THE ANALYSIS OF UNREPLICATED FACTORIAL EXPERIMENTS

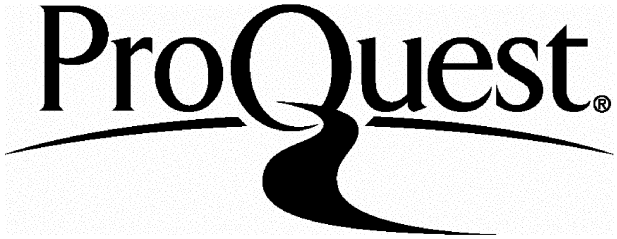
Thesis submitted to the University of London for the degree of Doctor of Philosophy in the Faculty of Science

by

Jorge Manuel Olguín Uribe

University College London

August, 1994



Abstract

This thesis addresses the problem of analyzing fractional factorial and other orthogonal experimental arrangements that are carried out without replication, so that the usual estimate of standard error computed from replications cannot be used.

Although the usefulness of unreplicated factorial experiments in industry was recognized soon after R. A. Fisher developed the basic ideas of statistical design of experiments, the influence of the Japanese ideas of quality improvement over the last few years has increased considerably the utilization of these experiments, and their analysis has become a major concern. Considering that these experiments are frequently analyzed by non-statisticians, it is desirable to provide easy to use and understand techniques to assess the significance of the estimates of the different effects involved in a particular experiment.

After a literature review, four methods for the analysis of unreplicated factorials are chosen, implemented, made comparable on an error rates basis, and applied to a substantial number of real experiments taken from the statistical literature and other sources.

In the process of making the methods comparable for the most common sizes of the experiments, one of the methods, which is found to be miscalibrated, is corrected.

On the basis of the result of the comparison of those methods, a modification of a procedure of statistical testing with half-normal plots is presented. The method

proposed, besides having the advantages of the half-normal plot, seems to be more powerful than other comparable methods, for experiments of small and moderate sizes.

Finally some ideas to make a Bayesian procedure proposed by Box and Meyer more general and robust are explored for situations suggested by 30 examples of small experiments.

Acknowledgments

I want to thank Tom Fearn who first interested me in the subject and who, as supervisor, was patient and encouraging.

My gratitude also goes to fellow students, staff of the Department of Statistical Science at UCL, and other friends who helped and supported me in various ways throughout my research. I would like to mention specially Lenita Turchi, Francisco Aranda (in memoriam), Jim Burridge, Adam Crisp, Eduardo Gutiérrez, Federico O'Reilly and Belem Trejo.

Finally I wish to acknowledge that my studies in the UK would not have been possible without the financial support of the *Universidad Nacional Autónoma de México.*

Contents

1 The quality revolution and unreplicated factorial experiments 8

1.1 The Japanese quality movement 8

1.2 Brief historical review of the use of unreplicated factorials 12

1.3 Example 14

2 Methods for assessing the significance of contrasts for unreplicated factorials 22

2.1 Assumptions and description of the problem 22

2.2 Traditional pooling 23

2.3 60% pooling 24

2.4 t margins of error 25

2.5 Inference with half-normal plots 26

2.6 A Bayesian approach 29

2.7 The chain-pooling method 30

2.8 Testing normality and outliers 31

2.9 A procedure based on homogeneity of mean squares 32

2.10 A method based on the coefficient R^2 32

2.11 Discussion 32

3 Examples and implementation of selected methods **36**

3.1 Examples collected 37

3.2 Implementation of selected methods 38

3.2.1 Bayesian approach (BM) 38

3.2.2 Zahn's method (ZA) 39

3.2.3 60% procedure (BP) 42

3.2.4 Lenth's method (LE) 42

3.3 Example 43

4 Comparison of four selected methods **47**

4.1 Error rates 47

4.2 Corrected critical values for LE 50

4.3 Error rate characteristics for comparison criteria 51

4.4 Comparative performance on real data 54

4.4.1 Case $m = 7$ 54

4.4.2 Case $m = 15$ 56

4.4.3 Case $m = 31$ 57

4.4.4 Case $m = 63$ 58

4.5 Concluding remarks 59

5 An alternative method of inference using half-normal plots **61**

5.1 Analysis of alternatives 61

5.1.1 Case $m = 7$ 63

5.1.2 Case $m = 15$ 65

5.1.3 Cases $m = 31$ and $m = 63$ 67

5.2 Method HP 67

5.3 Comparative performance of HP 68

5.4 Example 69

5.5 A note about power 71

5.6 Concluding remarks 71

6 Exploring extensions to the Bayesian approach BM **75**

6.1 Considering prior distributions for the parameters α and k 79

6.2 Assuming a heavier-tailed distribution for the active contrasts . . . 81

6.3 Concluding remarks 83

7 Discussion and Conclusions **85**

A List of examples **90**

B MINITAB macro HP **95**

References **98**

Chapter 1

The quality revolution and unreplicated factorial experiments

1.1 The Japanese quality movement

In 1980 for the first time in its history Japan produced more cars than any other country in the world. A great number of these cars were exported to the US, the former main producer (see e.g. Ealey, 1988). This is only an indicator of the manufacturing shares that in different industries the US and the so-called West have lost to Japan.

Threatened by the Japanese competitiveness, US companies (and more recently UK and other western industries) have been interested in the Japanese systems of quality improvement. As Box, et al. (1988) pointed out, "in Japan quality control has become a cost saving approach, in contrast with the prevailing view in the West that high quality is associated with higher cost".

Garvin (1988) relates extensively the evolution of the quality movement in

Japan. He recounts that within a month of Japan's surrender in the World War II, the Supreme Commander for the Allied Forces established the Civil Communication Section whose Industrial Division was assigned to work with Japanese manufacturers of communication equipment to improve production methods. Because the poor reliability of the national communications network, quality became a principal concern.

It is interesting to note that some of the key engineers of the Industrial Division had work at Bell laboratories in the US, the place where in the 1920s and 1930s W. A. Shewhart developed the theoretical concepts of process in and out of control and the Shewhart chart.

Garvin reports that between 1945 and 1949 these engineers pursued a variety of activities, among them advising the new leaders of Japanese business on questions of production management with strong emphasis on quality.

When the Allied Command was disbanded, Japan was aware of the importance of quality for the future of its industry. In 1950 the Union of Japanese Scientists and Engineers (JUSE) invited the American expert in quality control W. Edwards Deming for a series of seminars (he returned to Japan in 1951 and 1952).

Deming had also worked at Bell Laboratories where he had been a leading disciple of W. A. Shewhart. And, as Cox (1990) notes, Deming was also well known for a number innovative technical ideas.

Deming pushed top managers to become actively involved in their companies' quality improvement programs and urged them to focus on problems of variability and their causes. One of Deming's (1982) main technical issues is No. 3 of his 14 points for management.

> *Cease dependence on inspection to achieve quality. Build quality into the product during the product development stage. Mass inspection cannot compensate for bad design. Do it right the first time so there is no need for rectification later.*

In 1951 JUSE established the Deming Prize which won national acclaim. Some authors have credited Deming with leading the Japanese quality revolution.

Two other American quality control experts who influenced the Japanese quality movement were J. M. Juran and A. Feigenbaum. Invited by JUSE, Juran arrived in 1954 and conducted seminars for top- and mid-level executives, whereas Feigenbaum, as head of quality at General Electric, had extensive contact with such companies as Toshiba and Hitachi. Some of his publications in the 1950s were translated into Japanese.

A key Japanese individual in this movement is the engineer Kaoru Ishikawa. As secretary general of JUSE, Ishikawa was founder of the magazine *Quality Control for the Foreman*, later renamed *FQC*. Its editorial encouraged the formation of the QC circles, groups of study and discussion headed by a foreman and participated in by his subordinate workers. Thousands of workers were given the so-called seven tools: check sheets, the Pareto chart, cause-effect diagrams, histograms, stratification, scatter plots and graphs (which include control charts); and QC circle conferences were promoted and became an annual event. In 1984 there were more than 180,000 such circles registered at the national level (Garvin, 1988). Ishikawa is also the author of several texts of quality control.

