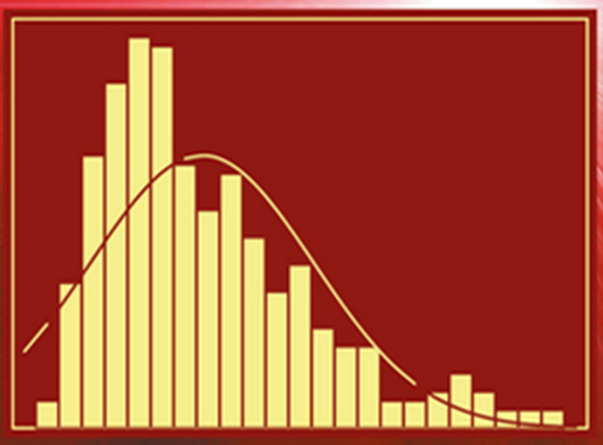
Myke King

Statistics for Process Control Engineers



A Practical Approach

WILEY

Statistics for Process Control Engineers

Statistics for Process Control Engineers

A Practical Approach

Myke King

Whitehouse Consulting, Isle of Wight, UK

WILEY

This edition first published 2017

Registered Office(s)
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Office
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Library of Congress Cataloging-in-Publication Data

Names: King, Myke, 1951–
Title: Statistics for process control engineers : a practical approach / Myke King.
Description: First edition. | Hoboken, NJ : Wiley, [2017] | Includes bibliographical references and index.
Identifiers: LCCN 2017013231 (print) | LCCN 2017015732 (ebook) | ISBN 9781119383482 (pdf)
ISBN 9781119383529 (epub) | ISBN 9781119383505 (cloth)
Subjects: LCSH: Process control–Mathematical models. | Engineering–Statistical methods. | Engineering mathematics.
Classification: LCC TS156.8 (ebook) | LCC TS156.8 .K487 2018 (print) | DDC 620.001/51–dc23
LC record available at https://lccn.loc.gov/2017013231

Cover design by Wiley
Cover Images: (Background) © kagankiris/Gettyimages;
(Graph) Courtesy of Myke King

Set in 10/12pt Times by SPi Global, Pondicherry, India

10 9 8 7 6 5 4 3 2 1

Contents

Preface

There are those that have a very cynical view of statistics. One only has to search the Internet to find quotations such as those from the author Mark Twain:

There are three kinds of lies: lies, damned lies, and statistics.
Facts are stubborn, but statistics are more pliable.

From the American humourist Evan Esar:

Statistics is the science of producing unreliable facts from reliable figures.

From the UK's shortest-serving prime minister George Canning:

I can prove anything by statistics except the truth.

And my personal favourite, attributed to many – all quoting different percentages!

76.3% of statistics are made up.

However, in the hands of a skilled process control engineer, statistics are an invaluable tool. Despite advanced control technology being well established in the process industry, the majority of site managers still do not fully appreciate its potential to improve process profitability. An important part of the engineer's job is to present strong evidence that such improvements are achievable or have been achieved. Perhaps one of the most insightful quotations is that from the physicist Ernest Rutherford.

If your experiment needs statistics, you ought to have done a better experiment.

Paraphrasing for the process control engineer:

If you need statistics to demonstrate that you have improved control of the process, you ought to have installed a better control scheme.

Statistics is certainly not an exact science. Like all the mathematical techniques that are applied to process control, or indeed to any branch of engineering, they need to be used alongside good engineering judgement. The process control engineer has a responsibility to ensure that statistical methods are properly applied. Misapplied they can make a sceptical manager even more sceptical about the economic value of improved control. Properly used they can turn a sceptic into a champion. The engineer needs to be well versed in their application. This book should help ensure so.

After writing the first edition of *Process Control: A Practical Approach*, it soon became apparent that not enough attention was given to the subject. Statistics are applied extensively at every stage of a process control project from estimation of potential benefits, throughout control design and finally to performance monitoring. In the second edition this was partially addressed by the inclusion of an additional chapter. However, in writing this, it quickly became apparent that the subject is huge. In much the same way that the quantity of published process control theory far outstrips more practical texts, the same applies to the subject of statistics – but to a much greater extent. For example, the publisher of this book currently offers over 2,000

titles on the subject but fewer than a dozen covering process control. Like process control theory, most published statistical theory has little application to the process industry, but within it are hidden a few very valuable techniques.

Of course, there are already many statistical methods routinely applied by control engineers – often as part of a software product. While many use these methods quite properly, there are numerous examples where the resulting conclusion later proves to be incorrect. This typically arises because the engineer is not fully aware of the underlying (incorrect) assumptions behind the method. There are also too many occasions where the methods are grossly misapplied or where licence fees are unnecessarily incurred for software that could easily be replicated by the control engineer using a spreadsheet package.

This book therefore has two objectives. The first is to ensure that the control engineer properly understands the techniques with which he or she might already be familiar. With the rapidly widening range of statistical software products (and the enthusiastic marketing of their developers), the risk of misapplication is growing proportionately. The user will reach the wrong conclusion about, for example, the economic value of a proposed control improvement or whether it is performing well after commissioning. The second objective is to extract, from the vast array of less well-known statistical techniques, those that a control engineer should find of practical value. They offer the opportunity to greatly improve the benefits captured by improved control.

A key intent in writing this book was to avoid unnecessarily taking the reader into theoretical detail. However the reader is encouraged to brave the mathematics involved. A deeper understanding of the available techniques should at least be of interest and potentially of great value in better understanding services and products that might be offered to the control engineer. While perhaps daunting to start with, the reader will get the full value from the book by reading it from cover to cover. A first glance at some of the mathematics might appear complex. There are symbols with which the reader may not be familiar. The reader should not be discouraged. The mathematics involved should be within the capabilities of a high school student. Chapters 4 to 6 take the reader through a step-by-step approach introducing each term and explaining its use in context that should be familiar to even the least experienced engineer. Chapter 11 specifically introduces the commonly used mathematical functions and their symbology. Once the reader's initial apprehension is overcome, all are shown to be quite simple. And, in any case, almost all exist as functions in the commonly used spreadsheet software products.

It is the nature of almost any engineering subject that the real gems of useful information get buried among the background detail. Listed here are the main items worthy of special attention by the engineer because of the impact they can have on the effectiveness of control design and performance.

- Control engineers use the terms 'accuracy' and 'precision' synonymously when describing the confidence they might have in a process measurement or inferential property. As explained in Chapter 4, not understanding the difference between these terms is probably the most common cause of poorly performing quality control schemes.

- The histogram is commonly used to help visualise the variation of a process measurement. For this, both the width of the bins and the starting point for the first bin must be chosen. Although there are techniques (described in this book) that help with the initial selection, they provide only a guide. Some adjustment by trial and error is required to ensure the resulting chart shows what is required. Kernel density estimation, described in Chapter 6, is a simple-to-apply, little-known technique that removes the need for this selection. Further it

generates a continuous curve rather than the discontinuous staircase shape of a histogram. This helps greatly in determining whether the data fit a particular continuous distribution.

- Control engineers typically use a few month's historical data for statistical analysis. While adequate for some applications, the size of the sample can be far too small for others. For example, control schemes are often assessed by comparing the average operation post-commissioning to that before. Small errors in each of the averages will cause much larger errors in the assessed improvement. Chapter 7 provides a methodology for assessing the accuracy of any conclusion arrived at with the chosen sample size.

- While many engineers understand the principles of significance testing, it is commonly misapplied. Chapter 8 takes the reader through the subject from first principles, describing the problems in identifying outliers and properly explaining the impact of repeatability and reproducibility of measurements.

- In assessing process behaviour it is quite common for the engineer to simply calculate, using standard formulae, the mean and standard deviation of process data. Even if the data are normally distributed, plotting the distribution of the actual data against that assumed will often reveal a poor fit. A single data point, well away from the mean, will cause the standard deviation to be substantially overestimated. Excluding such points as outliers is very subjective and risks the wrong conclusion being drawn from the analysis. Curve fitting, using all the data, produces a much more reliable estimate of mean and standard deviation. There are a range of methods of doing this, described in Chapter 9.

- Engineers tend to judge whether a distribution fits the data well by superimposing the continuous distribution on the discontinuous histogram. Such comparison can be very unreliable. Chapter 6 describes the use of quantile–quantile plots, as a much more effective alternative that is simple to apply.

- The assumption that process data follows the normal (Gaussian) distribution has become the de facto standard used in the estimation of the benefits of improved control. While valid for many datasets, there are many examples where there is a much better choice of distribution. Choosing the wrong distribution can result in the benefit estimate being easily half or double the true value. This can lead to poor decisions about the scope of an improved control project or indeed about whether it should be progressed or not. Chapter 10 demonstrates that while the underlying process data may be normally distributed, derived data may not be. For example, the variation in distillation column feed composition, as source of disturbance, might follow a normal distribution, but the effect it has on product compositions will be highly asymmetrical. Chapter 12 describes a selection of the more well-known alternative distributions. All are tested with different sets of real process data so that the engineer can see in detail how they are applied and how to select the most appropriate. A much wider range is catalogued in Part 2.

- While process control is primarily applied to continuous processes, there are many examples where statistics can be applied to assess the probability of an undesirable event. This might be important during benefit estimation, where the improvement achievable by improved control is dependent on other factors – for example, the availability of feed stock or of a key piece of process equipment. Failure to take account of such events can result in benefits being overestimated. Event analysis can also be applied to performance monitoring. For example, it is

common to check an inferential property against the latest laboratory result. Estimating the probability of a detected failure being genuine helps reduce the occasions where control is unnecessarily disabled. Chapter 12 and Part 2 describe the wide range of statistical methods that are applicable to discrete events. Again, the description of each technique includes a worked example intended to both illustrate its use and inspire the engineer to identify new applications.

- One objective of control schemes is to prevent the process operating at conditions classed as extreme. Such conditions might range from a minor violation of a product specification to those causing a major hazard. Analysing their occurrence as part of the tail of distribution can be extremely unreliable. By definition the volume of data collected in this region will be only a small proportion of the data used to define the whole distribution. Chapter 13 describes techniques for accurately quantifying the probability of extreme process behaviour.

- Rarely used by control engineers, the hazard function described in Chapter 14 is simply derived from basic statistical functions. It can be beneficial in assessing the ongoing reliability of the whole control scheme or of individual items on which it depends. It can be a very effective technique to demonstrate the need to invest in process equipment, improved instrumentation, additional staff and training.

- Engineers often overlook that some process conditions have 'memory'. It is quite reasonable to characterise the variability of a product composition by examining the distribution of a daily laboratory result. However the same methodology should not be applied to the level of liquid in a product storage tank. If the level yesterday was very low, it is extremely unlikely to be high today. The analysis of data that follow such a time series is included in Chapter 15. The technique is equally applicable to sub-second collection frequencies where it can be used to detect control problems.

- Regression analysis is primarily used by process control engineers to build inferential properties. While sensibly performed with software, there are many pitfalls that arise from not fully understanding the techniques it uses. The belief that the Pearson R coefficient is a good measure of accuracy is responsible for a very large proportion of installed inferentials, on being commissioned, causing unnoticed degradation in quality control. Chapter 16 presents the whole subject in detail, allowing the engineer to develop more effective correlations and properly assess their impact on control performance.

- Engineers, perhaps unknowingly, apply time series analysis techniques as part of model identification for MPC (multivariable predictive control). Often part of a proprietary software product, the technique is not always transparent. Chapter 17 details how such analysis is performed and suggests other applications in modelling overall process behaviour.

- Process control engineers frequently have to work with inconsistent data. An inferential property will generate a different value from that recorded by an on-stream analyser which, in turn, will be different from the laboratory result. Mass balances, required by optimisation strategies, do not close because the sum of the product flows differs from the measured feed flow. Data reconciliation is a technique, described in Chapter 18, which not only reconciles such differences but also produces an estimate that is more reliable than any of the measurements. Further, it can be extended to help specify what additional instrumentation might be installed to improve the overall accuracy.

- Much neglected, because of the perception that the mathematics are too complex to be practical, Fourier transforms are invaluable in helping detect and diagnose less obvious control problems. Chapter 19 shows that the part of the theory that is applicable to the process industry is quite simple and its application well within the capabilities of a competent engineer.

To logically sequence the material in this book was a challenge. Many statistical techniques are extensions to or special cases of others. The intent was not to refer to a technique in one chapter unless it had been covered in a previous one. This proved impossible. Many of the routes through the subject are circular. I have attempted to enter such circles at a point that requires the least previous knowledge. Cross-references to other parts of the book should help the reader navigate through a subject as required.

A similar challenge arose from the sheer quantity of published statistical distributions. Chapter 12 includes a dozen, selected on the basis that they are well known, are particularly effective or offer the opportunity to demonstrate a particular technique. The remainder are catalogued as Part 2 – offering the opportunity for the engineer to identify a distribution that may be less well known but might prove effective for a particular dataset. Several well-known distributions are relegated to Part 2 simply because they are unlikely to be applicable to process data but merit an explanation as to why not. A few distributions, which have uses only tenuously related to process control, are included because I am frequently reminded not to underestimate the ingenuity of control engineers in identifying a previously unconsidered application.

As usual, I am tremendously indebted to my clients' control engineers. Their cooperation in us together applying published statistical methods to their processes has helped hugely in proving their benefit. Much of the material contained in this book is now included in our training courses. Without the feedback from our students, putting what we cover into practice, the refinements that have improved practicability would never have been considered.

Finally, I apologise for not properly crediting everyone that might recognise, as theirs, a statistical technique reproduced in this book. While starting with the best of intentions to do so, it proved impractical. Many different statistical distributions can readily be derived from each other. It is not always entirely clear who thought of what first, and there can be dozens of papers involved. I appreciate that academics want to be able to review published work in detail. Indeed, I suspect that the pure statistician might be quite critical of the way in which much of the material is presented. It lacks much of the mathematical detail they like to see, and there are many instances where I have modified and applied their techniques in ways of which they would not approve. However this book is primarily for practitioners who are generally happy just that a method works. A simple Internet search should provide more detailed background if required.

