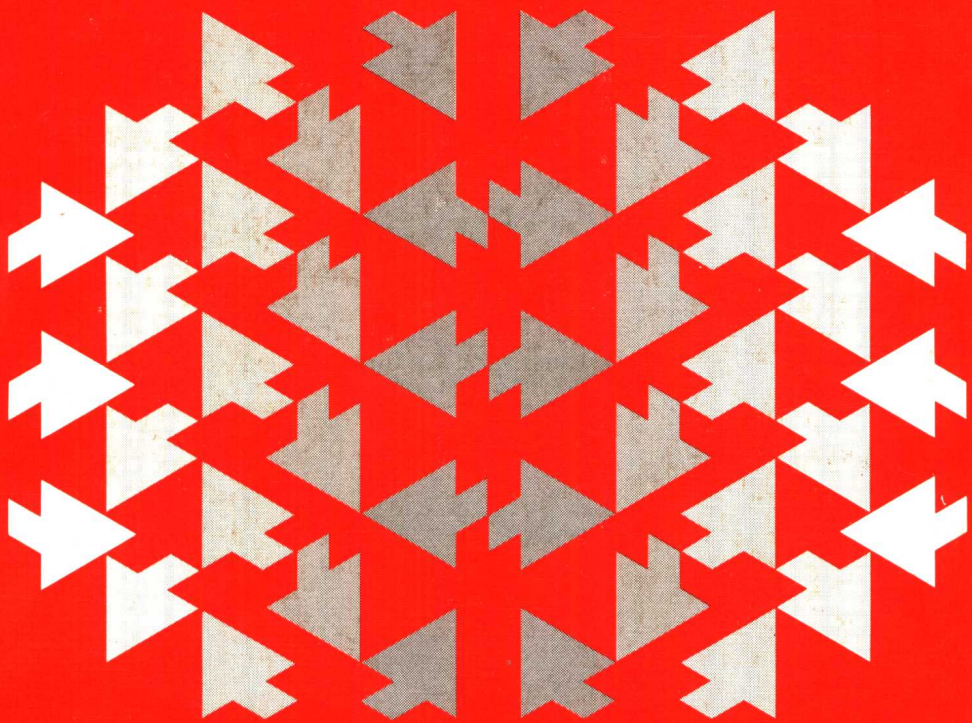
DATA HANDLING IN SCIENCE AND TECHNOLOGY

3

Experimental design: a chemometric approach

STANLEY N. DEMING
STEPHEN L. MORGAN



ELSEVIER

DATA HANDLING IN SCIENCE AND TECHNOLOGY – VOLUME 3

Experimental design: a chemometric approach

STANLEY N. DEMING

Department of Chemistry, University of Houston, Houston, TX 77004, U.S.A.

and

STEPHEN L. MORGAN

Department of Chemistry, University of South Carolina, Columbia, SC 29208, U.S.A.

ELSEVIER

Amsterdam – Oxford – New York – Tokyo **1987**

ELSEVIER SCIENCE PUBLISHERS B.V.
Sara Burgerhartstraat 25
P.O. Box 211, 1000 AE Amsterdam, The Netherlands

Distributors for the United States and Canada:

ELSEVIER SCIENCE PUBLISHING COMPANY INC.
655, Avenue of the Americas
New York, NY 10010, U.S.A.

First edition 1987
Second impression 1988

Library of Congress Cataloging-in-Publication Data

Deming, Stanley N., 1944-
Experimental design.

(Data handling science and technology ; v. 3)
Bibliography: p.
Includes index.
1. Chemistry, Analytic--Statistical methods.
2. Experimental design. I. Morgan, Stephen L., 1949- . II. Title. III. Series.
QD75.4.S6D46 1987 543'.007'.24 86-32625

ISBN 0-444-42734-1 (Vol. 3)
ISBN 0-444-42408-3 (Series)

Printed in The Netherlands

To
Bonnie, Stephanie, and Michael,
and to
Linda

Preface

As analytical chemists, we are often called upon to participate in studies that require the measurement of chemical or physical properties of materials. In many cases, it is evident that the measurements to be made will not provide the type of information that is required for the successful completion of the project. Thus, we find ourselves involved in more than just the measurement aspect of the investigation – we become involved in carefully (re)formulating the questions to be answered by the study, identifying the type of information required to answer those questions, making appropriate measurements, and interpreting the results of those measurements. In short, we find ourselves involved in the areas of experimental design, data acquisition, data treatment, and data interpretation.