Another influential Japanese quality control expert is the engineer Genichi Taguchi. After some success in Japan, in 1980 Taguchi went to the US and visited AT&T Bell Laboratories to promote his *sui generis* approach to quality control. First his ideas found echo in few people, but as Nair (1992) points out, in the span of two years the interest grew, due perhaps to the widespread enthusiasm in the US for Japanese quality practices in the early 1980s, and a few individuals from AT&T, Ford, ITT, Xerox, and other places and organizations were instrumental in promoting the application of Taguchi ideas in industry. A few years later, Taguchi's quality engineering approach (baptised in the US as *Taguchi Methods*) has had a great impact in western industry and has generated much discussion and controversy among statisticians. The methods employed by Taguchi include fractional factorials and other experimental orthogonal arrays that are used in the

design phases of a product or process (this accounts for the term *off-line quality control*). In these experiments Taguchi suggests including *noise factors* that are environmental variables and other factors that are difficult or expensive to control and so are not to be controlled when the production is in progress. The idea is to find the conditions where the product or process is robust to variation of the *noise factors.* The quality approach to achieve robustness of products or processes by means of factorial experimentation is called *parameter design* or *robust design.* Parameter design intends to fulfil point No. 3 of Deming's plan quoted above by designing quality into the product and process prior the manufacturing stage. It incorporates the idea that a product may be made robust to the variations in the user's environment. Moreover the process which produces the product may be made robust to variations in materials, components and manufacturing before the normal production starts (for an extensive discussion of these ideas see Logothetis and Wynn, 1989).

The contribution to quality improvement that Taguchi has made with some of these ideas is widely recognized. However, some of the unusual statistical techniques that he has promoted for the analysis of factorial experiments are unfortunate. The use of signal to noise ratios, accumulation analysis, minute analysis, and unusual applications of analysis of variance have been widely criticized; either as inadequate, or as unnecessarily complicated and inefficient (see e.g. Box, 1988; Box and Bisgaard 1987; Box and Jones, 1986; Gunter, 1987; Hunter, 1987; Nair 1986, 1992).

Nowadays the ultimate goal of a good quality system is to build in quality into every product and process at the design stage and to follow up every stage. According to Logothetis and Wynn (1989) "building in quality at the design stage represents the latest phase in the evolution of quality systems". The use of Off-line quality control techniques can avoid re-design during production, inspection, and recall of a product after distribution; with the consequent reduction of costs.

An important element is the extensive and innovative utilisation of factorial experiments. As the number of factors involved is usually large, these experiments

are often carried out without replication, hence the term *unreplicated factorials.*

1.2 Brief historical review of the use of unreplicated factorials

In the 1920s, while working at Rothamsted Agricultural Station in the UK, R. A. Fisher developed the basic ideas of statistical design of experiments and introduced the concept of *factorial experimentation* where several factors may be studied simultaneously instead of experimenting with them one at a time.

When many factors are involved, replicates of a factorial experiment may become difficult to carry out because of the large number of experimental units needed. However, it was recognized that often it would still be valuable to use a single replicate or even a fractional replicate. In the latter every contrast can be associated with one or more treatment effects. Therefore, the use of a fractional factorial design usually requires the assumption that interactions of certain orders are negligible, so previous knowledge and careful planning are needed to ensure that the confounding patterns do not destroy the value of the experiment.

One early example of the utilization of a fractional factorial design in industrial applications out of the agricultural field, is an experiment reported by Tippett (1935) who employed a 125th fraction of a 5^5 to discover the causes of problems in a cotton spinning machine.

Important contributions have been made by Yates (1935) who presented a comprehensive approach to full (complete) factorials and some ideas of fractional factorials. Fisher (1942) systematically constructed classes of fractional factorials, where each factor had the same prime number of levels. Finney (1945) provided a formal approach to fractional factorial designs.

Plackett and Burman (1946) gave a theory to construct saturated designs[1] involving only two-level factors and provided initial settings and formulas to construct these designs for n observations with $n = 4k$; $k = 1, 2, \ldots, 25$; except for $n = 92$. For the cases where $n = 2^p$, Plackett and Burman designs are 2-level fractional factorials of resolution[2] III denoted as 2^p_{III}. General and special methods for constructing orthogonal arrays were provided by Rao (1946, 1947).

After the second world war, complete (full) factorials as well as fractional factorials and Plackett and Burman designs were used consistently in industrial applications. Among sources of numerous examples are: Davies (1956), Johnson and Leone (1964), Daniel (1976) and Box, et al. (1978).

Factorial experimentation also plays an important role in response surface methodology, where factorial experiments are used sequentially to explore the behaviour of the response over the factor space in order to find the conditions that optimize the response. A variety of special designs including centre points and star points have been suggested in this area. Examples of early works on response surfaces are Box (1954), Box and Hunter (1957) and Box and Draper (1959). This methodology is described in considerable detail in Box and Draper (1987).

During the last decade, the Japanese ideas of quality improvement have increased considerably the awareness of the value that factorial experimental designs have in industry. Many companies have incorporated the use of factorial designs in their quality systems. Taguchi's deficient statistical procedures are being overtaken while the good ideas prevail. Factorial experiments are used at the design phases of projects in order to achieve quality by minimizing variability. It is common to study the effect of the factors on the variance as well as on the mean; here again, adequate transformations widely studied in the West have shown their

[1] In a saturated experimental design, the number of observations is equal to the number of parameters to be estimated. These parameters are the main effects and the mean (see Dueker, 1988).

[2] The resolution of a two-level factor design is defined by the confounding or alias structure (see Box, et al., 1978). In a resolution III design no main effect is confounded (aliased) with another main effect, although main effects may be aliased with two-factor interactions.

value.

As these experiments are frequently conducted without replication, the analysis of unreplicated factorials has become a major concern. Considering that these experiments are frequently analyzed by non-statisticians, it is desirable to provide easy to use and understand procedures, preferably with graphical displays, to assess the significance of the estimates of the different effects involved in a particular experiment. This is the topic on which we will concentrate from the next chapter.

1.3 Example

In order to illustrate some of the characteristics of unreplicated factorials as well as a typical Taguchi experiment, a shortened version of a parameter design experiment presented by Byrne and Taguchi (1989) is analyzed here. The objective was to maximize the pull-off force of an elastomeric connector assembled to a nylon tube utilized in automotive engine components.

The idea in parameter design experiments is to find the optimal conditions (parameters) of *control factors* (factors easy to control) that minimise the performance variation of products and processes in the face of *noise factors* (factors difficult or expensive to control) that are controlled at the stage of development for experimentation.

In this case the experimenters identified four control factors (A-D) and three noise factors (E-G) that they felt could affect the pull-off force of the assembly. They decided to vary the control factors over three equally spaced levels and the noise factors, which are uncontrolled during normal operations, were controlled at two levels during the experiment. However Byrne and Taguchi (1989) observed that, in this particular example, the experimenters could have combined the three noise factors into one (N) at two levels representing two extremes of environmental conditions, which would have reduced the number of observations from 72 to 18, leading basically to the same conclusions. In order to simplify the exposition this

TABLE 1.1
Factors, levels and codes for the connector experiment

	Factors	Levels (codes)		
control				
A	Interference	Low (1)	Medium (2)	High (3)
B	Wall thickness	Thin (1)	Medium (2)	Thick (3)
C	Insertion depth	Shallow (1)	Medium (2)	Deep (3)
D	Percentage adhesive	Low (1)	Medium (2)	High (3)
Noise				
N	Environmental conditions	Level 1 (1)	Level 2 (2)	

shortened version of the experiment is presented here. The list of factors, their levels and codes is provided in Table 1.1.

