening section that "data" simply comprise a collection of numerical values recording the magnitudes of various attributes of the objects under study. Then “data analysis” describes the processing of those data. Of course, one does not set out simply to analyse data. One always has some objective in mind: one wants to answer certain questions. These questions might be high level general questions, perhaps exploratory: for example, are there any interesting structures in the data? Are any records anomalous? Can we summarise the data in a convenient way? Or the questions might be more specifically confirmatory: Is this group different from that one? Does this attribute change over time? Can we predict the value of this attribute from the measured values of these? And so on.

Orthogonal to the exploratory/confirmatory distinction, we can also distinguish between descriptive and inferential analyses. A descriptive (or summarising) analysis is aimed at making a statement about the data set to hand. This might consist of observations on the entirety of a population (all employees of a corporation, all species of beetle which live in some locality), with the aim being to answer questions about that population: what is the proportion of females? How many of the beetle species have never been observed elsewhere? In contrast, an inferential analysis is aimed at trying to draw conclusions which have more general validity. What can we say about the likely proportion of females next year? Is the number of beetle species in this locality declining? Often inferential studies are based on samples from some population, and the aim is to try to make some general statement about the broader population, most (or some) of which has not been observed. Often it is not possible to observe all of the population (indeed, this may not always be well-defined - the population of London changes minute by minute).

The sorts of tools required for exploratory and confirmatory analyses differ, just as they do for descriptive and inferential analyses. Of course, there is considerable overlap - we are, at base, analysing data. Often, moreover, a tool which appears common is used in different ways. Take something as basic as the mean of a sample as an illustration. As a description of the sample, this is fixed and accurate and is the value - assuming no errors in the computation, of course. On the other hand, as a value derived in an inferential process, it is an estimate of the parameter of some distribution. The fact that it is based on a sample - that it is an estimate - means that it is not really what we are interested in. In some sense we expect it to be incorrect, to be subject to change (if we had taken a different sample, for example, we would expect it to be different), and to have distributional properties in its own right. The single number which has emerged from the computational process of calculating the mean will be used in different ways according to whether one is interested in description or inference. The fact that the mean of sample A is larger than the mean of sample B is an observed fact - and if someone asks which sample has the larger mean we reply “A”. This may be different from what we would reply to the question “Which population has the larger mean, that from which A was drawn or that from which B was drawn?” This is an inferential question, and the variability in the data (as measured by, for example, the standard deviations of the samples) may mean we