These four areas are not separate and distinct, but instead blend together in practice. For example, data interpretation must be done in the context of the original experimental design, within the limitations of the measurement process used and the type of data treatment employed. Similarly, data treatment is limited by the experimental design and measurement process, and should not obscure any information that would be useful in interpreting the experimental results. The experimental design itself is influenced by the data treatment that will be used, the limitations of the chosen measurement process, and the purpose of the data interpretation.

Data acquisition and *data treatment* are today highly developed areas. Fifty years ago, measuring the concentration of fluoride ion in water at the parts-per-million level was quite difficult; today it is routine. Fifty years ago, experimenters dreamed about being able to fit models to large sets of data; today it is often trivial.

Experimental design is also today a highly developed area, but it is not easily or correctly applied. We believe that one of the reasons "experimental design" is not used more frequently (and correctly) by scientists is because the subject is usually taught from the point of view of the statistician rather than from the point of view of the researcher. For example, one experimenter might have heard about factorial designs at some point in her education, and applies them to a system she is currently investigating; she finds it interesting that there is a "highly significant interaction" between factors A and B, but she is disappointed that all this work has not revealed to her the particular combination of A and B that will give her optimal results from her system. Another experimenter might be familiar with the least squares fitting of straight lines to data; the only problems he chooses to investigate are those that can be reduced to straight-line relationships. A third experimenter might be asked to do a screening study using Plackett-Burman designs; instead, he transfers out of the research division.

We do not believe that a course on the design of experiments must necessarily be preceded by a course on statistics. Instead, we have taken the approach that both subjects can be developed simultaneously, complementing each other as needed, in a course that presents the fundamentals of experimental design.

It is our intent that the book can be used in a number of fields by advanced undergraduate students, by beginning graduate students, and (perhaps more important) by workers who have already completed their formal education. The material in this book has been presented to all three groups, either through regular one-semester courses, or through intensive two- or three-day short courses. We have been pleased by the confidence these students have gained from the courses, and by their enthusiasm as they study further in the areas of statistics and experimental design.

The text is intended to be studied in one way only – from the beginning of Chapter 1 to the end of Chapter 12. The chapters are highly integrated and build on each other: there are frequent references to material that has been covered in previous chapters, and there are many sections that hint at material that will be developed more fully in later chapters.

The text can be read "straight through" without working any of the exercises at the ends of the chapters; however, the exercises serve to reinforce the material presented in each chapter, and also serve to expand the concepts into areas not covered by the main text. Relevant literature references are often given with this latter type of exercise.

The book has been written around a framework of linear models and matrix least squares. Because we authors are so often involved in the measurement aspects of investigations, we have a special fondness for the estimation of purely experimental uncertainty. The text reflects this prejudice. We also prefer the term "purely experimental uncertainty" rather than the traditional "pure error", for reasons we as analytical chemists believe should be obvious.

One of the important features of the book is the *sums of squares and degrees of freedom tree* that is used in place of the usual ANOVA tables. We have found the "tree" presentation to be a more effective teaching tool than ANOVA tables by themselves.

A second feature of the book is its emphasis on degrees of freedom. We have tried to remove the "magic" associated with knowing the source of these numbers by using the symbols n (the total number of experiments in a set), p (the number of parameters in the model), and f (the number of distinctly different factor combinations in the experimental design). Combinations of these symbols appear on the "tree" to show the degrees of freedom associated with various sums of squares (e.g., $n - f$ for SS_{pe}).

A third feature is the use of the $\boldsymbol{J}$ matrix (a matrix of mean replicate response) in the least squares treatment. We have found it to be a useful tool for teaching the effects (and usefulness) of replication.

We are grateful to a number of friends for help in many ways. Grant Wernimont and L. B. Rogers first told us why statistics and experimental design should be

important to us as analytical chemists. Ad Olansky and Lloyd Parker first told us why a *clear presentation* of statistics and experimental design should be important to us; they and Larry Bottomley aided greatly in the early drafts of the manuscript. Kent Linville provided many helpful comments on the early drafts of the first two chapters. A large portion of the initial typing was done by Alice Ross; typing of the final manuscript was done by Lillie Gramann. Their precise work is greatly appreciated.