Taguchi suggests the use of two orthogonal arrays, one for the control factors (inner array) and one for the noise factors (outer array). The two arrays are multiplied, i.e. for each factor-level combination in the inner array all factor-level combinations of the outer array are run to produce the variability due to noise factors. In Taguchi's context most interactions among control and among noise factors can often be assumed to be negligible. However interactions between control and noise factors are looked for; the basic idea is to use these interactions to identify appropriate settings of control factors at which the variability of the quality characteristic (response) is minimized, then use control factors that affect the mean but not the variability to bring the quality characteristic on to target.

In this experiment, the control factors were assigned to the orthogonal array $OA_9(3^4)$ (Taguchi's L_9) to generate a 3^{4-2} fractional factorial for the *inner array.* The outer array for the shortened experiment consists of two levels of the noise factor N. So only 18 out of the 72 observations originally considered are used. These are presented in Table 1.2.

Note that by using the array $OA_9(3^4)$, the experimenters are assuming that the effects of first and higher order interactions among the four control factors are all negligible, otherwise the experiment will be of little value.

In order to analyze *parameter design experiments* Taguchi suggests different

TABLE 1.2
Experimental layout and results for the connector experiment

control factors				Noise factor N	
A	B	C	D	1	2
1	1	1	1	15.6	19.1
1	2	2	2	15.0	21.9
1	3	3	3	16.3	20.4
2	1	2	3	18.3	24.7
2	2	3	1	19.7	25.3
2	3	1	2	16.2	24.7
3	1	3	2	16.1	21.6
3	2	1	3	14.2	24.4
3	3	2	1	16.1	28.6

performance measures called *signal-to-noise ratios* or SN ratios. For each setting of the control factors the SN is computed from the observations of the noise array. The specific expression for the SN ratio depends on the ideal value of the quality characteristic (see e.g., Taguchi, 1986; Bendell et al., 1989). When, as in this example, the quality characteristic (in this case pull-off force) is of the *larger-is-better* type, the signal to noise ratio recommended by Taguchi is:

$$SN_l = -10\log_{10}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{y_i^2}\right)\right] \tag{1.1}$$

where n is the number of settings in the outer array. In this example $n = 2$.

In order to decide which effects are significant Taguchi recommends estimating error variance by pooling the smallest mean squares, up to about half of the degrees of freedom, and then performing F-tests. Then the results are used to find the conditions of the control factors that maximize SN. This is supposedly a good compromise between minimizing the variability provoked by the noise factor and maximizing the mean response.

SN ratios have been criticized for being unnecessarily complicated and restrictive (see e.g. Box, 1988). Moreover, Shoemaker, et al. (1991) have shown the advantages of combining inner and outer arrays and modelling the response as a joint function of the control and noise factors.

On the other hand, the bias introduced by the nonstandard application of the analysis of variance recommended by Taguchi is well known. To show it, Box (1988) simulated five null experiments of size 15 and conducted an ANOVA by pooling the seven smallest sums of squares to test the remaining eight. The average percentage of effects found to be significant at 5% in the five random samples was 31%.

A natural alternative is to combine control and noise factors in a single design matrix and model the response instead of a performance measure. If we make the usual independence and normality assumptions, a standard linear model with dummy variables as predictors can be used. As each of the control factors is studied at three equally spaced levels, it is convenient to use orthogonal polynomials to obtain separately contrasts measuring linear and quadratic effects. Let A_l and A_q denote the linear and quadratic effects of factor A, and consider similar notation for factors B, C and D, and let N denote the effect of the noise factor. Then there are eight main effects of interest for the control factors and one for the noise factor. But the experimenters were also interested in interactions between noise and control factors. As the noise factor was experimented at its two levels for each setting of the control factors, it is possible to study the first order interactions between the noise factor and each component of the control factors. These are denoted by putting together the names of the elements involved, e.g. the interaction between C_q and N is denoted C_qN. All these effects are considered in the following regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\mu} + \mathbf{e} \tag{1.2}$$

where $\mathbf{y}$ is the 18×1 vector of observations of pull-off force, $\mathbf{X}$ is the 18×18 matrix of predictors including a column of 1s, $\boldsymbol{\mu} = (\mu_0, \mu_1, \ldots, \mu_{17})'$ is a vector of unknown coefficients, and $\mathbf{e}$ is an 18×1 vector of independent random variables from a normal distribution with mean zero and variance σ^2.

Table 1.3 shows the values of $\mathbf{y}$, $\mathbf{X} = (\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{17})$ and the names of the effects of interest. The latter are presented at the top of the columns of $\mathbf{X}$ that they are associated with.

TABLE 1.3
Effect names, $\mathbf{X}$ matrix and pull-off force data $\mathbf{y}$ for the connector experiment

	A_l	A_q	B_l	B_q	C_l	C_q	D_l	D_q	N	A_lN	A_qN	B_lN	B_qN	C_lN	C_qN	D_lN	D_qN	
$\mathbf{x}_0$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$	$\mathbf{x}_7$	$\mathbf{x}_8$	$\mathbf{x}_9$	$\mathbf{x}_{10}$	$\mathbf{x}_{11}$	$\mathbf{x}_{12}$	$\mathbf{x}_{13}$	$\mathbf{x}_{14}$	$\mathbf{x}_{15}$	$\mathbf{x}_{16}$	$\mathbf{x}_{17}$	$\mathbf{y}$
1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	15.6
1	-1	1	0	-2	0	-2	0	-2	-1	1	-1	0	2	0	2	0	2	15.0
1	-1	1	1	1	1	1	1	1	-1	1	-1	-1	-1	-1	-1	-1	-1	16.3
1	0	-2	-1	1	0	-2	1	1	-1	0	2	1	-1	0	2	-1	-1	18.3
1	0	-2	0	-2	1	1	-1	1	-1	0	2	0	2	-1	-1	1	-1	19.7
1	0	-2	1	1	-1	1	0	-2	-1	0	2	-1	-1	1	-1	0	2	16.2
1	1	1	-1	1	1	1	0	-2	-1	-1	-1	1	-1	-1	-1	0	2	16.1
1	1	1	0	-2	-1	1	1	1	-1	-1	-1	0	2	1	-1	-1	-1	14.2
1	1	1	1	1	0	-2	-1	1	-1	-1	-1	-1	-1	0	2	1	-1	16.1
1	-1	1	-1	1	-1	1	-1	1	1	-1	1	-1	1	-1	1	-1	1	19.1
1	-1	1	0	-2	0	-2	0	-2	1	-1	1	0	-2	0	-2	0	-2	21.9
1	-1	1	1	1	1	1	1	1	1	-1	1	1	1	1	1	1	1	20.4
1	0	-2	-1	1	0	-2	1	1	1	0	-2	-1	1	0	-2	1	1	24.7
1	0	-2	0	-2	1	1	-1	1	1	0	-2	0	-2	1	1	-1	1	25.3
1	0	-2	1	1	-1	1	0	-2	1	0	-2	1	1	-1	1	0	-2	24.7
1	1	1	-1	1	1	1	0	-2	1	1	1	-1	1	1	1	0	-2	21.6
1	1	1	0	-2	-1	1	1	1	1	1	1	0	-2	-1	1	1	1	24.4
1	1	1	1	1	0	-2	-1	1	1	1	1	1	1	0	-2	-1	1	28.6

The least squares estimate of μ is

$$\hat{\mu} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{1.3}$$

As $\mathbf{X}$ is orthogonal, each element of $\hat{\mu}$ can be computed using only the corresponding vector of $\mathbf{X}$,

$$\hat{\mu}_i = \frac{\mathbf{x}_i'\mathbf{y}}{\mathbf{x}_i'\mathbf{x}_i}, \; i = 0, \ldots, 17. \tag{1.4}$$

Note that $\hat{\mu}_1, \ldots, \hat{\mu}_{17}$ are contrasts that estimate the 17 effects of interest $A_l, \ldots, D_qN$. In fact each of these contrasts estimates a string of effects which are confounded with the effect of interest but that have been assumed to be negligible. The problem now is to decide which, if any, of the 17 contrasts are active (i.e. have nonzero mean). According to experience only a small proportion of these contrasts are expected to be active. A popular analysis for unreplicated factorials is by visually inspecting a half-normal plot of the contrasts. This technique was proposed by Daniel (1959) to interpret 2^{p-q} factorials. However this procedure can be used also for experiments including factors with more than two levels by scaling the full set of orthogonal contrasts so that all of them have the same variance. Consider the arbitrary contrast $\hat{\mu}_i$ $(i > 0)$ in (1.4). Its variance is

$$V(\hat{\mu}_i) = V\left(\frac{\mathbf{x}_i'\mathbf{y}}{\mathbf{x}_i'\mathbf{x}_i}\right) = \frac{\sigma^2}{\mathbf{x}_i'\mathbf{x}_i}. \tag{1.5}$$

Thus $\hat{\mu}_i$ can be scaled or standardized by multiplying it by $(\mathbf{x}_i'\mathbf{x}_i)^{1/2}$, to give the standardized contrast

$$U_i = \frac{\mathbf{x}_i'\mathbf{y}}{\sqrt{\mathbf{x}_i'\mathbf{x}_i}} \tag{1.6}$$

whose variance is σ^2.