Myke King
Isle of Wight
June 2017

About the Author

Myke King is the founder and director of Whitehouse Consulting, an independent consulting organisation specialising in process control. He has over 40 years' experience working with over 100 clients in 40 countries. As part of his consulting activities Myke has developed training courses covering all aspects of process control. To date, around 2,000 delegates have attended these courses. He also lectures at several universities. To support his consulting activities he has developed a range of software to streamline the design of controllers and to simulate their use for learning exercises. Indeed, part of this software can be downloaded from the companion web site for this book.

Myke graduated from Cambridge University in the UK with a Master's degree in chemical engineering. His course included process control taught as part of both mechanical engineering and chemical engineering. At the time he understood neither. On graduating he joined, by chance, the process control section at Exxon's refinery at Fawley in the UK. Fortunately he quickly discovered that the practical application of process control bore little resemblance to the theory he had covered at university. He later became head of the process control section and then moved to operations department as a plant manager. This was followed by a short period running the IT section.

Myke left Exxon to co-found KBC Process Automation, a subsidiary of KBC Process Technology, later becoming its managing director. The company was sold to Honeywell where it became their European centre of excellence for process control. It was at this time Myke set up Whitehouse Consulting.

Myke is a Fellow of the Institute of Chemical Engineers in the UK.

Supplementary Material

To access supplementary materials for this book please use the download links shown below.

There you will find valuable material designed to enhance your learning, including:

- Excel spreadsheets containing data used in the examples. Download directly from www.whitehouse-consulting.com/examples.xlsx
- Executable file (www.whitehouse-consulting.com/statistician.exe) to help you perform your own data analysis

For Instructor use only:

Powerpoint slides of figures can be found at

http://booksupport.wiley.com

Please enter the book title, author name or ISBN to access this material.

Part 1

The Basics

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.

2

Application to Process Control

Perhaps more than any other engineering discipline, process control engineers make extensive use of statistical methods. Embedded in proprietary control design and monitoring software, the engineer may not even be aware of them. The purpose of this chapter is to draw attention to the relevance of statistics throughout all stages of implementation of improved controls – from estimation of the economic benefits, throughout the design phase, ongoing performance monitoring and fault diagnosis. Those that have read the author's first book *Process Control: A Practical Approach* will be aware of the detail behind all the examples and so most of this has been omitted here.

2.1 Benefit Estimation

The assumption that the variability or standard deviation (σ) is halved by the implementation of improved regulatory control has become a de facto standard in the process industry. It has no theoretical background; indeed it would be difficult to develop a value theoretically that is any more credible. It is accepted because post-improvement audits generally confirm that it has been achieved. But results can be misleading because the methodology is being applied, as we will show, without a full appreciation of the underlying statistics.

There are a variety of ways in which the benefit of reducing the standard deviation is commonly assessed. The *Same Percentage Rule*[1,2] is based on the principle that if a certain percentage of results already violate a specification, then after improving the regulatory control, it is acceptable that the percentage violation is the same. Halving the standard deviation permits the average giveaway to be halved.

$$\Delta\bar{x} = 0.5\left(x_{\text{target}} - \bar{x}\right) \tag{2.1}$$

This principle is illustrated in Figure 2.1. Using the example of diesel quality data that we will cover in Chapter 3, shown in Table A1.3, we can calculate the mean as 356.7°C and the standard deviation as 8.4°C. The black curve shows the assumed distribution. It shows that the probability of the product being on-grade, with a 95% distillation point less than 360°C, is 0.65.

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

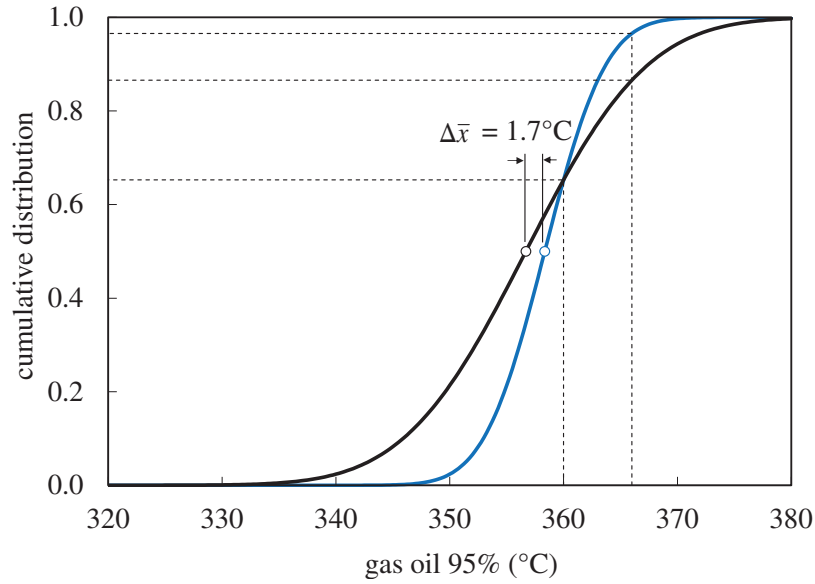


Figure 2.1 *Same percentage rule*

In other words, we expect 35% of the results to be off-grade. Halving the standard deviation, as shown by the coloured curve, would allow us to increase the mean while not affecting the probability of an off-grade result. From Equation (2.1), improved control would allow us to more closely approach the specification by 1.7°C.

We will show later that it is not sufficient to calculate mean and standard deviation from the data. Figure 2.2 again plots the assumed distribution but also shows, as points, the distribution

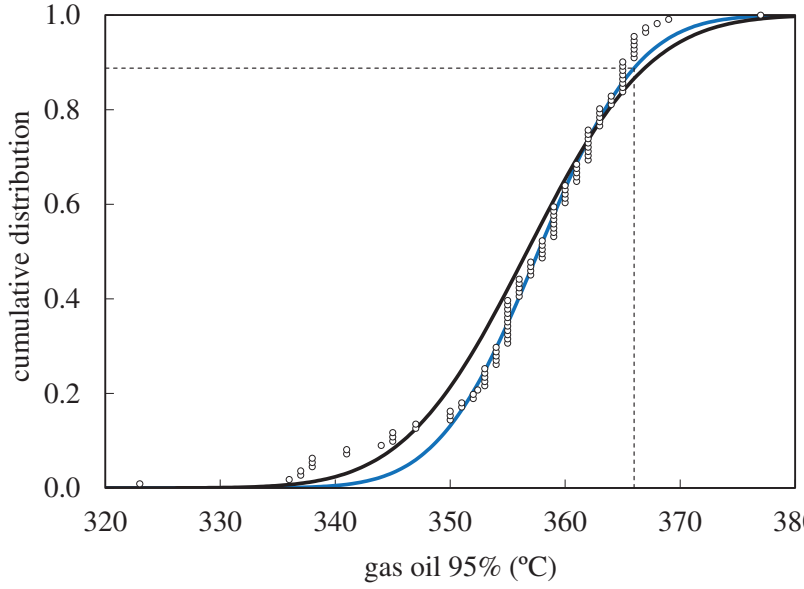


Figure 2.2 *Properly fitting a distribution*

of the actual data. The coloured curve is the result of properly fitting a normal distribution to these points, using the method we will cover in Chapter 9. This estimates the mean as 357.7°C and the standard deviation as 6.9°C. From Equation (2.1), the potential improvement is now 1.2°C. At around 30% less than the previous result, this represents a substantial reduction in the benefit achievable.

A second potential benefit of improved control is a reduction in the number of occasions the gasoil must be reprocessed because the 95% distillation point has exceeded 366°C. As Figure 2.1 shows, the fitted distribution would suggest that the probability of being within this limit is 0.888. This would suggest that, out of the 111 results, we would then expect the number of reprocessing events to be 12. In fact there were only five. It is clear from Figure 2.2 that the assumed distribution does not match the actual distribution well – particularly for the more extreme results. The problem lies now with the choice of distribution. From the large number of alternative distributions it is likely that a better one could be chosen. Or, even better, we might adopt a discrete distribution suited to estimation of the probability of events. We could also apply an extreme value analytical technique. Both these methods we cover in Chapter 13.

2.2 Inferential Properties

A substantial part of a control engineer's role is the development and maintenance of inferential property calculations. Despite the technology being well established, not properly assessing their performance is the single biggest cause of benefits not being fully captured. Indeed, there are many examples where process profitability would be improved by their removal.

Most inferentials are developed through regression of previously collected process data. Doing so employs a wide range of statistical techniques. Regression analysis helps the engineer identify the most accurate calculation but not necessarily the most practical. The engineer has to apply other techniques to assess the trade-off between complexity and accuracy.

While there are 'first-principle' inferentials, developed without applying statistical methods, once commissioned both types need to be monitored to ensure the accuracy is maintained. If an accuracy problem arises, then the engineer has to be able to assess whether it can be safely ignored as a transient problem, whether it needs a routine update to its bias term or whether a complete redesign is necessary. While there is no replacement for relying on the judgement of a skilled engineer, statistics play a major role in supporting this decision.

2.3 Controller Performance Monitoring

Perhaps the most recent developments in the process control industry are process control performance monitoring applications. Vendors of MPC packages have long offered these as part of a suite of software that supports their controllers. But more recently the focus has been on monitoring basic PID control, where the intent is to diagnose problems with the controller itself or its associated instrumentation. These products employ a wide range of statistical methods to generate a wide range of performance parameters, many of which are perhaps not fully understood by the engineer.

2.4 Event Analysis

Event analysis is perhaps one of the larger opportunities yet to be fully exploited by process control engineers. For example, they will routinely monitor the performance of advanced control – usually reporting a simple service factor. Usually this is the time that the controller is in service expressed as a fraction of the time that it should have been in service. While valuable as reporting tool, it has limitations in terms of helping improve service factor. An advanced control being taken out of service is an example of an event. Understanding the frequency of such events, particularly if linked to cause, can help greatly in reducing their frequency.

Control engineers often have to respond to instrument failures. In the event of one, a control scheme may have to be temporarily disabled or, in more extended cases, be modified so that it can operate in some reduced capacity until the fault is rectified. Analysis of the frequency of such events, and the time taken to resolve them, can help justify a permanent solution to a recurring problem or help direct management to resolve a more general issue.

Inferential properties are generally monitored against an independent measurement, such as that from an on-stream analyser or the laboratory. Some discrepancy is inevitable and so the engineer will have previously identified how large it must be to prompt corrective action. Violating this limit is an event. Understanding the statistics of such events can help considerably in deciding whether the fault is real or the result of some circumstance that needs no attention.

On most sites, at least part of the process requires some form of sequential, rather than continuous, operation. In an oil refinery, products such as gasoline and diesel are batch blended using components produced by upstream continuous processes. In the polymers industry plants run continuously but switch between grades. Downstream processing, such as extrusion, has to be scheduled around extruder availability, customer demand and product inventory. Other industries, such as pharmaceuticals, are almost exclusively batch processes. While most advanced control techniques are not applicable to batch processes, there is often the opportunity to improve profitability by improved scheduling. Understanding the statistical behaviour of events such as equipment availability, feedstock availability and even the weather can be crucial in optimising the schedule.

Many control engineers become involved in alarm studies, often following the guidelines[3] published by Engineering Equipment and Materials Users' Association. These recommend the following upper limits per operator console:

- No more than 10 *standing alarms*, i.e. alarms which have been acknowledged
- No more than 10 *background alarms* per hour, i.e. alarms for information purposes that may not require urgent attention
- No more than 10 alarms in the first 10 minutes after a major process problem develops

There are also alarm management systems available that can be particularly useful in identifying repeating *nuisance* and *long-standing* alarms. What is less common is examination of the probability of a number of alarms occurring. For example, if all major process problemshave previously met the criterion of not more than 10 alarms, but then one causes 11, should this prompt a review? If not, how many alarms would be required to initiate one?

2.5 Time Series Analysis

Often overlooked by control engineers, feed and product storage limitations can have a significant impact on the benefits captured by improved control. Capacity utilisation is often the major source of benefits. However, if there are periods when there is insufficient feed in storage or insufficient capacity to store products, these benefits would not be captured. Indeed, it may be preferable to operate the process at a lower steady feed rate rather than have the advanced control continuously adjust it. There is little point in maximising feed rate today if there will be a feed shortage tomorrow.

Modelling the behaviour of storage systems requires a different approach to modelling process behaviour. If the level in a storage tank was high yesterday, it is very unlikely to be low today. Such levels are *autoregressive*, i.e. the current level (L_n) is a function of previous levels.

$$L_n = a_0 + a_1 L_{n-1} + a_2 L_{n-2} + \ldots \tag{2.2}$$

The level is following a *time series*. It is not sufficient to quantify the variation in level in terms of its mean and standard deviation. We need also to take account of the sequence of levels.

Time series analysis is also applicable to the process unit. Key to effective control of any process is understanding the process dynamics. *Model identification* determines the correlation between the current process value (PV_n), previous process values (PV_{n-1}, etc.) and previous values of the manipulated variable (MV) delayed by the process deadtime (θ). If ts is the data collection interval, the *autoregressive with exogenous input* (*ARX*) model for a single MV has the form

$$PV_n = a_0 + a_1 PV_{n-1} + a_2 PV_{n-2} + \ldots + b_1 MV_{n-\theta/ts} + b_2 MV_{n-1-\theta/ts} \ldots \tag{2.3}$$

For a *first order* process, this model will include only one or two historical values. Simple formulae can then be applied to convert the derived coefficients to the more traditional parametric model based on process gain, deadtime and lag. These values would commonly be used to develop tuning for basic PID controllers and for advanced regulatory control (ARC) techniques. Higher order models can be developed by increasing the number of historical values and these models form the basis of some proprietary MPC packages. Other types of MPC use the time series model directly.