We are grateful also to the Literary Executor of the late Sir Ronald A. Fisher, F. R. S., to Dr. Frank Yates, F. R. S., and to Longman Group Ltd., London, for permission to partially reprint Tables III and V from their book *Statistical Tables for Biological, Agricultural and Medical Research*, 6th ed., 1974.

Finally, we would like to acknowledge our students who provided criticism as we developed the material presented here.

S. N. Deming
Houston, Texas
August 1986

S. L. Morgan
Columbia, South Carolina

Contents

CHAPTER 1

System Theory

General system theory is an organized thought process to be followed in relating cause and effect. The system is treated as a bounded whole with inputs and outputs external to the boundaries, and transformations occurring within the boundaries. Inputs, outputs, and transformations can be important or trivial – the trick is to determine which. The system of determination involves a study and understanding of existing theory in the chosen field; a study and understanding of past observations and experiments more specific to the chosen problem; and new experiments specific to the problem being studied.

General system theory is a highly versatile tool that provides a useful means of investigating many research and development projects. Although other approaches to research and development often focus on the detailed internal structure and organization of the system, our approach here will be to treat the system as a whole and to be concerned with its overall behavior.

1.1. Systems

A *system* is defined as a regularly interacting or interdependent group of items forming a unified whole. A system is described by its borders, by what crosses the borders, and what goes on inside. Thus, we often speak of a solar system when referring to a sun, its planets, their moons, etc.; a thermodynamic system when we are describing compounds in equilibrium with each other; and a digestive system if we are discussing certain parts of the body. Other examples of systems are ecological systems, data processing systems, and economic systems. We even speak of *the* system when we mean some part of the established order around us, some regularly interacting or interdependent group of items forming a unified whole of which we are a part.

General system theory views a system as possessing three basic elements – *inputs*, *transforms*, and *outputs* (see Figure 1.1). An example of a simple system is the mathematical relationship

$$y = x + 2. \tag{1.1}$$

This algebraic system is diagrammed in Figure 1.2. The *input* to the system is the independent variable x. The *output* from the system is the dependent variable y.

Figure 1.1. General system theory view of relationships among inputs, transforms, and outputs.

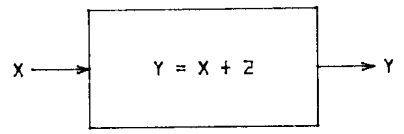


Figure 1.2. General system theory view of the algebraic relationship $y = x + 2$.

The *transform* that relates the output to the input is the well defined mathematical relationship given in Equation 1.1. The mathematical equation transforms a given value of the input, x, into an output value, y. If $x = 0$, then $y = 2$. If $x = 5$, then $y = 7$, and so on. In this simple system, the transform is known with certainty.

Figure 1.3 is a system view of a wine-making process. In this system, there are

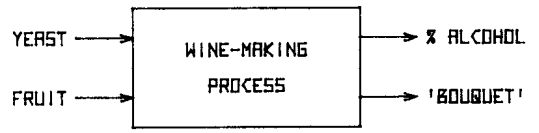


Figure 1.3. General system theory view of a wine-making process.

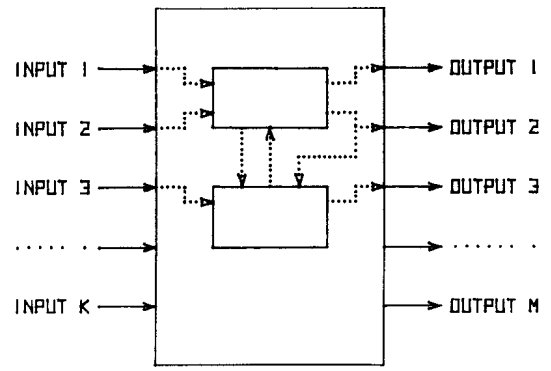


Figure 1.4. General system theory view emphasizing internal structures and relationships within a system.

two inputs (yeast and fruit), *two* outputs (percent alcohol and "bouquet"), and a transform that is probably *not* known with certainty.