Note that the standardization could have been done in (1.2) by multiplying each element of the column $\mathbf{x}_i$ of $\mathbf{X}$ by $(\mathbf{x}_i'\mathbf{x}_i)^{-1/2}$ $(i = 1, \ldots, 17)$.

A half-normal plot is constructed by plotting the $n - 1 = m$ ordered absolute contrasts against the approximate expected order statistics for a sample of size m from the half-normal distribution. If all contrasts are null, the full collection should

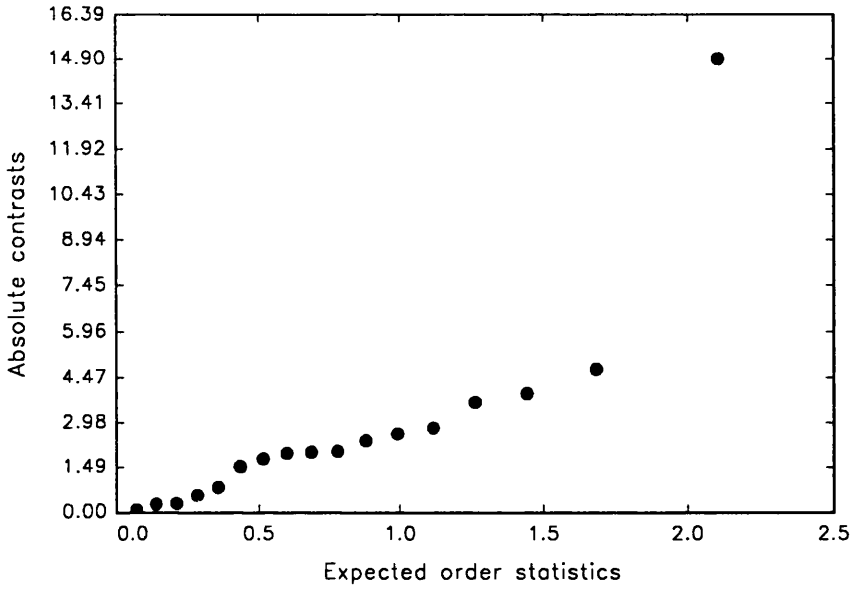


Figure 1.1: Half-normal plot of the standardized contrasts for the elastomeric connector example

fall near a straight line through the origin; the slope of the line is the standard deviation of the null contrasts. If a small number of real (active) contrasts are present, the corresponding absolute values should plot well off the straight line.

The half-normal plot of the 17 standardized contrasts of the elastomeric connector example is shown in Figure 1.1. Only the largest absolute contrast, which measures the effect of the noise factor N, is obviously significant. Therefore there is no evidence that any particular setting of the control factors, within the ranges studied, could either dampen the effect of the noise factor, or affect the mean pull-off force.

The most frequently noted disadvantage of making conclusions based on visual inspections is that these procedures are somewhat subjective. Although in this example most people surely would coincide in deeming only the estimated contrast measuring the effect of factor N to be significant, in other not so clear situations two people judging the same plot might well arrive at different conclusions. Consequently, it has been suggested that these plots should be supplemented with a

more formal method.

In the next chapter a literature review of methods for assessing the significance of estimated contrasts for unreplicated factorials is presented.

Chapter 2

Methods for assessing the significance of contrasts for unreplicated factorials

2.1 Assumptions and description of the problem

The attention will be restricted to unreplicated factorial experiments that can be described by the following linear model.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\mu} + \mathbf{e} \tag{2.1}$$

where $\mathbf{y}$ is an $n \times 1$ vector of independent observations at n experimental conditions; $\mathbf{X}$ is an $n \times n$ orthogonal matrix with rank n, whose first column $\mathbf{x}_0$ is a column of 1s and the rest of the columns $\mathbf{x}_1 \ldots \mathbf{x}_m$ $(m = n - 1)$ are assumed to have been standardized so that $\mathbf{x}_i'\mathbf{x}_i = 1$, $i = 1, \ldots, m$; $\boldsymbol{\mu}$ is a vector of unknown coefficients, say $\boldsymbol{\mu}' = (\mu_0, \mu_1, \ldots, \mu_m)$; and $\mathbf{e}$ is a vector of n random variables independent and identically distributed according to the normal probability law with mean 0 and common variance σ^2. We shall use the standard notation and

refer to the elements of $\mathbf{e}$ as iid $N(0,\sigma^2)$.

The least squares estimators of $\mu_1, \ldots, \mu_m$ are the orthogonal contrasts

$$\hat{\mu}_i = U_i = \mathbf{x}_i'\mathbf{y}, \quad i = 1, \ldots, m. \tag{2.2}$$

Each one of the squares $Z_i = (\mathbf{x}_i'\mathbf{y})^2$, $i = 1, \ldots, m$ has one degree of freedom and it may be verified that

$$\mathbf{y}'\mathbf{y} - n\bar{y}^2 = Z_1 + \cdots + Z_m. \tag{2.3}$$

A common assumption for these experiments is that of *effect sparsity*, the idea that only a small number of contrasts are expected to have nonzero means, those being called real or active contrasts.

The relevant problem is to infer, on the basis of one set of observations $y_1, \ldots, y_n$, which, if any, of the orthogonal contrasts, have mean different from zero.

Note that (2.1) covers unreplicated two-level fractional factorials and Plackett and Burman designs for which the column vectors of the design matrix $\mathbf{X}$ do not need to be standardized. The assumption of standardized columns (with the exception of column $\mathbf{x}_0$) has been included in order to cover also fractional factorials and other experimental orthogonal arrangements with factors with more than two levels and those combining factors with different levels. Care should be taken in choosing the convenient standardized orthogonal contrasts to separate the desired comparisons. When the levels are equally spaced points of a quantitative characteristic, standardized orthogonal polynomials are specially useful. Orthogonal comparisons for different kind of factors can be found, e.g., in Cochran and Cox (1957).

2.2 Traditional pooling

Perhaps the most traditional practice for assessing the significance of contrasts in unreplicated factorials is to pool some arbitrary number of sums of squares into

an estimate of error variance and then use the F-test in the usual ANOVA way.

It is common to pool (when available) high-order interactions on the grounds that effects of this kind are often negligible. According to Davies (1956), in order not to bias the procedure, one should decide before the experiment is performed which interactions are likely to be null so their sums of squares are to be used to estimate the error variance.

Although this method is often considered as suitable practice, it is frequently violated as soon as the data are seen. If, for example, one of the high-order interactions happens to be large the experimenter might be led to exclude it in order to obtain an "improved" estimate of error, biasing the procedure in this way. For further common violations of this procedure see Daniel (1976, p.72).

One of the disadvantages of this procedure is that the experimenters are not always able to decide a priori which contrasts should be used to estimate error.

2.3 60% pooling

Berk and Picard (1991) investigated an approach based on the standard ANOVA table, but without involving tests based on the F distribution. In the absence of an independent estimate of error variance, they suggest pooling 60% of the smallest mean squares into a baseline and forming F-like ratios of large mean squares with the mean square for the baseline as denominator. By simulation of the empirical distribution of such ratios they obtained critical values for testing individual contrasts at significance levels of 0.05, 0.01 and 0.001 for experiments with $n = 8,\ 12,\ 16,\ 20$ and 32 observations.