There is a wide range of proprietary model identification software products. Control engineers apply them without perhaps fully understanding how they work. Primarily they use *regression analysis* but several other statistical techniques are required. For example, increasing the number of historical values will always result in a model that is mathematically more accurate. Doing so, however, will increasingly model the noise in the measurements and reduce the robustness of the model. The packages include statistical techniques that select the optimum model length. We also need to assess the reliability of the model. For example, if the process disturbances are small compared to measurement noise or if the process is highly nonlinear, there may be little confidence that the identified model is reliable. Again the package will include some statistical technique to warn the user of this. Similarly statistical methods might also be used to remove any suspect data before model identification begins.

3

Process Examples

Real process data has been used throughout to demonstrate how the techniques documented can be applied (or not). This chapter simply describes the data and how it might be used. Where practical, data are included as tables in Appendix 1 so that the reader can reproduce the calculations performed. All of the larger datasets are available for download.

The author's experience has been gained primarily in the oil, gas and petrochemical industries; therefore much of the data used come from these. The reader, if from another industry, should not be put off by this. The processes involved are relatively simple and are explained here. Nor should the choice of data create the impression that the statistical techniques covered are specific to these industries. They are not; the reader should have no problem applying them to any set of process measurements.

3.1 Debutaniser

The debutaniser column separates C_{4-} material from naphtha, sending it to the de-ethaniser. Data collected comprises 5,000 hourly measurements of reflux (R) and distillate (D) flows. Of interest is, if basic process measurements follow a particular distribution, what distribution would a derived measurement follow? In Chapter 10 the flows are used to derive the reflux ratio (R/D) to demonstrate how the ratio of two measurements might be distributed.

3.2 De-ethaniser

The overhead product is a gas and is fed to the site's fuel gas system, along with many other sources. Disturbances to the producers cause changes in fuel gas composition – particularly affecting its molecular weight and heating value. We cover this later in this chapter.

The bottoms product is mixed LPG (propane plus butane) and it routed to the splitter. The C_2 content of finished propane is determined by the operation of the de-ethaniser. We cover, later in this chapter, the impact this has on propane cargoes.

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

3.3 LPG Splitter

The LPG splitter produces sales grade propane and butane as the overheads and bottoms products respectively. Like the debutaniser, data collected includes 5,000 hourly measurements of reflux and distillate flows. These values are used, along with those from the debutaniser, to explore the distribution of the derived reflux ratio.

The reflux flow is normally manipulated by the composition control strategy. There are columns where it would be manipulated by the reflux drum level controller. In either case the reflux will be changed in response to almost every disturbance to the column. Of concern on this column are those occasions where reflux exceeds certain flow rates. Above 65 m^3/hr the column can flood. A flow above 70 m^3/hr can cause a pump alarm. Above 85 m^3/hr, a trip shuts down the process.

Figure 3.1 shows the variation in reflux over 5,000 hours. Figure 3.2 shows the distribution of reflux flows. The shaded area gives the probability that the reflux will exceed 65 m^3/hr. We will show, in Chapter 5, how this value is quantified for the normal distribution and, in subsequent chapters, how to apply different distributions.

Alternatively, a high reflux can be classed as an event. Figure 3.1 shows 393 occasions when the flow exceeded 65 m^3/hr and 129 when it exceeded 70 m^3/hr. The distribution can then be based on the number of events that occur in a defined time. Figure 3.3 shows the distribution of the number of events that occur per day. For example, it shows that the observed probability of the reflux not exceeding 70 m^3/hr in a 24 hour period (i.e. 0 events per day) is around 0.56. Similarly the most likely number of violations, of the 65 m^3/hr limit, is two per day, occurring on approximately 29% of the days. We will use this behaviour, in Chapter 12 and Part 2, to show how many of the discrete distributions can be applied. Another approach is to analyse the variation of time between high reflux events. Figure 3.4 shows the observed distribution of the interval between exceeding 65 m^3/hr. For example, most likely is an interval of one hour – occurring

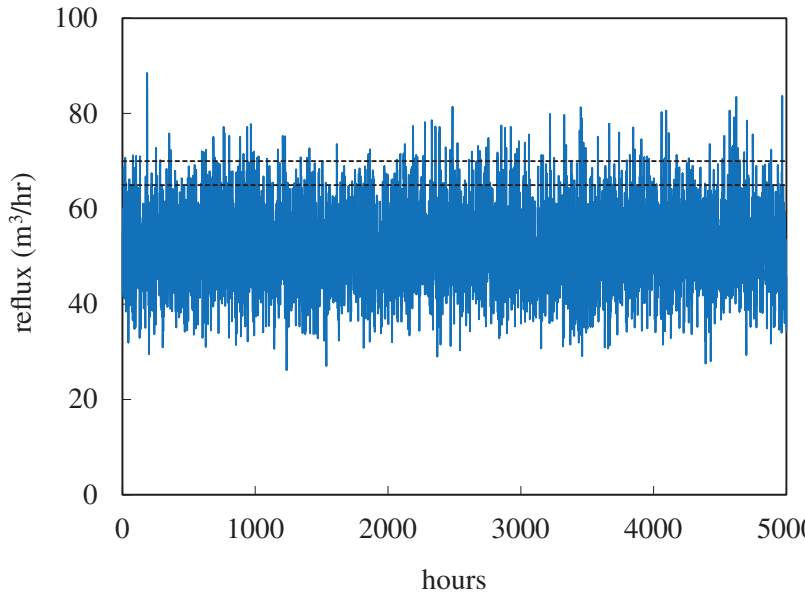


Figure 3.1 Variation in LPG splitter reflux

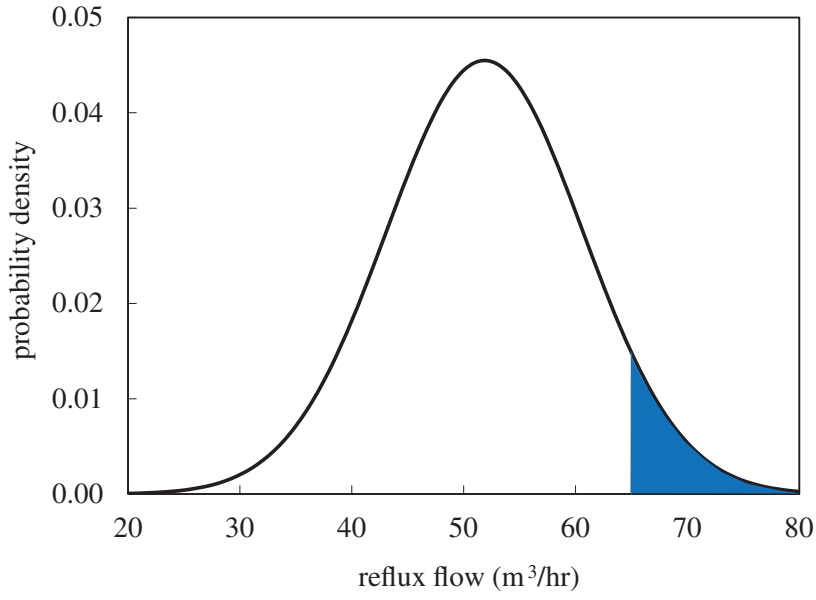


Figure 3.2 *Distribution of reflux flow*

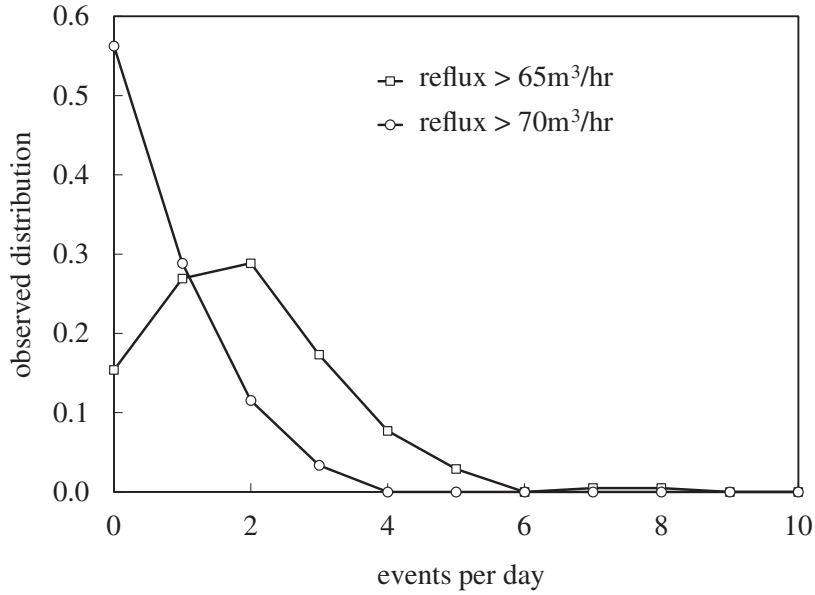


Figure 3.3 *Distribution of high reflux events*

on about 9% of the occasions. In this form, continuous distributions can then be applied to the data.

Table A1.1 shows the C_4 content of propane, not the finished product but sampled from the rundown to storage. It includes one year of daily laboratory results, also shown in Figure 3.5. Of interest is the potential improvement to composition control that will increase the C_4 content

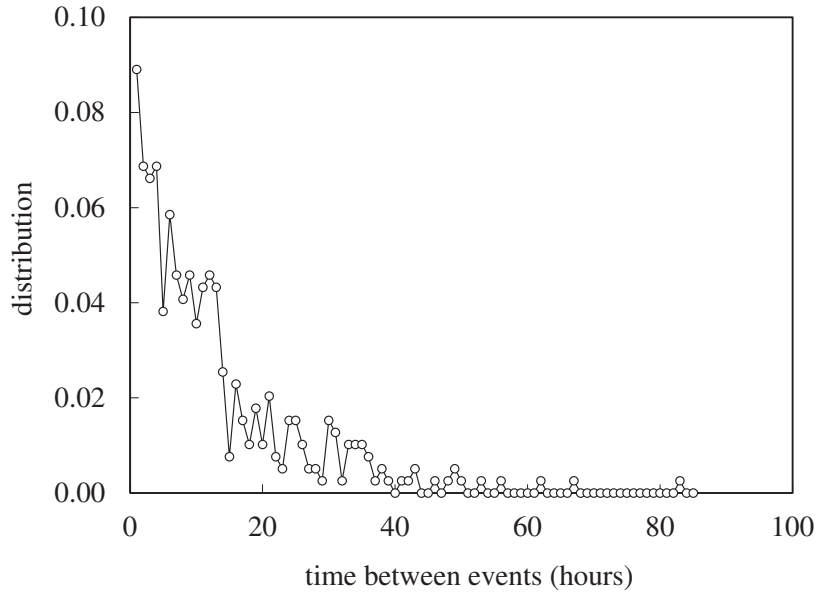


Figure 3.4 *Distribution of intervals between reflux events*

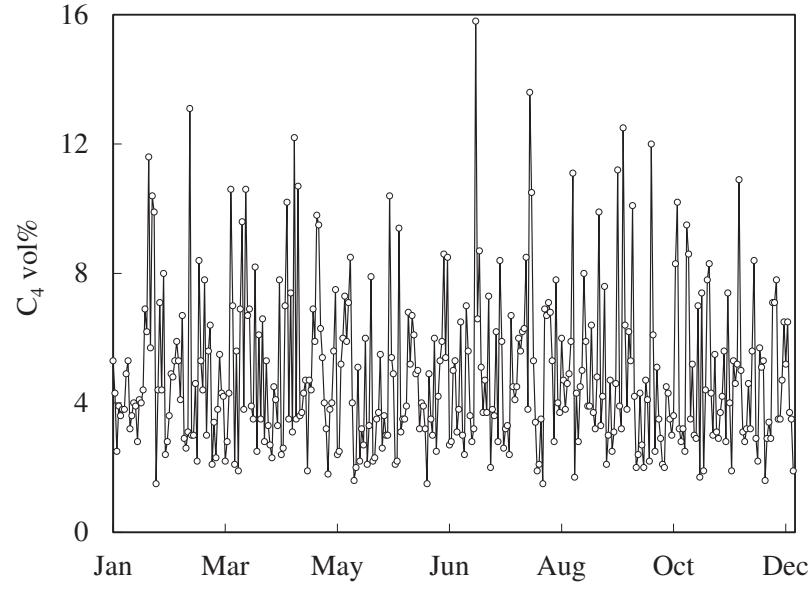


Figure 3.5 *Laboratory results for C_4 content of propane*

closer to the specification of 10 vol%. To determine this we need an accurate measure of the current average content and its variation. The key issue is choosing the best distribution. As Figure 3.6 shows, the best fit normal distribution does not match well the highly skewed data. Indeed, it shows about a 4% probability that the C_4 content is negative. We will clearly need to select a better form of distribution from the many available.