Most systems are much more complex than the simple examples shown here. In general, there will be many inputs, many outputs, many transforms, and considerable subsystem structure. A more realistic view of most systems is probably similar to that shown in Figure 1.4.

1.2. Inputs

We will define a system *input* as a quantity or quality that might have an influence upon the system.

The definition of a system input is purposefully broad. It could have been made narrower to include only those quantities and qualities that *do* have an influence upon the system. However, because a large portion of the early stages of much research and development is concerned with determining which inputs do have an influence and which do not, such a narrow definition would assume that a considerable amount of work had already been carried out. The broader definition used here allows the inclusion of quantities and qualities that might eventually be shown to have no influence upon the system, and is a more useful definition for the early and speculative stages of most research.

We will use the symbol x with a subscript to represent a given input. For example, x_1 means "input number one", x_2 means "input number two", x_i means "the ith input", and so on.

The intensity setting of an input is called a *level*. It is possible for an input to be at different levels at different times. Thus, x_1 might have had the value 25 when we were first interested in the system; now we might want x_1 to have the value 17. To designate these different sets of conditions, a second subscript is added. Thus, $x_{11} = 25$ means that in the first instance, $x_1 = 25$; and $x_{12} = 17$ means that in the second instance, $x_1 = 17$.

Ambiguity is possible with this notation: e.g., x_{137} might refer to the level of input number one under the 37th set of conditions, or it might refer to the level of input 13 under the seventh set of conditions, or it might refer to input number 137. To avoid this ambiguity, subscripts greater than nine can be written in parentheses or separated with commas. Thus, $x_{1(37)}$ and $x_{1,37}$ refer to the level of input number one under the 37th set of conditions, $x_{(13)7}$ and $x_{13,7}$ refer to the level of input 13 under the seventh set of conditions, and $x_{(137)}$ and x_{137} refer to input number 137.

Input variables and factors

A system *variable* is defined as a quantity or quality associated with the system that may assume any value from a set containing more than one value. In the

algebraic system described previously, "x" is an *input variable*: it can assume any one of an infinite set of values.

"Yeast" and "fruit" are input variables in the wine-making process. In the case of yeast, the amount of a given strain could be varied, or the particular type of yeast could be varied. If the variation is of extent or quantity (e.g., the use of one ounce of yeast, or two ounces of yeast, or more) the variable is said to be a *quantitative variable*. If the variation is of type or quality (e.g., the use of *Saccharomyces cerevisiae*, or *Saccharomyces ellipsoideus*, or some other species) the variable is said to be a *qualitative variable*. Thus, "yeast" could be a qualitative variable (if the amount added is always the same, but the type of yeast is varied) or it could be a quantitative variable (if the type of yeast added is always the same, but the amount is varied). Similarly, "fruit" added in the wine-making process could be a qualitative variable or a quantitative variable. In the algebraic system, x is a quantitative variable.

A *factor* is defined as one of the elements contributing to a particular result or situation. It is an input that *does* have an influence upon the system.

In the algebraic system discussed previously, x is a factor; its value determines what the particular result y will be. "Yeast" and "fruit" are factors in the wine-making process; the type and amount of each contributes to the alcohol content and flavor of the final product.

In the next several sections we will further consider factors under several categorizations.

Known and unknown factors

In most research and development projects it is important that as many factors as possible be known. Unknown factors can be the witches and goblins of many projects – unknown factors are often uncontrolled, and as a result such systems appear to behave excessively randomly and erratically. Because of this, the initial phase of many research and development projects consists of screening a large number of input variables to see if they are factors of the system; that is, to see if they have an *effect* upon the system.

The proper identification of factors is clearly important (see Table 1.1). If an input variable *is* a factor and it *is* identified as a factor, the probability is increased for the success of the project. If an input variable truly *is* a factor but it *is not*

TABLE 1.1
Possible outcomes in the identification of factors.

Type of input variable	Identified as a factor	Not identified as a factor
A factor	Desirable for research and development	Random and erratic behavior
Not a factor	Unnecessary complexity	Desirable for research and development

included as an input variable and/or *is not* identified as a factor, random and erratic behavior might result. If an input variable *is not* a factor but it is falsely identified as a factor, an unnecessary input variable will be included in the remaining phases of the project and the work will be unnecessarily complex. Finally, if an input variable *is not* a factor and *is not* identified as a factor, ignoring it will be of no consequence to the project.