Berk and Picard carried out a simulation study involving different numbers of real effects of several sizes and compared some operating characteristics of this method with those proposed by Zahn (1975a), Box and Meyer (1986), Voss (1988) and Lenth (1989), all of which are described or commented on below. Berk and

Picard concluded that based on the situations simulated, all the procedures seemed to perform "quite comparably".

2.4 t margins of error

Lenth (1989) proposed a technique based on the idea of using the usual t-test. Let U_i, $i = 1, \ldots, m$, be the contrasts as defined before, and let

$$S_0 = 1.5 \cdot \underset{i}{\text{median}} |U_i|. \tag{2.4}$$

The *pseudo standard error* (PSE) of the contrasts is defined to be

$$\text{PSE} = 1.5 \cdot \underset{|U_i|<2.5S_0}{\text{median}} |U_i|. \tag{2.5}$$

Lenth shows that, under the assumption of effect sparsity, PSE is a good estimate of σ.

Fitting the empirical distribution of PSE^2 by scaled chi-squared distributions, Lenth considers the appropriate number of degrees of freedom to be $d = m/3$ so, with the tables provided in Lenth (1989), approximate confidence intervals may be computed in the natural way.

It is suggested to compute a *margin of error* (ME) as

$$\text{ME} = t_d^{(.975)} \cdot \text{PSE} \tag{2.6}$$

where $t_d^{(.975)}$ is the 0.975th quantile of a t distribution with d degrees of freedom. And taking into account that several inferences are being made simultaneously it is also suggested to compute a *simultaneous margin of error* (SME) as

$$\text{SME} = t_d^{(\gamma)} \cdot \text{PSE} \tag{2.7}$$

where $\gamma = (1 + 0.95^{1/m})/2$.

If $|U_i| > \text{SME}$ the respective contrast is declared to be active; if $|U_i| < \text{ME}$ the respective contrast is declared to be null. Those in between are indeterminate and become a matter for the experimenter's judgement.

Lenth suggested presenting this information in a bar graph in which the magnitude and sign of each contrast are displayed along with reference lines at $\pm$ME and at $\pm$SME.

The tables provided allow the computation of ME and SME for experiments with $m = 7$, 15, 31, 63, 127 and 255.

Stephenson (1991) presented a computer program for the application of this procedure.

2.5 Inference with half-normal plots

Along with the visual inspection of a half-normal plot, Daniel (1959) suggested a more formal procedure for assessing sequentially the significance of the largest contrasts. Let $U_1, \ldots, U_m$ denote the orthogonal standardized contrasts from an experimental array with $n = m + 1$ observations. Under the assumptions stated in Section 2.1 $U_1, \ldots, U_m$ are independent normally distributed random variables with respective means $\mu_1, \ldots, \mu_m$ and unknown variance σ^2. The problem is to infer on the basis of one set of observed values of $U_1, \ldots, U_m$ which, if any, of the means are nonzero.

Note that if Y is $N(0, \sigma^2)$, $X = |Y|$ is a half-normal random variable with parameter σ^2 and $P(X \leq \sigma) = 0.683$.

Let $V_{(1)}, \ldots, V_{(m)}$ denote the ordered absolute contrasts. The test statistics suggested by Daniel are

$$T_k = \frac{V_{(k)}}{V_{(a)}}, \; k = m, m-1, \ldots, a+1 \tag{2.8}$$

with $V_{(a)}$ such that $(a-0.5)/m$ is most nearly 0.683. As is standard, Daniel approximated the expected order statistics by the percentiles $(i - 0.5)/m$, $i = 1, \ldots, m$, thus $V_{(a)}$ is the order statistic that is expected to be closest to σ. Simulating the empirical distributions of (2.8), Daniel obtained critical values and presented

special grids for the cases m = 15, 31, 63 and 127 (a = 11, 22, 44, and 88) including *guard rails* to detect up to $r = m - a$ real contrasts. These critical values were corrected by Zahn (1975a) who also investigated alternative methods (Zahn, 1975b).

Daniel's procedure is as follows.

i) Compute $V_{(i)}/V_{(a)}$; $i = 1, \ldots, m$ and plot them on the appropriate half-normal grid.

ii) Examine $V_{(m)}/V_{(a)}$. If it is plotted above the corresponding guard rail, declare the contrast significant and examine $V_{(m-1)}/V_{(a)}$. Proceed until a value plotted beneath the guard rail is encountered or until $V_{(a)}/V_{(a)}$ is met. Declare that contrast and all the smaller ones insignificant.

Regarding Daniel's critical values, Zahn argued that in order to control the *probability of a nonzero family error rate* (PER)[1], using the terminology of Miller (1966), to be no greater than α in Daniel's procedure, the critical values should be as follows.

Let $V_{h,m}$ denote the hth order statistic from a sample of size m, and let α denote the probability of at least one false positive in a null experiment. Zahn's α-level critical values $c_m, c_{m-1}, \ldots, c_{m-r+1}$ are such that

$$P\left(\frac{V_{m-i,m-i}}{V_{a,m}} < c_{m-i} \mid \mu_1 = \ldots = \mu_{m-i} = 0\right) = 1 - \alpha, \quad i = 0, \ldots, r-1. \qquad (2.9)$$

Zahn (1975a) computed the set of critical values defined by (2.9) for α = 0.05, 0.20 and 0.40 revising in this way those given by Daniel (1959), and made other minor improvements to Daniel's procedure.

With these critical values, Daniel's procedure will give (in the null situation) one false positive if and only if the largest contrast is declared significant, i.e. if $V_{m,m}/V_{m,a} > c_m$. Hence to control the PER = α in a null experiment c_m would

[1]When inferences about p parameters are carried out, one statement is made for each parameter. The PER is defined as the probability of at least one incorrect statement in the family of p statements

suffice. The set of critical values c_{m-i} for $i = 1, \ldots, r-1$ are used during the procedure for determining which $\mu's$ will be considered different from zero when the null case has been rejected. The smaller these critical values, the higher the detection rates in non-null situations but the larger the expected number of false positives. With the set of critical values defined by (2.9) Zahn intended to control the $PER \leq \alpha$ even in non-null situations. In fact this is analogous to the multiple modulus tests in the sense of Newman and Keuls (see Miller, 1966).

By means of simulation Zahn (1975b) compared empirically the operating characteristics of several variations of Daniel's corrected procedure for the case $m = 15$, and concluded that the method that he called *version S* (and we will call Zahn's method) seemed superior at least for $m = 15$, and he provided critical values for significance levels of 0.05, 0.20 and 0.40. Zahn's method differs from Daniel's corrected method with respect to the denominator of the test statistics used during the procedure. Let $W_{1:m}, \ldots, W_{m:m}$, denote the expected values of the order statistics of a sample of size m from the standard ($\sigma = 1$) half-normal distribution. Zahn's suggested that better results may be obtained if the statistics (2.8) are replaced by

$$T'_k = \frac{V_{(k)}}{\mathrm{SL}(a,m)}, \; k = m, m-1, \ldots, a+1 \tag{2.10}$$

where $\mathrm{SL}(a,m)$ is the slope of the ordinary least squares regression through the origin of V on W using the points $(W_{i:m}, V_{(i)})$, $i = 1, \ldots, a$, i.e.

$$\mathrm{SL}(a,m) = \frac{\sum_{i=1}^{a} V_{(i)} W_{i:m}}{\sum_{i=1}^{a} W_{i:m}^2}. \tag{2.11}$$

The idea that motivated this alternative is that in the null case $\mathrm{SL}(a,m)$ estimates σ and the ratios $T'_i = V_{(i)}/\mathrm{SL}(a,m)$ estimate $W_{i:m}$, $i = 1, \ldots, m$.

A limitation of these methods is that they assume that at most $r = 4$ real contrasts may be present in an experiment of size 15.