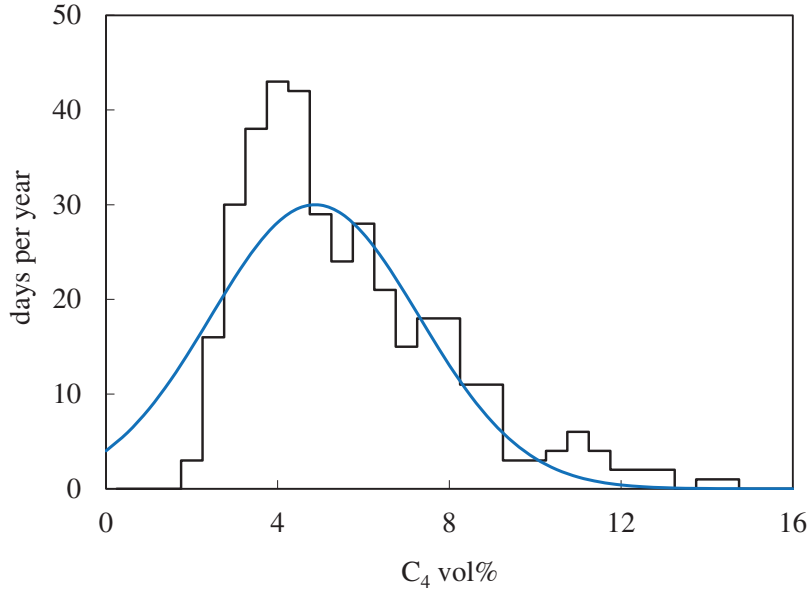


Figure 3.6 *Skewed distribution of C_4 results*

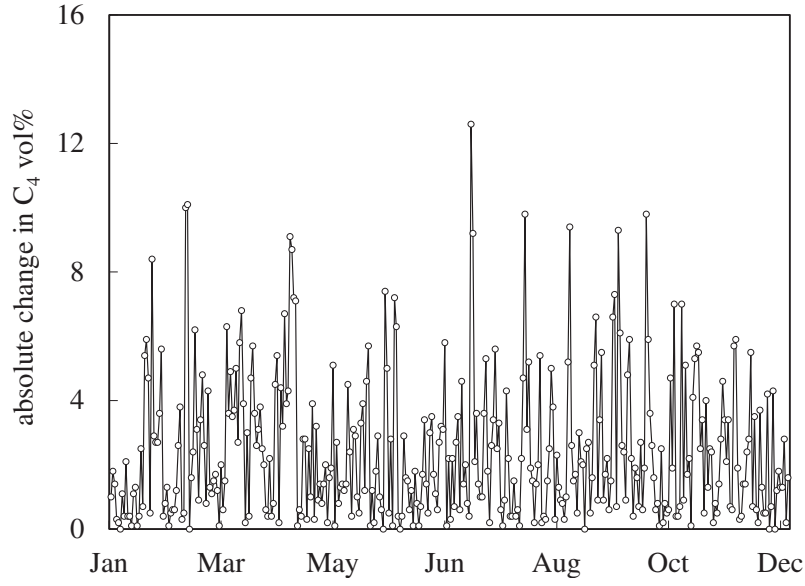


Figure 3.7 *Changes in C_4 content*

Also of concern are the occasional very large changes to the C_4 content, as shown by Figure 3.7, since these can cause the product, already in the rundown sphere, to be put off-grade. We will show how some distributions can be used to assess the impact that improved control might have on the frequency of such disturbances.

There is an analyser and an inferential property installed on the bottoms product measuring the C_3 content of butane. Figure 3.8 shows data collected every 30 minutes over 24 days, i.e.

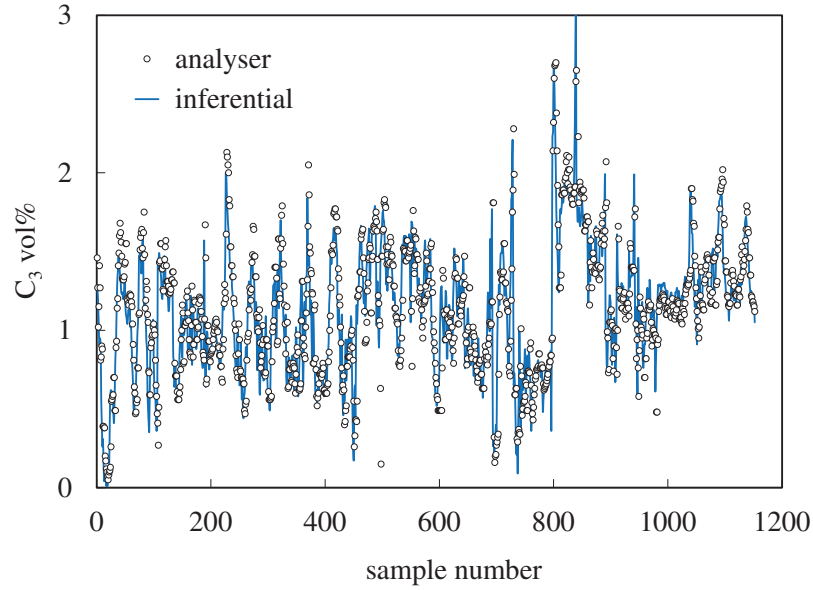


Figure 3.8 *Comparison between analyser and inferential*

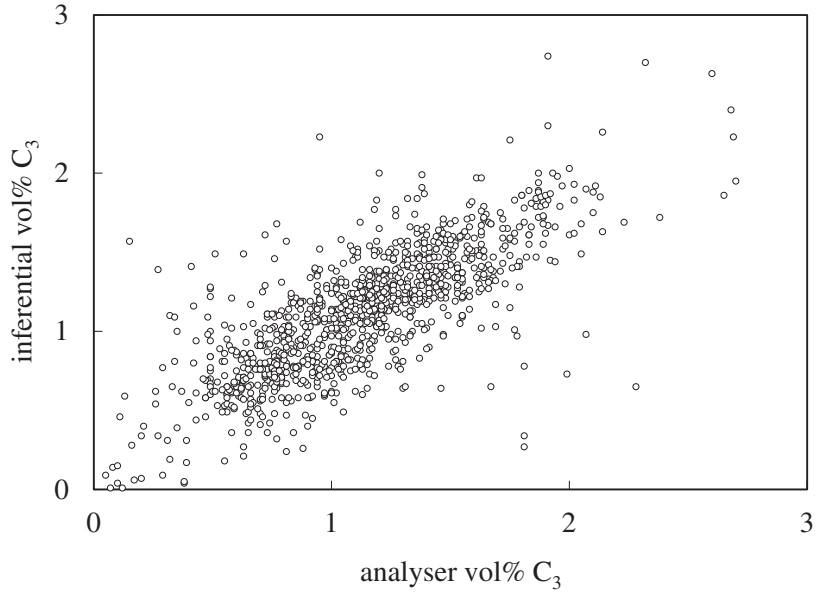


Figure 3.9 *Scatter plot of inferential against analyser*

1,152 samples. Such line plots are deceptive in that they present the inferential as more accurate than it truly is. Figure 3.9 plots the same data as a scatter diagram showing that, for example, if the analyser is recording 1 vol%, the inferential can be in error by ±0.5 vol%. Further, there is tendency to assume that such errors follow the normal distribution. Figure 3.10 shows the best fit normal distribution. The actual frequency of small errors is around double that suggested by

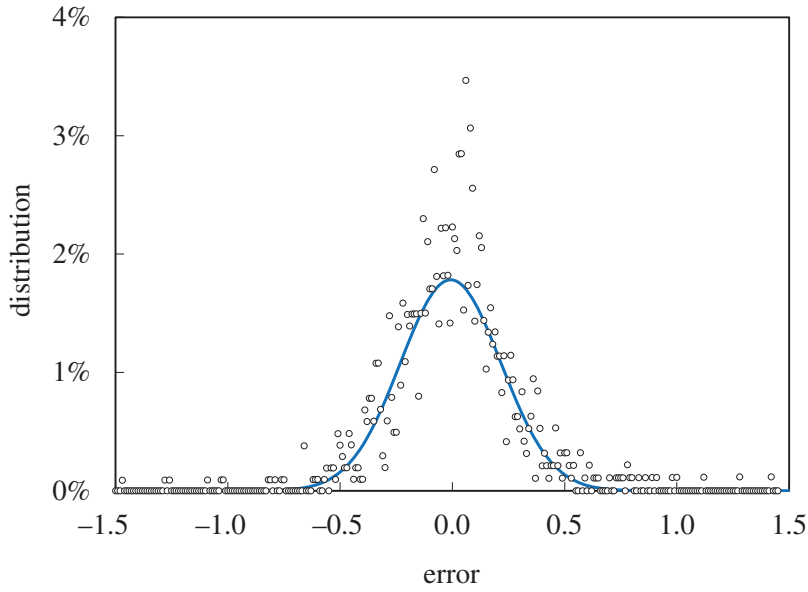


Figure 3.10 *Actual distribution very different from normal*

the normal distribution. We will look at how other distributions are better suited to analysing this problem.

3.4 Propane Cargoes

Propane from the LPG splitter is routed to one of two spheres. Once a sphere is full, production is switched to the other sphere. Cargoes are shipped from the filled sphere once the laboratory has completed a full analysis for the certificate of quality. The maximum C_2 content permitted by the propane product specification is 5 vol%. Table A1.2 includes the results, taken from the certificates of quality, for 100 cargoes. While primarily used to illustrate, in Chapter 6, methods of presenting data, it is also used to draw attention to the difference between analysing finished product data as opposed to data collected from the product leaving the process.

3.5 Diesel Quality

A property commonly measured for oil products is the *distillation point*. Although its precise definition is more complex, in principle it is the temperature at which a certain percentage of the product evaporates. Gasoil, for example, is a component used in producing diesel and might have a specification that 95% must evaporate at a temperature below 360°C. There is an economic incentive to get as close as possible to the specification.

Table A1.3 shows the results of 111 daily laboratory samples taken from the gasoil rundown. The product is routed to a storage tank which is well-mixed before being used in a blend. A certain amount of off-grade production is permissible provided it is balanced by product in giveaway, so that the filled tank is within specification. Indeed, as Figure 3.11 shows, 40 of the results violate the specification. But the simple average, represented by the coloured

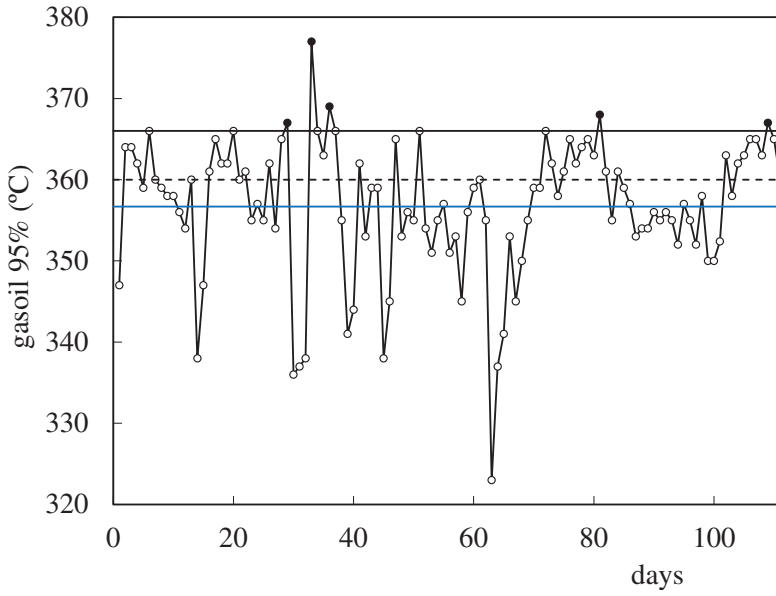


Figure 3.11 *Laboratory results for gasoil 95% distillation point*

line, shows giveaway of 3.3°C. Improving composition control would reduce the variation and so allow closer approach to the specification and an increase in product yield. We will use these data to show how to more properly estimate the average operation and its variation.

Of course there are industries where no amount of off-grade production is permitted. Most notable are the paper and metals industries where the off-grade material cannot be blended with that in giveaway. Our example has a similar situation. At a distillation point above 366°C undesirable components can be present in the product that cannot be sufficiently diluted by mixing in the tank. Any material this far off-grade must be reprocessed. The figures highlighted in Table A1.3 and Figure 3.11 show the five occasions when this occurred. Improved control would reduce the number of occasions and so reduce reprocessing costs. We will use these data to explore the use of discrete distributions that might be used to determine the savings.

3.6 Fuel Gas Heating Value

In common with many sites all fuel gas from producers, such as the de-ethaniser, is routed to a header from which all consumers take their supply. A disturbance on any producer causes a change in NHV (net heating value) that then upsets a large number of fired heaters and boilers. Some consumers of fuel gas are also producers; a disturbance to the gas supplied to these units propagates through to further disturb the header NHV.

The data, included as Table A1.4, comprises laboratory results collected daily for a period of six months. It was collected to identify the best approach to handling variations to reduce the process disturbances. There are several solutions. One might be to install densitometers, use these to infer NHV and effectively convert the flow controllers on each consumer to duty controllers. Another might be to switch from conventional orifice plate type flowmeters to coriolis types on the principle that heating value measured on a weight basis varies much less than that

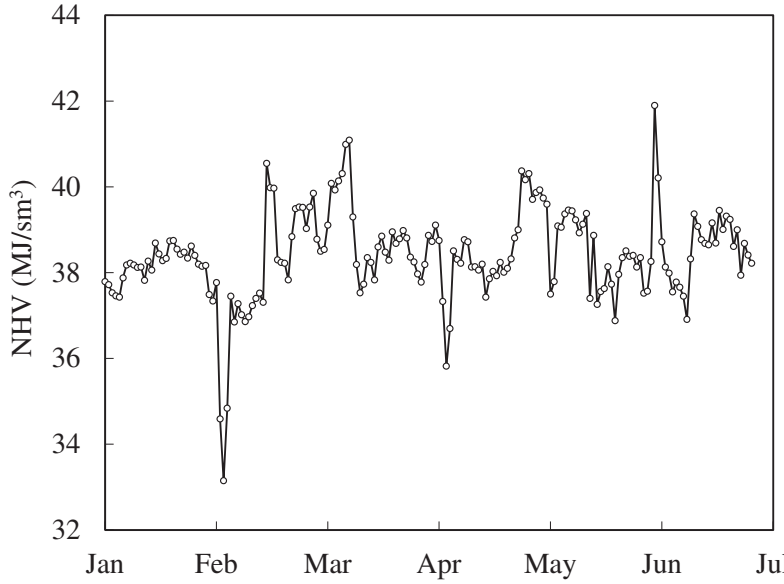


Figure 3.12 Variation in site's fuel gas heating value

measured on a volume basis. Understanding the variation in NHV permits each solution to be assessed in terms of the achievable reduction in disturbances.

