1

Introduction

Statistical methods have a very wide range of applications. They are commonplace in demographic, medical and meteorological studies, along with more recent extension into financial investments. Research into new techniques incurs little cost and, nowadays, large quantities of data are readily available. The academic world takes advantage of this and is prolific in publishing new techniques. The net result is that there are many hundreds of techniques, the vast majority of which offer negligible improvement for the process industry over those previously published. Further, the level of mathematics now involved in many methods puts them well beyond the understanding of most control engineers. This quotation from Henri Poincaré, although over 100 years old and directed at a different branch of mathematics, sums up the situation well.

> *In former times when one invented a new function it was for a practical purpose; today one invents them purposely to show up defects in the reasoning of our fathers and one will deduce from them only that.*

The reader will probably be familiar with some of the more commonly used statistical distributions – such as those described as uniform or normal (Gaussian). There are now over 250 published distributions, the majority of which are offspring of a much smaller number of parent distributions. The software industry has responded to this complexity by developing products that embed the complex theory and so remove any need for the user to understand it. For example, there are several products in which their developers pride themselves on including virtually every distribution function. While not approaching the same range of techniques, each new release of the common spreadsheet packages similarly includes additional statistical functions. While this has substantial practical value to the experienced engineer, it has the potential for an under-informed user to reach entirely wrong conclusions from analysing data.

Very few of the mathematical functions that describe published distributions are developed from a physical understanding of the mechanism that generated the data. Virtually all are empirical. Their existence is justified by the developer showing that they are better than a previously developed function at matching the true distribution of a given dataset. This is achieved by the

Statistics for Process Control Engineers: A Practical Approach, First Edition. Myke King.

inclusion of an additional fitting parameter in the function or by the addition of another non-linear term. No justification for the inclusion is given, other than it provides a more accurate fit. If applied to another dataset, there is thus no guarantee that the improvement would be replicated.

In principle there is nothing wrong with this approach. It is analogous to the control engineer developing an inferential property by regressing previously collected process data. Doing so requires the engineer to exercise judgement in ensuring the resulting inferential calculation makes engineering sense. He also has to balance potential improvements to its accuracy against the risk that the additional complexity reduces its robustness or creates difficult process dynamics. Much the same judgemental approach must be used when selecting and fitting a distribution function.