Voss (1988) studied another variant of Daniel's procedure using a generalized modulus-ratio statistic, but the detection rates obtained were poorer than those of Zahn's recommended version.

2.6 A Bayesian approach

Box and Meyer (1986) developed a Bayesian procedure. Starting with an apriori probability for a contrast to be active (real) and assuming a normal distribution for the active contrasts, a posterior probability that it is active is computed for each contrast.

Let $U_1, \ldots, U_m$ be the estimated contrasts. A random effects model is assumed, that is, for an inert (null) contrast $U_i = e_i$ and for an active contrast $U_i = \eta_i + e_i$, where the e_i are iid $N(0, \tau^2)$ and the η_i are iid $N(0, \tau_\eta^2)$.

Assume that a contrast is active with probability α and let $k^2 = (\tau^2 + \tau_\eta^2)/\tau^2$, then $U_1, \ldots, U_m$ are iid from the scale contaminated normal distribution denoted by

$$(1 - \alpha)N(0, \tau^2) + \alpha N(0, k^2\tau^2). \tag{2.12}$$

Bayes' theorem is used to compute the posterior probability that a single estimated contrast U_i comes from the distribution $N(0, k^2\tau^2)$ given τ. Then numerical integration over the posterior distribution of τ given $U_1, \ldots, U_m$ is used to compute the unconditional posterior probability p_i that a contrast U_i is active.

Two parameters are required: α, the probability of an active contrast, and k the inflation factor of the standard deviation produced by an active contrast.

From information on the behaviour of several unreplicated fractional factorials reported in the literature, Box and Meyer took α from the average proportion of contrasts declared significant by the authors, which was about 0.20; and k^2 from the average of the ratios of mean squared significant contrasts over the mean squared null contrasts, which gave k about 10.

The bias introduced by restricting attention to published examples is recognized, but Box and Meyer argued that the conclusions to be drawn from the analysis are usually insensitive to moderate changes in α and k.

They suggest presenting the results in a graphical form known as a Bayes plot

accompanied by normal plots.

2.7 The chain-pooling method

This is a quite complicated procedure presented by Holms and Berrettoni (1969). It is based on a modification of the Cochran (1941) method for testing homogeneity of variance. It does not require the effect sparsity assumption as only a few contrasts $s \geq 1$ are assumed to be null.

Let $Z_1, \ldots, Z_m$ be the ordered mean squares of the contrasts; the test statistics are

$$R_j = jZ_j/(Z_1 + \cdots + Z_j) \quad j = 2, \ldots, m. \tag{2.13}$$

Chain pooling is carried out in two steps:

i) Testing starts with $j = s + 1$ and a large nominal significance level α_p (e.g. 0.25 or 0.50). Proceeding sequentially, all Z_j testing nonsignificant remain in the denominator until some Z_j is significant. As α_p is large, this is expected to occur early in the chain.

ii) A new significance level $\alpha_f < \alpha_p$ (e.g. 0.01 or 0.05) is imposed and the same Z_j is tested at level α_f. If Z_j is significant again, all Z_k $(k > j)$ are concluded to be significant. If Z_j is not significant, continue testing $Z_{j+1}, \ldots$ until some Z_k is significant, but only the first $j - 1$ mean squares, plus the one being tested, are pooled into the denominator. The test statistic for the kth mean square is then

$$R_j = jZ_k/(Z_1 + \cdots + Z_{j-i} + Z_k). \tag{2.14}$$

The selection of the values (s, α_p, α_f) is called the *statistician's strategy.*

Holms and Berretoni presented a summary of an extensive Monte Carlo study made by Holms (1966) about the operating characteristics of different strategies for the 2^4 experiment. The aim was to find robust strategies that produce reasonably

small type I and type II error rates for different numbers and sizes of real contrasts. Tables and charts were given in order to help the statistician to choose his/her own strategy.

The overall recommended procedure starts with $s = 1$ and a rough guess about the number of null contrasts η, this leads to an estimate $\hat{\eta}$. A second chain pooling is then performed by selecting s in accordance with $\hat{\eta}$ and using the charts to select α_p and α_f. One limitation of this procedure is that it is not feasible to predetermine a bound on the probability of a type I error.

2.8 Testing normality and outliers

Benski (1989) presented a sequential method that uses Olson's (1979) version of the Shapiro-Wilk tests of normality, supplemented by a test of outliers. Let $U_{(1)}, \ldots, U_{(m)}$ be the ordered contrasts; Olson's test statistic is

$$T = \frac{\sum(w_i U_{(i)})^2}{\sum w_i^2 \sum(U_{(i)} - \bar{U})^2}, \tag{2.15}$$

where w_i are the approximate expected values of $U_{(i)}$, $i = 1, \ldots, m$, and $\bar{U} = \sum U_i/m$.

To detect the presence of outliers a robust test described by Hoaglin (1983) is suggested.

Testing starts with T which gives a significance level P_1. If P_1 is small, the outliers test is performed giving a significance level P_2. P_1 and P_2 are then combined by Fisher's combination method (Stephens, 1986). This gives a significance level P_c which is obtained using the fact that, under the null hypothesis, $2\ln(1/P_1P_2)$ has approximately a χ^2 distribution with 4 degrees of freedom. If P_c is small the largest absolute contrast is removed and T is computed with the remaining values. The procedure continues until either P_1 or P_c is not small. The contrasts removed during the procedure are declared significant.

2.9 A procedure based on homogeneity of mean squares

Bissell (1989) proposed "tentatively" a procedure based on a modification by Box (1949) of Bartlett's (1937) test for variance homogeneity. Let $Z_1, \ldots, Z_m$ be the mean squares of the m contrasts, the test statistic is

$$\mathrm{B} = \ln\left(\frac{1}{m}\sum_{i=1}^{m} Z_i\right) - \frac{1}{m}\sum_{i=1}^{m} \ln Z_i. \tag{2.16}$$

The procedure consists of computing B and comparing it with critical values given in Bissell's paper. If B is significant, the largest mean square is deleted and B is computed with the remaining mean squares. One then proceeds until B is not significant. The contrasts corresponding to the mean squares removed during the procedure are considered to be active.

2.10 A method based on the coefficient R^2

Hamada and Wu (1991) suggested the following procedure based on the statistic R^2. A straight line is fitted through the smallest contrasts (over 50% in number) and its R^2 value is observed. Then the remaining contrasts are added one at a time, fitting a new line each time and looking for an R^2 drop of 0.1 or more from the last fitted line. At the first such drop, the currently added contrast and the remaining larger contrasts are identified as significant.

2.11 Discussion

At this stage no best approach to the analysis of unreplicated factorials is apparent. The operating characteristics of some procedures have been compared by simulation of a variety of situations in several studies. Zahn (1975b) compared

different methods based on Daniel's half normal plots for experiments with 15 contrasts. Berk and Picard (1991) obtained critical values for the 60% method giving the same probabilities of error as those used for the procedures given by Box and Meyer (1986), Lenth (1991) and Zahn (1975a), and presented results of operating characteristics for experiments with 7 and 15 contrasts.

Taking the case of 15 contrasts, which has been studied most extensively, the following points summarize the main features of the results obtained so far. In all cases the operating characteristics were obtained from 1000 samples simulated.

1.- Zahn's recommended version for half-normal plots showed larger detection rates (power) than Daniel's corrected version, especially in situations with 4 real contrasts where the difference was "often by as much as 0.07" (Zahn, 1975b).

2.- In all the situations in which they were compared, the detection rates obtained by Zahn's method (Zahn, 1975b; Berk and Picard, 1991) are either equal to or greater than those obtained by the 60% procedure. In one of the situations the difference obtained was as large as 0.11. Zahn's procedure however, is restricted to testing no more than 4 contrasts, so for situations with 5 or 6 real contrasts it is ineffective.

3.- The detection rates of the procedure suggested by Box and Meyer (1986) exceeded those of the 60% method in situations with 2 and 3 real contrasts while in situations with 6 real contrasts the 60% method seems to be more powerful.