Figure 3.12 shows the variation of the NHV of fuel gas routed to a fired heater. The disturbances (x) are determined from Table A1.4 as

$$x_n = \frac{NHV_n - NHV_{n-1}}{NHV_{n-1}} \times 100 \quad n > 1 \tag{3.1}$$

These disturbances are plotted as Figure 3.13. Figure 3.14 shows that the best fit normal distribution is unsuitable since it has tails much fatter than the distribution of the data. The data will therefore be used to help assess the suitability of alternative distributions, some of which do not accommodate negative values. For this type of data, where it would be reasonable to assume that positive and negative disturbances are equally likely, one approach is to fit to the absolute values of the disturbances.

Table A1.5 comprises analyses of 39 of the samples of fuel gas showing the breakdown by component. These we will use to illustrate how a multivariate distribution might be applied.

3.7 Stock Level

While the control engineer may feel that there is little application of process control to product storage, understanding its behaviour can be crucial to properly estimating benefits of improved control and in assessing what changes might be necessary as a result of a control improvement. For example, the final product from many sites is the result of blending components. Properly controlling such blending can substantially improve profitability but, in estimating the benefits of a potential improvement we need to assess the availability not only of the blend components but also available capacity for finished product. There are many projects that have been justified

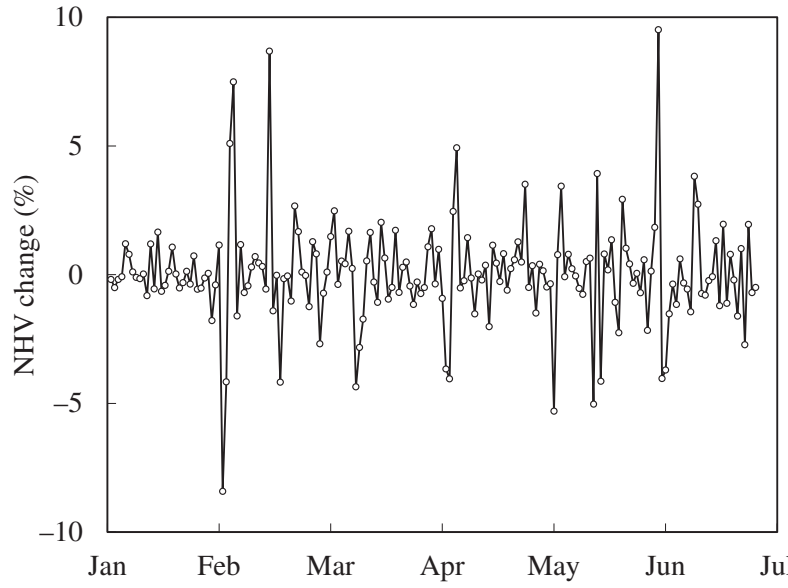


Figure 3.13 Disturbances to heating value

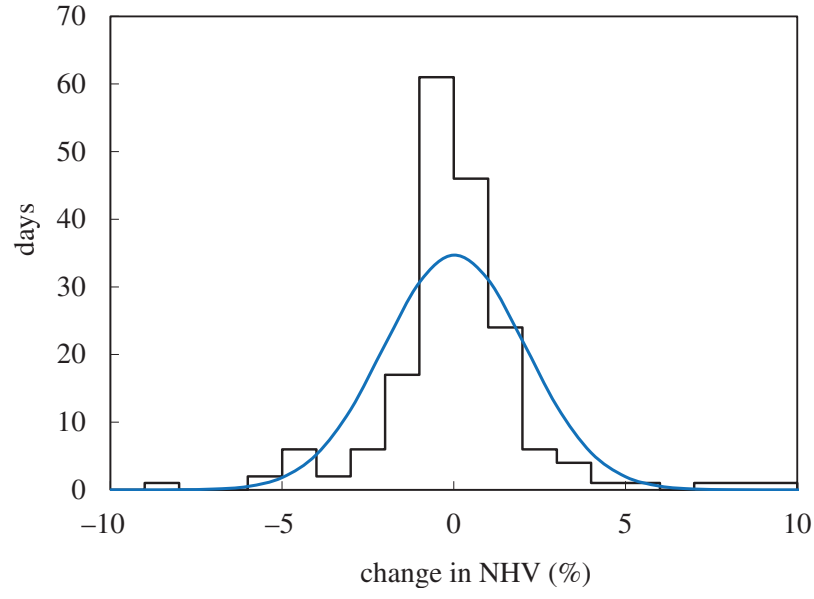


Figure 3.14 Small tails compared to normal distribution

on increasing production rates only to find that this cannot be fully accommodated by the storage facilities.

The data, included in Table A1.6, is the daily stock level of a key component collected over a seven month period or 220 days. The variation is also shown in Figure 3.15. Figure 3.16 shows

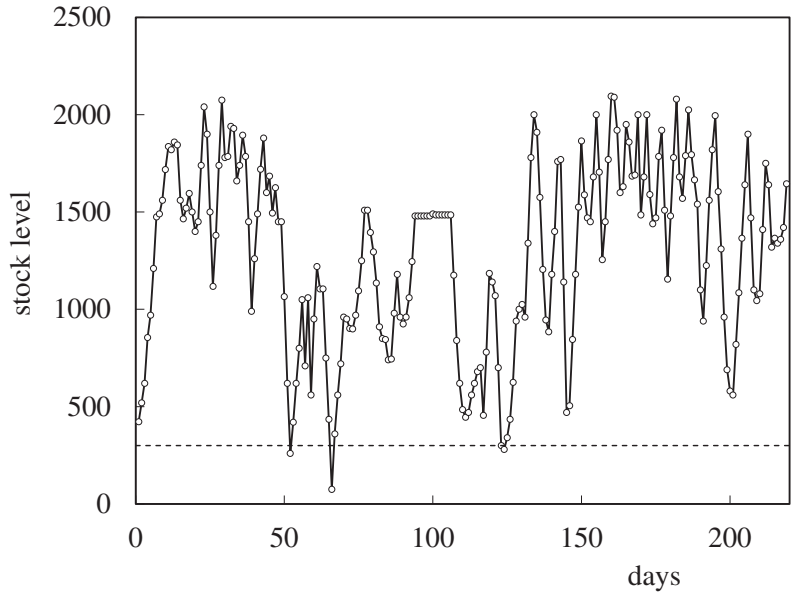


Figure 3.15 Variation in stock level

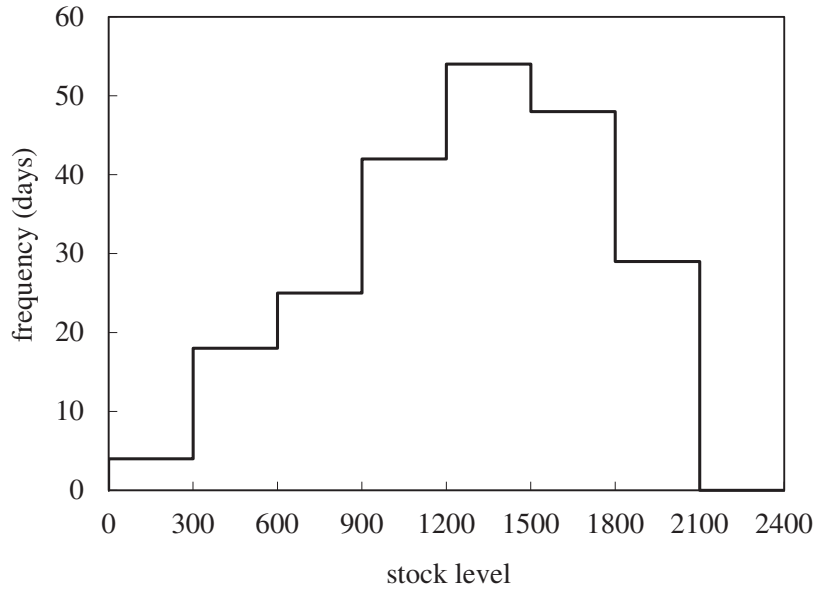


Figure 3.16 Skewed distribution of stock levels

the non-symmetrical distribution of the data. The main concern is those occasions when the inventory drops to 300, below which a blend cannot be started. There are three occasions, highlighted as bold in Table A1.6, when this occurred. The data will be used to demonstrate how discrete distributions might estimate the probability of such an occurrence in the future. We will also apply a time series technique to predict the future variation in level.

3.8 Batch Blending

Batch blending is common to many industries, including quite large continuous processes such as oil refineries. In this particular example, a batch is produced by blending components and then taking a laboratory sample of the product. The key specification is 100; if the laboratory result is less than this, a corrective trim blend is added to the product and it is then retested. Each blend takes a day, including the laboratory analysis of its quality. This is repeated as necessary until the product is on-grade. 100 m^3 is then routed to a downstream process and then to sales. The material remaining in the blend vessel forms part of the next batch.

There is large incentive to improve the control and so reduce the number of trim blends. This would permit an increase in production and reduction in storage requirements. Table A1.7 and Figure 3.17 show the intermediate and finished laboratory results for the 78 blends resulting in 44 finished batches. This example is used to explore the use of both continuous and discrete distributions in assessing the improvement that might arise from improved control. We will also show how to use some distributions to explore what change in storage facilities might be required if production is increased.

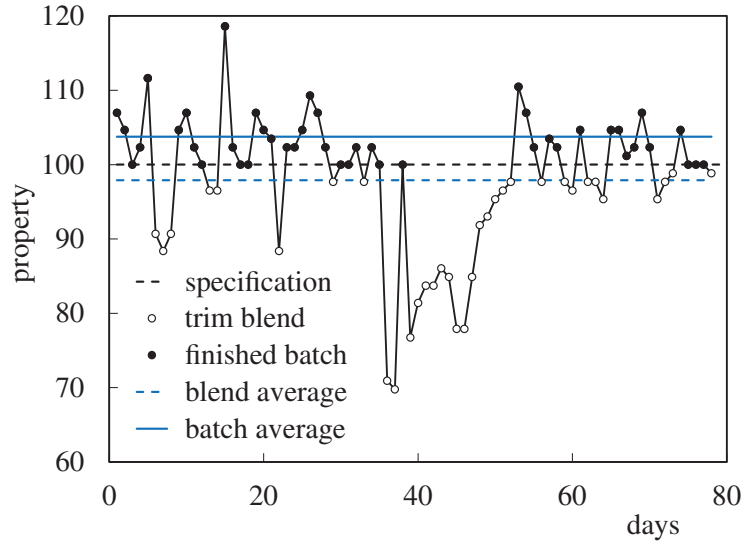


Figure 3.17 *Variation in blend property*

4

Characteristics of Data

4.1 Data Types

Data fall into one of three types:

Dichotomous data can have one of only two values. In the process industry examples might include pass/fail status of a check against a product specification. Similarly it might be the pass/fail status from validating a measurement from an on-stream analyser or inferential. Such data can be averaged by assigning a numerical value. For example, we might assign 1 if a MPC application is in use and 0 if not. Averaging the data would then give the service factor of the application.

Nominal data have two or more categories but cannot be ranked sensibly. For example, the oil industry produces multiple grades of many products. Specifications can vary by application, by season and by latitude. For example, a data point might have the value 'summer grade European regular gasoline'. Only limited mathematical manipulation is possible with such data. For example, it would be possible to determine what percentage of cargoes fell into this category.

Cardinal data have two or more categories that can be ranked. Most process data fall into this category. Within this type, data can be *continuous* or *discrete*. Process measurements might generally be considered as continuous measurements. Strictly a DCS can only generate discrete values although, for statistical purposes, the resolution is usually such that they can be treated as continuous. Laboratory results are usually discrete values. This arises because testing standards, for example, those published by the ASTM, specify the resolution to which they should be reported. This is based on the accuracy of the test. For example, the flash point of products like jet fuel is quoted to the nearest 0.5°C, whereas the cloud point of diesel is quoted to the nearest 3°C. Another common example of a discrete variable is the number of events that occur in a fixed time. In principle, the reciprocal of this, which is the elapsed time between events, is a continuous variable. However, real-time databases historise data at a fixed interval, for example one minute, and so even time can then be treated as a discrete variable.

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

4.2 Memory

Data can arise from a process that has *memory*. This occurs when the current measurement is in any way dependent on the preceding measurement. For example, we might wish to assess the probability of violating the maximum permitted inventory in a storage tank. It is extremely unlikely that the level in the tank will be very low today if it was very high yesterday. Today's inventory will be largely influenced by what it was yesterday.

The same might apply to assessing the likelihood of equipment failure. As equipment ages it might become more prone to failure. The time to failure is no longer an independent variable; it becomes shorter as the length of memory increases.

Process events, such as alarms, can also show memory. The condition that activates the alarm might also trigger several others. Analysis of alarm frequency would show that, rather than follow a random pattern, they tend to occur in batches. In other words the likelihood of an alarm increases if one has just occurred.

Most process data do not have memory. For example, product composition can change quite rapidly on most processes. While the composition now will be closely related to what it was a few seconds ago, it will show little correlation with what it was an hour ago. If composition data were collected as daily laboratory measurements, the process will almost certainly appear *memoryless* or *forgetful*. However measuring a property that changes slowly over time, such as catalyst activity, will show memory.

4.3 Use of Historical Data

We have seen that the process control engineer requires historical process data for a wide range of applications. Primarily these fall into two categories – assessing current performance and predicting future performance.