The first and last of the above four possibilities are the desired outcomes. The second and third are undesirable outcomes, *but undesirable for different reasons and with different consequences*. The third possibility, falsely identifying an input variable as a factor, is unfortunate but the consequences are not very serious: it might be expensive, in one way or another, to carry the variable through the project, but its presence will not affect the ultimate outcome. However, the second possibility, *not* identifying a factor, can be very serious: omitting a factor can often cause the remaining results of the project to be worthless.

In most research and development, the usual approach to identifying important factors uses a statistical test that is concerned with the risk (α) of stating that an input variable is a factor when, in fact, it is not – a risk that is of relatively little consequence (see Table 1.1). Ideally, the identification of important factors should also be concerned with the potentially much more serious risk (β) of stating that an input variable is not a factor when, in fact, it is a factor (see Table 1.1). This subject is discussed further in Chapter 6.

In our representations of systems, a known factor will be shown as a solid arrow pointing toward the system; an unknown factor will be shown as a dotted arrow pointing toward the system (see Figure 1.5).

Controlled and uncontrolled factors

The word "control" is used here in the sense of exercising restraint or direction over a factor – that is, the experimental setting of a factor to a certain quantitative or qualitative value.

Controllable factors are desirable in experimental situations because their effects can usually be relatively easily and unambiguously detected and evaluated. Examples of individual controllable factors include x, yeast, fruit, temperature, concentration, time, amount, number, and size.

Uncontrollable factors are undesirable in experimental situations because their effects cannot always be easily or unambiguously detected or evaluated. Attempts

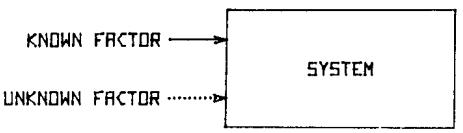


Figure 1.5. Symbols for a known factor (solid arrow) and an unknown factor (dotted arrow).

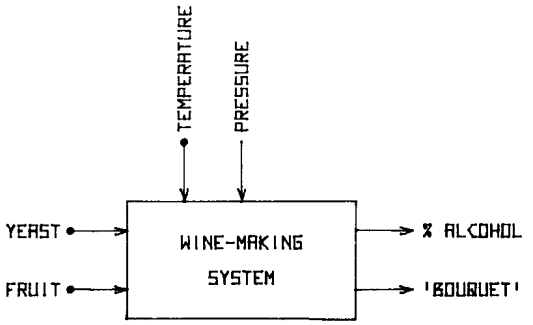


Figure 1.6. Symbols for controlled factors (arrows with dot at tail) and an uncontrolled factor (arrow without dot).

are often made to minimize their effects statistically (e.g., through randomization of experiment order – see Section 12.2) or to separate their effects, if known, from those of other factors (e.g., by measuring the level of the uncontrolled factor during each experiment and applying a "known correction factor" to the experimental results). Examples of individual uncontrollable factors include incident gamma ray background intensity, fluctuations in the levels of the oceans, barometric pressure, and the much maligned "phase of the moon".

A factor that is uncontrollable by the experimenter might nevertheless be controlled by some other forces. Incident gamma ray background intensity, fluctuations in the levels of the oceans, barometric pressure, and phase of the moon cannot be controlled by the experimenter, but they are "controlled" by the "Laws of Nature". Such factors are usually relatively constant with time (e.g., barometric pressure over a short term), or vary in some predictable way (e.g., phase of the moon).

A controlled factor will be identified as an arrow with a dot at its tail; an uncontrolled factor will not have a dot. In Figure 1.6, temperature is shown as a controlled known factor; pressure is shown as an uncontrolled known factor. "Yeast" and "fruit" are presumably controlled known factors.

Intensive and extensive factors

Another categorization of factors is based upon their dependence on the size of a system. The value of an *intensive factor* is not a function of the size of the system. The value of an *extensive factor* is a function of the size of the system.

The temperature of a system is an intensive factor. If the system is, say, 72°C, then it is 72°C independent of how large the system is. Other examples of intensive factors are pressure, concentration, and time.

The mass of a system, on the other hand, does depend upon the size of the