4.- Regarding comparisons with the method proposed by Lenth (1989), the 60% procedure seems more powerful in situations with 1 and 2 real contrasts while Lenth's appeared more powerful when there were 4 and 6 real contrasts.

Although these results comprise important contributions, they are not conclusive and may lead to contradictory interpretations. For instance, although the effect sparsity assumption, which is based on the experience accumulated throughout the years, is generally accepted, there is no general agreement about the degree of effect sparsity that should be considered. Thus if it is thought that in experi-

ments with 15 contrasts there will be no more than 4 real contrasts, Zahn's method will be preferred over any other procedure. But if it is thought that there may be 5 or 6 then Zahn's method should be avoided.

Similarly, the lack of knowledge about the numbers and sizes of real contrasts that are likely to occur prevents the definite preference of any of the other procedures compared.

It would be possible to extend the simulation studies to a richer range of situations. However this would imply making more assumptions about the situations that are likely to arise in practice, and will not help very much to discriminate among the methods.

Instead, it was decided to study the performance of the four procedures mentioned above on real data using for this purpose a substantial number of experiments published in the statistical literature and other sources. The four procedures are the ones proposed by Box and Meyer (1986), Berk and Picard (1991), Lenth (1989), and Zahn (1975a).

Although the selection of these four procedures is somewhat arbitrary, they have been chosen according to the following criteria: once implemented, they are easy to apply and interpret; the interpretation can be effectively helped by graphical displays; and they seem to have been accepted by the statistical community as the most promising.

The exercise of comparing these methods on real data cannot be expected to answer all the questions that have not been answered by simulation studies. In fact, since the numbers of null and real contrasts and the sizes of the latter ones will remain unknown, it cannot not be expected that, at the end, it will be possible to make any statement that implied their previous knowledge.

What is expected instead, is that the problems that usually arise in practice will appear, showing for example, the limitations of one procedure compared with the other ones. And hopefully, some other points that arise only when one is

dealing with real data will emerge.

In practice the model assumptions are rarely wholly satisfied. For example, it would be naive to think that all the assumed null contrasts are actually null, and although in most cases they may be small enough to be considered as being negligible for practical purposes, those small nonzero contrasts might affect the effectiveness of the procedures in different ways.

It is expected that in most of the examples the procedures will coincide, so attention will be paid to those cases where they do not agree, looking for clues to the possible reasons. At the end, it will be interesting to analyse the global performance of each procedure compared with the others. The results of the simulation studies described above might be useful as a reference background.

It is important that the procedures are applied using the same probability error rates, either for individual or multiple inferences, but also it is desirable that they are applied in the way they have been proposed. Thus it may require compromise between these two ideas to make the procedures at least roughly comparable and try to keep as close as possible to the original proposals. The examples selected as well as the implementation of the methods are presented in the next chapter.

Chapter 3

Examples and implementation of selected methods

In the discussion at the end of the last chapter it was decided to investigate four methods by applying them to a substantial number of real experiments. The methods are: The Bayesian procedure suggested by Box and Meyer (1986) which will be referred to as BM; the 60% method by Berk and Picard (1991) which will be referred to as BP; Lenth's (1989) method which will be referred to as LE; and Zahn's version of Daniel's procedure for half-normal plots, which will be referred to as ZA.

The results of the search for real examples of unreplicated factorials are presented in Section 3.1. Some technical details of the implementation and supplementation of the four procedures to make them readily available and applicable to all sizes of experiments collected are exhibited in Section 3.2. In Section 3.3 graphical displays of the four methods are illustrated with an example.

3.1 Examples collected

A search for examples of unreplicated factorials in the statistical literature and other sources was carried out. There is a great variety of examples but the most numerous are of the type 2^{p-q}, with $p - q = 3, 4, 5, 6$. Thus it was decided to use two-level factor examples to study the performance of the four procedures selected. The examples were carefully scrutinized in order to avoid repetitions and making the best effort not to include simulated data. The final list contained 102 examples of factorial experiments distributed as follows.

- **30 examples with 7 contrasts.** Design and, within brackets, number of examples with that design.

  2^3 (12), 2^{4-1}_{IV} (6), 2^{5-2}_{III} (5), 2^{6-3}_{III} (2), 2^{7-4}_{III} (5).

- **54 examples with 15 contrasts.**

  2^4 (18), 2^{5-1}_{V} (6), 2^{6-2}_{IV} (6), 2^{7-3}_{IV} (5), 2^{8-4}_{IV} (10), 2^{9-5}_{III} (4), 2^{10-6}_{III} (2), 2^{15-11}_{III} (2), Plackett and Burman (1).

- **9 examples with 31 contrasts.**

  2^5 (6), 2^{6-1}_{VI} (1), 2^{8-3}_{IV} (1), 2^{9-4}_{IV} (1).

- **9 examples with 63 contrasts.**

  2^6 (3), 2^{8-2}_{V} (5), 2^{16-10}_{III} (1).

The complete list of examples is presented in Appendix A, where they have been labelled as [1], [2], ..., [102] for reference purposes.

The set of contrasts for each example was computed and all the sets were organized in a data base for their manipulation and processing.

3.2 Implementation of selected methods

The four methods were implemented in the matrix programming language GAUSS. The characteristics and amount of the mathematical and computing work involved varied from method to method as is shown below.

3.2.1 Bayesian approach (BM)

In the Bayesian method proposed by Box and Meyer (1986) outlined in the previous chapter, the standardized contrasts $U_1, \ldots, U_m$ are considered as a sample of independent random variables with density

$$\begin{aligned} f(U_i|\xi_i, \tau) &= \xi_i \left[\frac{1}{k\tau\sqrt{2\pi}} \exp\left(\frac{-U_i^2}{2k^2\tau^2} \right) \right] \\ &+ (1-\xi_i) \left[\frac{1}{\tau\sqrt{2\pi}} \exp\left(\frac{-U_i^2}{2\tau^2} \right) \right], \ (i = 1, \ldots, m) \end{aligned} \tag{3.1}$$

where ξ_i is a random variable that takes the value 1 when the contrast i is active and 0 when it is inert, and for which the prior probability distribution is

$$f(\xi_i) = \alpha^{\xi_i}(1-\alpha)^{(1-\xi_i)}, \ \ \xi_i = 0, 1. \tag{3.2}$$

Direct application of Bayes' theorem leads to the posterior probability that a particular contrast i is active given U_i and the standard deviation τ

$$\Pr[\xi_i = 1|U_i, \tau] = \frac{\alpha \left\{ \frac{1}{k\tau\sqrt{2\pi}} \exp\left(\frac{-U_i^2}{2k^2\tau^2} \right) \right\}}{\alpha \left\{ \frac{1}{k\tau\sqrt{2\pi}} \exp\left(\frac{-U_i^2}{2k^2\tau^2} \right) \right\} + (1-\alpha) \left\{ \frac{1}{\tau\sqrt{2\pi}} \exp\left(\frac{-U_i^2}{2\tau^2} \right) \right\}}. \tag{3.3}$$

Note that this is independent of the remaining $U_1, \ldots, U_{i-1}, U_{i+1}, \ldots, U_m$.

To compute the unconditional posterior probability that the contrast T_i is active, the parameter τ must be integrated out of (3.3) over its posterior distribution $f(\tau|\mathbf{U})$, where $\mathbf{U} = (U_1, \ldots, U_m)$.