There are three basic methods of using the data for prediction:

- The simplest is to assume that future operation will be identical to the past. Process data are used directly. For example, in studying how a new control scheme might react to changes in feed composition, it is assumed that it will have to deal with exactly the same pattern of changes as it did in the past.

- The second method is to analyse historical data to identify parameters that accurately describe the distribution of the data as a *probability density function* (*PDF*) or *cumulative density function* (*CDF*). This distribution is then used in assessing future process behaviour. This is perhaps the most common method and a large part of this book presents these density functions in detail.

- The third approach is *Monte Carlo simulation*. This uses the derived distribution to generate a very large quantity of data that have the same statistics as the historical data. The synthesised data is then used to study the likely behaviour of a process in the future. For example, it is commonly used in simulating imports to and exports from product storage to determine what storage capacity is required. Provided the simulated imports and exports have the same statistical distribution as the real situation then the *law of large numbers* tells us the average of the results obtained from a large enough number of trials should be close to the real result.

Key to the success of the latter two methods is accurately defining the shape of the distribution of historical data.

4.4 Central Value

A dataset requires two key parameters to characterise its properties. Firstly, data generally show *central tendency* in that they are clustered around some central value. Secondly, as we shall see in the next section, a parameter is needed to describe how the data is dispersed around the central value. The most commonly used measure of the central value is the *mean* – more colloquially called the *average*. There are many versions of the mean. There are also several alternative measures of the central value. Here we define those commonly defined and identify which might be of value to the control engineer.

The *arithmetic* mean of the population (μ) of a set of N values of x is defined as

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \tag{4.1}$$

We will generally work with samples of the population. A sample containing n values will have the mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{4.2}$$

For example, as part of benefits study, we might examine the average giveaway against the maximum amount of C_4 permitted in propane product. If propane attracts a higher price than butane, we would want to maximise the C_4 content. We would normally take a large number of results to calculate the mean but, as an illustration, let us assume we have only three results of 3.9, 4.7 and 4.2. From Equation (4.2) we calculate the mean as 4.27. If the maximum permitted content is 5, the average giveaway is 0.73. Knowing the annual production of propane, we could use this to determine how much additional C_4 could be sold at the propane price rather than at the lower butane price.

However we should more properly use the *weighted mean*. Imagine that the three results were collected for three cargoes, respectively sized 75, 25 and 50 km^3. If w is the *weighting factor* (in this case the cargo size) then the mean butane content is

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{4.3}$$

The true mean C_4 content is therefore 4.13 and the giveaway 0.87. Calculating how much more C_4 could be included in propane would, in this example, give a result some 19% higher than that based on the simple mean. Equation (4.3) is effectively the total C_4 contained in all the cargoes divided by the total propane shipped. The additional C_4 that could have been included in propane is therefore 1.3 km^3 – given by

$$\frac{5\sum_{i=1}^{n} w_i - \sum_{i=1}^{n} w_i x_i}{100} \tag{4.4}$$

We might to keep track of the mean C_4 content over an extending period. Imagine that the three results are the first in a calendar year and we want to update the mean as additional cargoes are shipped to generate a year-to-date (YTD) giveaway analysis. We could of course recalculate

the mean from all the available results. Alternatively we can simply update the previously determined mean. In the case of the simple mean, the calculation would be

$$\bar{x}_{n+1} = \frac{n\bar{x}_n + x_{n+1}}{n+1} = \bar{x}_n + \frac{x_{n+1} - \bar{x}_n}{n+1} \tag{4.5}$$

For example, the fourth cargo of the year contains 3.7 vol% C_4. The year-to-date mean then becomes

$$\bar{x}_{n+1} = 4.27 + \frac{3.7 - 4.27}{3+1} = 4.13 \tag{4.6}$$

Note that this is different from a *rolling average*, in which the oldest result is removed when a new one is added. If is the number of values in the average (m) is 3, then

$$\bar{x}_{n+1} = \bar{x}_n + \frac{x_{n+1} - x_{n-m+1}}{m} = 4.27 + \frac{3.7 - 3.9}{3} = 4.20 \tag{4.7}$$

While not normally used as part of improved control studies, the rolling average can be applied as a means of filtering out some of the random behaviour of the process and measurement. Indeed, in some countries, finished product specifications will permit wider variation in the property of a single cargo, provided the rolling average is within the true specification.

To update a weighted average

$$\bar{x}_{n+1} = \frac{\bar{x}_n \sum_{i=1}^{n} w_i + w_{n+1}x_{n+1}}{\sum_{i=1}^{n} w_i + w_{n+1}} = \bar{x}_n + \frac{w_{n+1}(x_{n+1} - \bar{x}_n)}{\sum_{i=1}^{n} w_i + w_{n+1}} \tag{4.8}$$

So, if the fourth cargo was 60 km^3

$$\bar{x}_{n+1} = 4.13 + \frac{60(3.7 - 4.13)}{150 + 60} = 4.01 \tag{4.9}$$

Table 4.1 shows the result of applying the calculations described as additional cargoes are produced.

In addition to the arithmetic mean there is the *harmonic mean*, defined as

$$\bar{x}_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \tag{4.10}$$

The *weighted harmonic mean* is

$$\bar{x}_h = \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} \frac{w_i}{x_i}} \tag{4.11}$$

Using heavy fuel oil as an example, its maximum permitted density is 0.991. Giveaway is undesirable because density is reduced by adding diluent, such as gasoil, that would otherwise be sold at a price higher than heavy fuel oil. Consider three cargoes sized 80, 120 and 100 kt with densities of 0.9480, 0.9880 and 0.9740. The weighted average, if calculated from Equation (4.3), would be 0.9727. However, density does not blend linearly on a weight basis.

Table 4.1 *Averaging C_4 content of propane cargoes*

C_4 vol%	YTD mean	rolling average $n = 3$	cargo km^3	YTD weighted mean
3.9	3.90		75	
4.7	4.30		25	
4.2	4.27	4.27	50	4.13
3.7	4.13	4.20	60	4.01
4.7	4.24	4.20	60	4.16
4.4	4.27	4.27	80	4.22
4.7	4.33	4.60	75	4.30
4.0	4.29	4.37	25	4.29
4.8	4.34	4.50	70	4.35
4.2	4.33	4.33	25	4.35
4.1	4.31	4.37	80	4.32
4.0	4.28	4.10	65	4.29
4.0	4.26	4.03	25	4.28
4.4	4.27	4.13	75	4.29
4.8	4.31	4.40	80	4.34
3.9	4.28	4.37	60	4.31
4.3	4.28	4.33	25	4.31
4.0	4.27	4.07	80	4.28
4.8	4.29	4.37	60	4.31
4.1	4.29	4.30	75	4.30

To properly calculate the mean we should first convert each of the cargoes to km^3. Volume is mass (w) divided by density (x), so Equation (4.11) effectively divides the total mass of the cargoes by their total volume and gives the mean density of 0.9724. While an error in the fourth decimal place may seem negligible, it may be significant when compared to the potential improvement. For example, if improved control increased the mean density to 0.9750, the increase would be about 13% higher than that indicated by Equation (4.3) and so too would be the economic benefit.

Table 4.2 shows how the harmonic mean changes as additional cargoes are produced.

There is also the *geometric mean* derived by multiplying the n values together and taking the n^{th} root of the result, i.e.

$$\bar{x}_g = \sqrt[n]{\prod_{i=1}^{n} x_i} \tag{4.12}$$

The geometric mean can be useful in determining the mean of values that have very different ranges. For example, in addition to density, heavy fuel oil is subject to a maximum viscosity specification of 380 cSt and a maximum sulphur content of 3.5 wt%. If only one diluent is available then it must be added until all three specifications are met. The most limiting specification will not necessarily be the same for every cargo since the properties of the base fuel oil and the diluent will vary. To assess giveaway we would have to divide cargoes into three groups, i.e. those limited on density, those limited on viscosity and those limited on sulphur. In principle we can avoid this by calculating, for each cargo, the geometric mean of the three properties. If all three specifications were exactly met, the geometric mean of the properties would be 10.96. If any property is in giveaway, for example by being 10% off target, then the geometric mean would be reduced by 3.2% – no matter which property it is. However, this should be considered

Table 4.2 *Averaging SG of heavy fuel oil cargoes*

SG	cargo kt	YTD harmonic mean
0.9480	80	0.9480
0.9880	120	0.9716
0.9740	100	0.9724
0.9830	120	0.9754
0.9510	100	0.9706
0.9520	80	0.9681
0.9830	80	0.9698
0.9560	100	0.9680
0.9640	80	0.9677
0.9560	120	0.9662
0.9840	100	0.9678
0.9700	100	0.9680
0.9600	80	0.9675
0.9770	100	0.9682
0.9550	80	0.9675
0.9700	120	0.9676
0.9730	80	0.9679
0.9730	80	0.9681
0.9770	100	0.9686
0.9840	100	0.9694
0.9830	80	0.9699
0.9580	120	0.9693

as only an indicative measure of the potential to reduce giveaway. The amount of diluent required to reduce density by 10% is not the same as that required to reduce viscosity by 10%. A more precise approach would be to calculate, for each cargo, exactly how much less diluent could have been used without violating any of the three specifications.

Laws of heat transfer, vaporisation and chemical reaction all include a logarithmic function. Taking logarithms of Equation (4.12)

$$\log\left(\bar{x}_g\right) = \frac{\sum_{i=1}^{n} \log(x_i)}{n} \tag{4.13}$$

Rather than taking the logarithm of each measurement before calculating the mean, we could instead calculate the geometric mean.

Not to be confused with this definition of the geometric mean, there is also the *logarithmic mean*. It is limited to determining the mean of two positive values. If these are x_1 and x_2 then it is defined as

$$\bar{x}_l = \frac{x_1 + x_2}{\ln(x_1) - \ln(x_2)} = \frac{x_1 + x_2}{\ln\left(\frac{x_1}{x_2}\right)} \tag{4.14}$$

It has limited application but is most notably used in calculating the *log mean temperature difference* (*LMTD*) used in heat exchanger design. While Equation (4.14) uses the natural logarithm, the logarithm to any base (e.g. 10) may be used.

Table 4.3 *Analyses of propane cargoes*

vol% C_2	vol% C_3	vol% C_4
1.3	97.8	0.9
3.2	95.2	1.6
1.1	96.2	2.7
2.9	96.0	1.1
0.8	95.8	3.4
0.5	95.1	4.4
1.5	95.2	3.3
0.3	97.7	2.0
2.4	95.4	2.2
2.5	96.9	0.6

The definition of mean can be extended to multidimensional space. Table 4.3 shows the composition of 10 cargoes of propane. Propane must be at least 95 vol% pure and so the total of the C_2 content and the C_4 content must not exceed 5 vol%. Both these components have a lower value than propane and so their content should be maximised. To quantify what improvement is possible we can determine the *centroid*. This is the point at which the *residual sum of the squares* (*RSS*) of the distances from it to the data points is a minimum. In our example we have three variables: vol% C_2 (x_1), vol% C_3 (x_2) and vol% C_4 (x_3). We therefore adjust the coordinates a, b and c to minimise the function

$$RSS = \sum_{i=1}^{n} (x_{1i} - a)^2 + \sum_{i=1}^{n} (x_{2i} - b)^2 + \sum_{i=1}^{n} (x_{3i} - c)^2 \tag{4.15}$$

To identify the minimum we set the partial derivatives of this function to zero. For example, partially differentiating with respect to x_1 gives

$$\frac{\partial RSS}{\partial x_1} = 2\sum_{i=1}^{n} (x_{1i} - a) = 0 \tag{4.16}$$

$$\therefore \sum_{i1}^{n} x_{1i} - na = 0 \quad \text{and so} \quad a = \frac{\sum_{i=1}^{n} x_{1i}}{n} = \bar{x}_1 \tag{4.17}$$

Therefore *RSS* will be at a minimum when the coordinates of the centroid are the arithmetic means of x_1, x_2 and x_3, i.e. (1.65, 96.13, 2.22). The centroid is therefore mathematically no different from calculating the means separately. Its main advantage, for the two-dimensional case, is the way it presents the data. This is shown by Figure 4.1, which plots two of the three dimensions. The coloured points are the compositions of the cargoes; the white point is the centroid. It shows that giveaway could be eliminated by increasing the C_2 content to 2.96 or increasing the C_4 content to 3.35. Of course any combination of increases in the two components is possible, provided they sum to 1.13. Indeed one of the components could be increased beyond this value, provided the other is reduced. The strategy adopted would depend on the relative values of C_2 and C_4 when routed to their next most valuable alternative disposition.

In the same way we can add weighting factors to the arithmetic mean, we can do so to the calculation of the centroid. Indeed its alternative name, *centre of mass*, comes from using it to calculate the position of the centre of gravity of distributed weights.

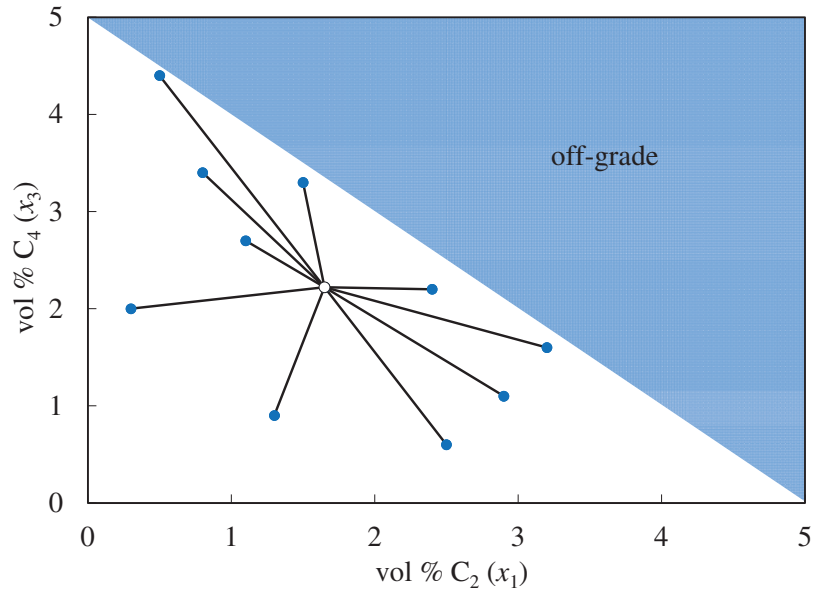


Figure 4.1 *Centroid of propane composition*

While using the mean makes good engineering sense in assessing opportunities for process improvements, it can give a distorted measure of performance. The inclusion of a single measurement that is very different from the others can, particularly if the sample size is small, significantly affect the estimate of the mean. An example might be a measurement collected under very unusual conditions, such as a process upset, or simply an error. We will present later how such *outliers* might be excluded. Doing so results in a *trimmed* (or *truncated*) mean. For example, we could simply exclude the lowest and highest values. So, after ranking the values of x from the lowest (x_1) to the highest (x_n), the trimmed arithmetic mean becomes

$$\bar{x}_t = \frac{\sum_{i=2}^{n-1} x_i}{n-2} \tag{4.18}$$

The *Winsorised mean* is based on a similar approach but the most outlying values are replaced with the adjacent less outlying values. For example, if we Winsorise the two most outlying values, the mean would be calculated as

$$\bar{x}_w = \frac{x_2 + x_{n-1} + \sum_{i=2}^{n-1} x_i}{n} \tag{4.19}$$

If n is 10, Equation (4.19) describes the *10% Winsorised mean*; we have removed 10% of the lower and upper outliers. If the sample size were increased to 20 then, to maintain the same level of Winsorisation, we would replace x_1 and x_2 with x_3, and x_{19} and x_{20} with x_{18}.

We will show later that exclusion of outliers carries the risk that an important aspect of process behaviour will be overlooked. An alternative approach is to define the centre of our sample using the *median*. In order to determine the median we again rank the dataset. If there is an odd number of measurements in the set then the median is the middle ranked value. If there is an

even number, it is the average of the two middle values. The addition of an outlier (no matter what its value) to the dataset will therefore have very little effect on the median – shifting it half the distance between two adjacent values ranked in the middle. The median can also be described as the 50 percentile, i.e. 50% of the values lie below (or above) the median.

We can also define *quartiles* – 25% of the values lie below the *first quartile* (Q_1) and 75% lie below the *third quartile* (Q_3). While the principle of determining the quartiles is clear, there are two different ways they can be calculated. These are known as the *inclusive* and *exclusive* methods. The inclusive method is based on the intervals between the ranked data. If there are n samples in the dataset, then there are $n - 1$ intervals. The first quartile is then the value $(n - 1)/4$ intervals from the start, i.e. the value ranked $(n - 1)/4 + 1$. The third quartile is the value ranked $3(n - 1)/4 + 1$. For example, in the dataset containing the values 10, 20, 30, 40 and 50, the first quartile is 20 and the third is 40. It is quite likely, however, that the calculated rank will not be an integer. If so, then we obtain the quartile by interpolating between the adjacent values. For example, if we were to add the value 60 to the dataset, the ranking for the first quartile becomes 2.25. Since this is nearer 2 than 3, we interpolate as $(0.75 \times 20) + (0.25 \times 30)$ or 22.5. The third quartile is 47.5.

The exclusive approach effectively adds a value ranked 0 to the data. The first quartile is then the value ranked $(n + 1)/4$. For example, if we were to increase n to 7 by adding 70 to the dataset, the first quartile would be the value ranked at 2, i.e. 20. The third quartile would be the value ranked $3(n + 1)/4$, i.e. 60. Non-integer results would similarly be dealt with by interpolation. For example, if we revert to the dataset without the 70, the first quartile is given by $(0.25 \times 10) + (0.75 \times 20)$, or 17.5. The third quartile is 52.5.

While each method gives very different results, this is primarily because the dataset is very small. The difference would be negligible if the dataset is large and well dispersed.

The median can also be described as the *second quartile* (Q_2). The average of the first and third quartile is known at the *midhinge* – another parameter representing the central value. There is also the *trimean*, defined as $(Q_1 + 2Q_2 + Q_3)/4$. Finally there is the *midrange*, which is simply the average of the smallest and largest values in the dataset.

There is a multidimensional equivalent to the median, the *geometric median*. It is similar to the centroid but is positioned to minimise the sum of the distances, not the sum of their squares. For the three-dimensional case the penalty function (F) is described by

$$F = \sum_{i=1}^{n} \sqrt{(x_{1i} - a)^2 + (x_{2i} - b)^2 + (x_{3i} - c)^2} \tag{4.20}$$

Unlike the centroid, the coordinates of the geometric median cannot be calculated simply. An iterative approach is necessary that adjusts a, b and c to minimise F. Applying this to the data in Table 4.3 gives the coordinates as (1.62, 96.02, 2.36). This is very close to the centroid determined from Equation (4.17). However, like its one-dimensional equivalent, the geometric median is less sensitive to outliers. For example, if we increase the value of the last C_2 content in Table 4.3 from 2.5 to 12.5, the mean increases from 1.65 to 2.65. The geometric median moves far less to (1.74, 95.77, 2.49).

Unlike the median, the geometric median will not be one of the data points. The data point nearest to the geometric mean is known as the *medoid*.

In process engineering, the mean value of a parameter has a true engineering meaning. For example, daily production averaged over a year can be converted to an annual production (simply by multiplying by 365) and any cost saving expressed per unit production can readily be converted to an annual saving. The same is not true of the median. While of some qualitative value in presenting data, it (and its related parameters) cannot be used in any meaningful calculation.

4.5 Dispersion

Once we have defined the central value, we need some measure of the dispersion of the data around it – often described as *variability* or *spread*. There are several simple parameters that we might consider. They include the *range*, which is simply the difference between the highest and lowest values in the dataset. It is sensitive to outliers, but we can deal with this in much the same way as we did in determining the mean. For example, we can use the *trimmed range* or *truncated range* where some criterion is applied to remove outliers. We can similarly *Winsorise* the range (by replacing outliers with adjacent values nearer the central value), although here this will give the same result as truncation. However these measures of dispersion, while offering a qualitative view, cannot readily be used in engineering calculations, for example, to assess potential improvements.

We can base measures of dispersion on the distance from the median. The most common of these is the *interquartile range* – the difference between the third and first quartile. It is equivalent to the 25% Winsorised range. There is also the *quartile deviation*, which is half of the interquartile range. There are measures based on *deciles*. For example, 10% of the values lie below the first decile (D_1) and 10% lie above the ninth decile (D_9). The difference between the two is another measure of dispersion. However, as described in the previous section, all such parameters are largely qualitative measures.

A better approach is to use the mean as the central value. We then calculate the deviation from the mean of each value in the dataset. In principle we could then sum these deviations as a measure of dispersion. However, from the definition of the arithmetic mean given by Equation (4.2), they will always sum to zero.

$$\sum_{i=1}^{n}(x_i-\bar{x})=\sum_{i=1}^{n}x_i-n\bar{x}=\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}x_i=0 \tag{4.21}$$

A possible solution is to use the absolute value (D) of the deviation from the mean. This suffers the problem that increasing the number of data points in the set will increase the sum of the absolute deviations, even if there is no increase in dispersion. To overcome this we divide by n to give the *mean absolute deviation*

$$\bar{D}=\frac{\sum_{i=1}^{n}|x_i-\bar{x}|}{n} \tag{4.22}$$

However, this does lend itself to use in further mathematical analysis. For example, if we know only the mean absolute deviation for each of two datasets, we cannot derive it for the combined set.

A more useful way of removing the sign of the deviation is to square it. The average sum of the squares of the deviations is the *variance*. Its positive square root is the *standard deviation* (σ).

$$\sigma^2=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n} \tag{4.23}$$

Expanding gives

$$\sigma^2=\frac{\sum_{i=1}^{n}x_i{}^2-2\bar{x}\sum_{i=1}^{n}x_i+n\bar{x}^2}{n} \tag{4.24}$$

Combining with Equation (4.2) gives an alternative way of calculating the variance.

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 \tag{4.25}$$

The main advantage of variances is that they are additive. If, in addition to the series of values for x, we have one for values of y then from Equation (4.25)

$$\sigma_{x+y}^2 = \frac{\sum_{i=1}^{n}(x_i + y_i)^2}{n} - (\bar{x} + \bar{y})^2 \tag{4.26}$$

$$= \frac{\sum_{i=1}^{n} x_i^2 + 2\sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2}{n} - \left(\bar{x}^2 + 2\bar{x}\bar{y} + \bar{y}^2\right) \tag{4.27}$$

$$= \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2\right) + \left(\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2\right) + 2\left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}\right) \tag{4.28}$$

$$= \sigma_x^2 + \sigma_y^2 + 2\sigma_{xy} \tag{4.29}$$

Following the same method we can determine the variance of the difference between two variables.

$$\sigma_{x-y}^2 = \sigma_x^2 + \sigma_y^2 - 2\sigma_{xy} \tag{4.30}$$

The term σ_{xy} is the *covariance*. We cover this in more detail in Section 4.9 but, if x and y are independent variables, their covariance will be zero. The variance of the sum of (or difference between) two variables will then be the sum of their variances.

Variance, because of the squaring of the deviation from the mean, is very sensitive to outliers. For example, the variance of C_2 vol% in Table 4.3 is 1.05. Increasing just the last value from 2.5 to 5.0 increases the variance to 2.11.

We may wish to compare dispersions that have different ranges or engineering units. Some texts suggest we modify the measure of dispersion to make it dimensionless. For example, dividing standard deviation by the mean gives the *coefficient of variation*. Other commonly documented dimensionless measures include the *coefficient of range* (the range divided by the sum of the lowest and highest values) and the *quartile coefficient* (the interquartile range divided by the sum of the first and third quartiles). In practice these values can be misleading. Consider a product that comprises mainly a component A and an impurity B. Assume the percentage of B in the product has a mean of 5 vol% and a standard deviation of 1%. The coefficient of variation is therefore 0.20. It follows therefore that the mean percentage of component A is 95 vol%. Its standard deviation will also be 1 vol% – giving a coefficient of variation of 0.01. This would appear to suggest that control of product purity is 20 times better than control of impurity where, in a two-component mixture, they must be the same.

4.6 Mode

Another measure commonly quoted is the *mode*. In principle it is the value that occurs most frequently in the dataset. However, if the values are continuous, it is quite possible that no two are the same. Instead we have to partition the data into ranges (known as *bins*); the mode

is then the centre of the most populated range. While in most cases the mode will be towards the centre of the dataset, there is no guarantee that it will be. It has little application in the statistics associated with process control. However it is important that we work with distributions that are *unimodal*, i.e. they have a single mode. A distribution would be *bimodal*, for example, if a second grade of propane was produced (say, with a minimum purity target of 90%) and the analyses included in the same dataset as the higher purity grade. A distribution can also be *multimodal* – having two or more modes.

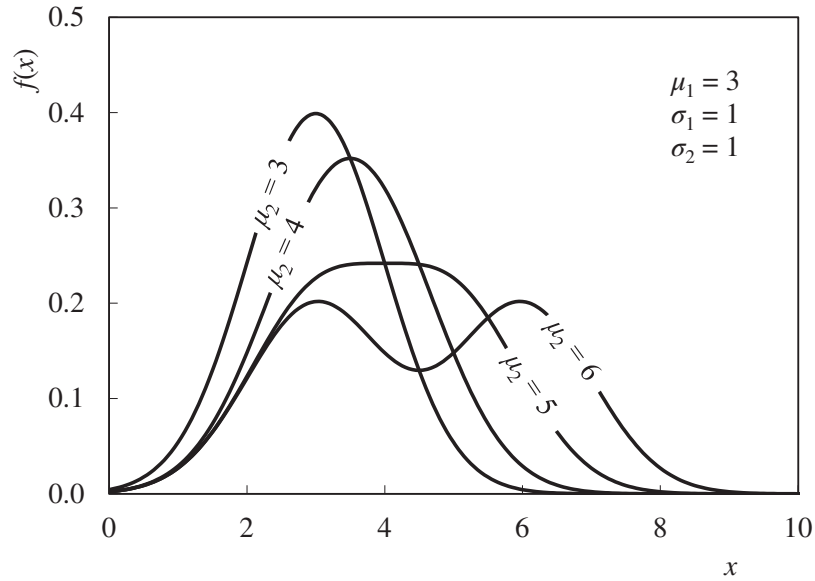


Figure 4.2 *Multimodal distributions*

A multimodal distribution can be easily mistaken for a unimodal one. Figure 4.2 illustrates this. Two distributions, with different means, have been combined. Both have a standard deviation of 1. In general, if the difference between the means is greater than double the standard deviation, the distribution will be clearly bimodal. In our example, where μ_2–μ_1 is 2, a second mode is not visible. Calculating the standard deviation without taking account of the modality would substantially overestimate the dispersion of the data. In this example it would be around 40% higher than the true value.

If a product is produced to different specifications, then the distribution will certainly be multimodal, even if not visibly so. A number of the distributions described in this book can be bimodal. However, their use is better avoided by segregating the data so that the standard deviation is determined separately for each grade of product. Alternatively, instead of obtaining the statistics for the propane purity, we obtain them for the deviation from target. Indeed this is exactly how we might assess the performance of a PID controller. We determine the standard deviation of the error, not the measurement. We need, however, to be more careful in assessing the performance of an inferential property. We might do so by monitoring the standard deviation of the difference between it and the laboratory. However, a change in operating mode might cause a bias error in the inferential. A bias error only contributes to standard deviation when it changes. Frequent changes in mode will therefore increase the standard deviation, suggesting that the inferential is performing poorly. To properly check its performance we should segregate the data for each operating mode.

Although not generally encountered in the process industry, a distribution can pass through a minimum, rather than a maximum. The minimum is known as the *anti-mode* – the value that occurs least often in the dataset. The distribution would then be described as *anti-modal*.

4.7 Standard Deviation

Standard deviation can be considered a measure of *precision*. Indeed, some texts define precision as the reciprocal of the variance (often using the term τ). Others define it as the reciprocal of the standard deviation (using the term τ'). Control engineers, of course, use τ to represent process lag. To avoid confusion, we avoid using precision as a statistical parameter. It does however have meaning. For example, the standard deviation of a variable that we wish to control is a measure of how precisely that variable is controlled. Indeed, reducing the standard deviation is often the basis of benefit calculations for process control improvements. If control were perfect the standard deviation would be zero. We similarly assess the performance of an inferential property from the standard deviation of the prediction error. However precision is not the same as *accuracy*. For example, if an inferential property consistently misestimates the property by the same amount, the standard deviation would be zero but the inferential would still be inaccurate. We have to distinguish between bias error and random error. Standard deviation is a measure only of random error.

We need to distinguish between *population* and *sample*. The population includes every value. For example, in the process industry, the population of daily production rates includes every measured rate since the process was commissioned until it is decommissioned. We clearly have no values for production rates between now and decommissioning. Records may not be available as far back as commissioning and, even if they are, the volume of data may be too large to retrieve practically. In practice, we normally work with a subset of the population, i.e. a sample. We need, of course, to ensure that the sample is representative by ensuring it includes values that are typical and that the sample is sufficiently large.

The standard deviation (σ_p) of the whole population of N values, if the *population mean* is μ, is given by

$$\sigma_p^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N} \tag{4.31}$$

When executing process control benefits studies we select a sample – a period for analysis comprising n data points. Those performing such analysis will likely have noticed that, to account for this, μ in Equation (4.31) is replaced by the *sample mean* and the denominator is replaced by $n - 1$. The following explains why.

From the data points collected in the period we estimate the sample mean.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{4.32}$$

Applying Equation (4.31) to a sample of the population will underestimate the true standard deviation. This is because the sum of the squared deviations of a set of values from their sample mean ($\bar{x}$) will always be less than the sum of the squared deviations from a different value, such

as the population mean (μ). To understand this consider the trivial example where we have a sample of two data points with values 1 and 5. The sample mean is 3 and the sum of the deviations from the mean is 8 ($2^2 + 2^2$). If, instead of using the sample mean, we choose to use a higher value of 4, the sum of the deviations will then be 10 ($3^2 + 1^2$). We would get the same result if we had chosen a lower value of 2 for the mean. The nonlinearity, caused by squaring, results in the increase in squared deviation in one direction being greater than the decrease in the other.

We do not know the mean of the whole population (μ). Applying Equation (4.32) to different samples selected from the population will give a number of possible estimates of the true mean. These estimates will have a mean of μ. Similarly we do not know the standard deviation of the whole population. Imagine the sample mean being determined by, before summing all the data points, dividing each by n. The standard deviation of the resulting values will therefore be n times smaller, i.e. σ_p/n, giving a variance of $(\sigma_p/n)^2$. We have seen that variances are additive. The sum of the n values, which will now be the sample mean, will therefore have a variance n times larger, i.e. ${\sigma_p}^2/n$. The square root of this value is sometimes described as the *standard error*.

The variance of the sample (σ^2) is

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} \tag{4.33}$$

The variance of the population will be the variance of the sample plus the variance of the sample mean.

$$\sigma_p^2 = \sigma^2 + \frac{\sigma_p^2}{n} \quad \text{or} \quad \sigma_p^2 = \frac{n}{n-1}\sigma^2 \tag{4.34}$$

Substituting for σ^2 from Equation (4.33) gives

$$\sigma_p^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \tag{4.35}$$

We use Equation (4.35) to generate a *sample-adjusted* or *unbiased* variance. This technique is known as *Bessel's correction*. The denominator is often described as the number of *degrees of freedom*. We will see it is used in number of distributions. By definition it is the number of values in the dataset that can be freely adjusted while retaining the value of any quantified statistical parameters. In this case we have quantified only one parameter – the mean. We can freely adjust $n - 1$ of the values, provided we adjust the n^{th} value to keep the total, and hence the mean, constant.

In practice, if the number of data points is sufficiently large, the error introduced is small. For example, with a value of 50 for n, the effect on σ^2 will be to change it by about 2%, with a change in σ of less than 1%.

To remove the need to first calculate the sample mean, Equation (4.35) can be rewritten as

$$\sigma_p^2 = \frac{\sum_{i=1}^{n} {x_i}^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2}{n-1} = \frac{n\sum_{i=1}^{n} {x_i}^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n(n-1)} \tag{4.36}$$

4.8 Skewness and Kurtosis

Like variance, *skewness* (γ) and *kurtosis* (κ) are used to describe the shape of the distribution. To mathematically represent the distribution of the data we first have to choose the form of the distribution. That chosen is known as the *prior distribution*. It will contain parameters (such as mean and variance) that are then adjusted to fit the real distribution as close as possible. The main use of skewness and kurtosis is to assess whether the actual distribution of the data is then accurately represented. They are examples of *moments*. The k^{th} *raw* moment (m) is defined as

$$m_k = \frac{\sum_{i=1}^{N} x_i^k}{N} \tag{4.37}$$

Although of little use, the *zeroth raw moment* ($k = 0$) has a value of 1. The *first raw moment* ($k = 1$) is the population mean (μ). *Central moments* (m') are calculated about the population mean

$$m'_k = \frac{\sum_{i=1}^{N} (x_i - \mu)^k}{N} \tag{4.38}$$

The *first central moment* will evaluate to zero.

$$m'_1 = \frac{\sum_{i=1}^{N} (x_i - \mu)}{N} = \frac{\sum_{i=1}^{N} x_i - N\mu}{N} = \mu - \mu = 0 \tag{4.39}$$

The second central moment is the population variance; replacing k in Equation (4.38) with 2 gives Equation (4.31). Higher moments are generally *normalised* or *standardised*, by dividing by the appropriate power of standard deviation of the population, so that the result is dimensionless.

$$m_k = \frac{\sum_{i=1}^{N} (x_i - \mu)^k}{N \sigma_p^k} \quad k > 2 \tag{4.40}$$

Skewness (γ) is the third central moment. If the number of data points in the sample is large then we can calculate it from Equation (4.40). Strictly, if calculated from a sample of the population, the formula becomes

$$\gamma = \frac{n}{(n-1)(n-2)} \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{\sigma_p^3} \quad n > 2 \tag{4.41}$$

If the skewness is greater than zero, it might indicate there are more values higher than the mean than there are below it. Or it might indicate that the values higher than the mean are further from it than the values below it. The value of skewness does not indicate the cause. It does not tell us whether the mean is less than or greater than the median.

As a simple example, consider a dataset containing the values 98, 99, 100, 101 and 107. The majority of the values are less than the mean of 101, indicating that the skew might be negative.

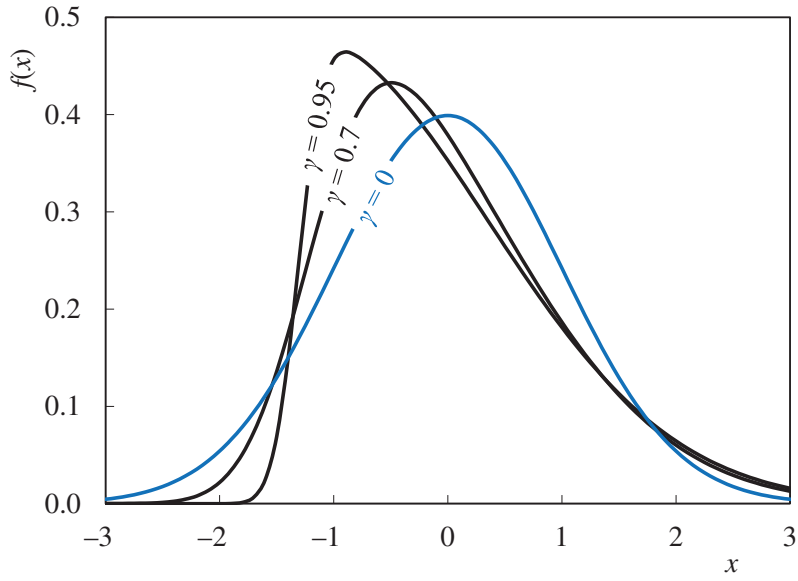


Figure 4.3 *Skewness*

However, the one value greater than the mean is far more distant from it than the others – possibly indicating a positive skew. This is confirmed by Equation (4.41), which gives the skewness as 1.7.

Figure 4.3 shows (as the coloured line) the normal distribution with a mean of 0 and standard deviation of 1. A source of confusion is that a positive skewness indicates a skew to the right. The black lines show increasing skewness while keeping the mean and standard deviation constant. The mode has moved to the left but, as skewness increases, values below the mean have approached the mean while some of values above the mean now form a more extended tail.

A normal distribution is symmetrical about the mean. Some texts therefore suggest that skewness lying between –0.5 and +0.5 is one of the indications that we can treat the distribution as normal. However, while a symmetrical distribution has a skewness of zero, the converse is not necessarily true. For example, a large number of values a little less than the mean might be balanced by a small number much higher than the mean. Skewness will be zero, but the distribution is clearly not symmetrical.

In any symmetrical distribution, the mean, median and mode will all have the same value.

Kurtosis (κ) is the fourth central moment given, for a sample of the population, by

$$\kappa = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{\sigma_p^4} - \frac{3(3n-5)}{(n-2)(n-3)} \qquad n > 3 \tag{4.42}$$

The kurtosis of a normal distribution is 3; for this reason many texts use the parameter γ_1 as skewness and γ_2 as *excess kurtosis*.

$$\gamma_1 = \gamma \qquad \gamma_2 = \kappa - 3 \tag{4.43}$$

Kurtosis is a measure of how flat or peaked is the distribution. It is a measure of how much of the variance is due to infrequent extreme deviations from the mean, rather than more frequent small deviations. This is apparent from examining Equation (4.42). If the deviation from the

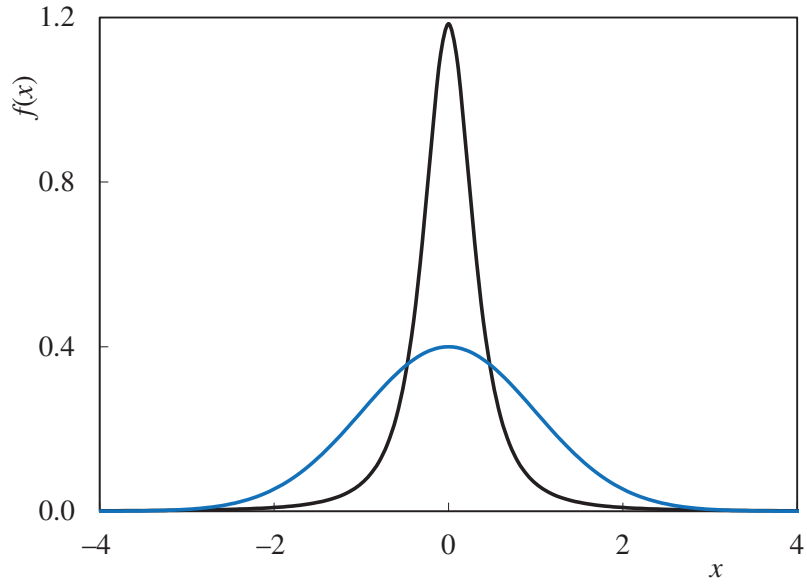


Figure 4.4 *Increasing kurtosis*

mean is less than the standard deviation then $(x_i - \bar{x})/\sigma_P$ will be less than 1. Raising this to the fourth power will make it considerably smaller and so it will contribute little to the summation. If the deviations are predominantly larger than the standard deviation (i.e. the distribution has long tails) then kurtosis will be large. Specifically, if excess kurtosis is positive ($\kappa > 3$) then the distribution is *leptokurtic*, i.e. it has a higher peak with long tails. If negative ($\kappa < 3$) then it is *platykurtic*, i.e. more flat with short tails. If excess kurtosis is zero ($\kappa = 3$) the distribution is described as *mesokurtic*. Indeed, if excess kurtosis is outside the range –0.5 to +0.5, we should not treat the distribution as normal. Many commonly used distributions, as we will see later, are leptokurtic and can be described as *super-Gaussian*. Platykurtic distributions can be described as *sub-Gaussian*.

Most spreadsheet packages and much statistical analysis software use excess kurtosis. To avoid confusion, and to keep the formulae simpler, this book uses kurtosis (κ) throughout.

Kurtosis is quite difficult to detect simply by looking at the distribution curve. Figure 4.4 shows (as the coloured line) the normal distribution – with a mean of 0 and a variance of 1. The black line has the same mean and variance but with kurtosis increased (to around 20). Figure 4.5 also shows the same normal distribution but this time the kurtosis is kept at zero and the variance reduced (to around 0.115). The dashed line looks almost identical to the solid line – although close inspection of the tails of the distribution shows the difference. Figure 4.6 shows the same distributions plotted on a cumulative basis and shows much the same difficulty. If, instead of plotting the function, the distribution were plotted from the data, it is even less likely that kurtosis could be seen. To detect it reliably (and to quantify it) kurtosis should at least be calculated as above, but preferably estimated from curve fitting.

Higher order moments can be defined but their interpretation is difficult and are rarely used. The fifth moment is *hyperskewness* and the sixth *hyperflatness*.

In addition to calculating the skewness and kurtosis from the data, we also need to determine them for the chosen distribution function. As we will see later, many distributions are documented with simple formulae but for some the calculations are extremely complex. Under these