The prior distribution for τ used by Box and Meyer is

$$f(\tau) \propto \frac{1}{\tau}. \tag{3.4}$$

Using Bayes' theorem again, the posterior distribution of τ given the contrasts $\mathbf{U}$, is

$$f(\tau|\mathbf{U}) = \frac{f(\tau, \mathbf{U})}{f(\mathbf{U})} \tag{3.5}$$

where

$$\begin{aligned} f(\tau, \mathbf{U}) = f(\tau)f(\mathbf{U}|\tau) &= \frac{1}{\tau}\prod_{j=1}^{m}\left\{\frac{1}{\tau\sqrt{2\pi}}\left[\frac{\alpha}{k}\exp\left(\frac{-U_j^2}{2k^2\tau^2}\right)\right.\right. \\ &+ \left.\left.(1-\alpha)\exp\left(\frac{-U_j^2}{2\tau^2}\right)\right]\right\} \end{aligned} \tag{3.6}$$

and

$$f(\mathbf{U}) = \int_0^\infty f(\tau, \mathbf{U})d\tau. \tag{3.7}$$

Finally, removing the conditioning on τ the posterior probability that an individual contrast i is active given $\mathbf{U}$ is

$$\begin{aligned} \Pr[\xi_i = 1|\mathbf{U}] &= \int_0^\infty \Pr[\xi_i = 1|U_i, \tau]f(\tau|\mathbf{U})d\tau \\ &= \frac{\int_0^\infty \Pr[\xi_i = 1|U_i, \tau]f(\tau, \mathbf{U})d\tau}{f(\mathbf{U})}. \end{aligned} \tag{3.8}$$

Rather than expanding the integrands in (3.8), the integrals are computed by numerical integration as suggested by Box and Meyer (1986), so the main computational work was the writing of the integrand's functions. The GAUSS routine used is INTQUAD1 which uses Gauss-Legendre quadrature.

3.2.2 Zahn's method (ZA)

As Zahn (1975a) presented critical values for ZA for the case of 15 contrasts only, it is necessary to compute critical values for the cases with 7, 31 and 63 contrasts.

The computation of the denominator SL of the statistic (2.10) requires the expected values of order statistics of samples of the standard half-normal distribution (HEOS). Since these are tabulated only for small sample sizes (Zahn, 1975a),

an expression to compute suitable close approximations for any sample size is obtained below.

A random variable is said to be standard half-normal if its probability distribution function is

$$g(x) = \begin{cases} \sqrt{2/\pi}\, \exp\{-x^2/2\}, & \text{for } x \geq 0 \\ 0, & \text{for } x < 0. \end{cases} \tag{3.9}$$

Let $G(x)$ denote its cumulative distribution function, i.e.

$$G(x) = \int_0^\infty g(x)dx. \tag{3.10}$$

Let $X_1 \leq X_2 \leq \cdots \leq X_m$ denote the order statistics of an independent random sample of size m from this distribution and let $F_i(x)$ $(i = 1, \ldots, m)$ denote the cumulative distribution function of the ith order statistic X_i, then

$$F_i(x) = \Pr[X_i \leq x] = \sum_{j=i}^{m} \binom{m}{j} [G(x)]^j [1 - G(x)]^{m-j}. \tag{3.11}$$

The expected value of X_i is (see David, 1970)

$$E[X_i] = m \binom{m-1}{i-1} \int_{-\infty}^{\infty} x[G(x)]^{i-1}[1 - G(x)]^{m-i} g(x)dx. \tag{3.12}$$

Instead of solving this equation by numerical integration, which would require extensive computational work, general results from David and Johnson (1954) may be applied. These authors obtained expressions to approximate the cumulants of order statistics from continuous populations up to order $(m+2)^{-3}$. For our particular case we found that using the expression for approximation to order $(m+2)^{-2}$ is enough to reproduce the 11 values (with three decimal places) used by Zahn for the case of $m = 15$. Thus the general expression for approximations to order $(m+2)^{-2}$, also presented by David (1970), is used here.

Let $p_i = i/(m+1)$, $q_i = 1 - p_i$, and $H_i = G^{-1}(p_i)$, where G^{-1} is the inverse of 3.10. The expected value of X_i is to order $(m+2)^{-2}$

$$E[X_i] = H_i + \frac{p_i q_i}{2(m+2)} H_i'' + \frac{p_i q_i}{2(m+2)^2} \left[\frac{q_i - p_i}{3} H_i''' + \frac{p_i q_i}{8} H_i'''' \right] \tag{3.13}$$

where H_i'', H_i''' and H_i'''' are the 2nd, 3rd and 4th derivatives of G^{-1} evaluated at $x = p_i$. To obtain them note that

$$1 = \frac{d}{dx}G(G^{-1}(x)) = g(G^{-1}(x))\frac{d}{dx}G^{-1}(x), \tag{3.14}$$

then

$$\frac{d}{dx}G^{-1}(x) = \frac{1}{g(G^{-1}(x))}. \tag{3.15}$$

Hence

$$\frac{d^2}{dx^2}G^{-1}(x) = \frac{d}{dx}\left[\frac{1}{g(G^{-1}(x))}\right] = \frac{-g'(G^{-1}(x))}{g^3(G^{-1}(x))} = \frac{G^{-1}(x)}{g^2(G^{-1}(x))} \tag{3.16}$$

as (3.9) is such that $g'(x) = -xg(x)$.

Continuing it is found that

$$\frac{d^3}{dx^3}G^{-1}(x) = \frac{1 + 2G^{-1}(x)}{g^3(G^{-1}(x))} \tag{3.17}$$

and

$$\frac{d^4}{dx^4}G^{-1}(x) = \frac{G^{-1}(x)\{7 + 6[G^{-1}(x)]^2\}}{g^4(G^{-1}(x))}. \tag{3.18}$$

Taking $x = p_i$ and replacing H_i'', H_i''' and H_i'''' in (3.13) we obtain the expression to approximate the expected values of order statistics from the samples of the standard half-normal distribution (HEOS) to order $(m+2)^{-2}$. This is

$$\begin{aligned} E(X_{i:m}) &= H_i + \frac{p_i q_i}{2(m+2)}\left[\frac{H_i}{g^2(H_i)}\right] \\ &+ \frac{p_i q_i}{(m+2)^2}\left[\frac{(q_i - p_i)(1 + 2H_i^2)}{3g^3(H_i)} + \frac{p_i q_i H_i(7 + 6H_i^2)}{8g^4(H_i)}\right]. \end{aligned} \tag{3.19}$$

The values H_i can be obtained from tables of the standard normal distribution or from most statistical packages using the relation

$$G^{-1}(p) = \Phi^{-1}\left(\frac{1+p}{2}\right), \ 0 \le p \le 1, \tag{3.20}$$

where $\Phi(x)$ denotes the standard normal cumulative distribution function.

Sets of HEOS that were computed using (3.19) are presented in Table 5.6 at the end of Chapter 5 for several values of m. The sets for $m = 7$, 15, 31 and 63 were used in obtaining, via simulation, critical values for the ratios involved in the sequential procedure described in Chapter 2 for PER at 0.05, 0.20 and 0.40. Further details are given in Chapter 4.

3.2.3 60% procedure (BP)

Berk and Picard (1991) presented critical values for significance levels $p = 0.05$, 0.01, 0.001 for testing individual contrasts for several sizes of experiments. However they did not present values for the case of 63 contrasts, so they were computed following the same calibration process followed by Berk and Picard which can be described as follows.

According to this procedure, for the case $m = 63$, 25 ratios are formed by dividing the 25 largest mean squares by the average of the smallest 38 mean squares. For a significance level p the critical value has to be such that the expected proportion of false rejections in a null experiment is $p \times 63$ out of 63. As only 25 are actually tested, the proportion of the 25 expected to be rejected is $(p \times 63)/25$.

One million sets of independent samples of 63 null contrasts were simulated to obtain the $1 - (p \times 63)/25$ percentiles for $p = 0.001$, 0.01, 0.05.

The procedure was then supplemented with this set of critical values and with a bar graph display of the ratios and reference lines at the critical values as shown in the example below.

3.2.4 Lenth's method (LE)

Programming Lenth's procedure is a straightforward task, however as will be seen in the next chapter, the values tabulated by Lenth (1989) are far from producing the error rates they are supposed to, making it appear too conservative for small

and moderate sizes of experiments.

3.3 Example

The graphical displays of the four procedures are illustrated using a 2^5 example (labelled [92] in Appendix 3A), reported by Kempthorne (1952). It is a classic experiment conducted at Rothamsted to study the effects of five fertilizers on the yield of mangolds. The fertilizers were amounts of sulphate of ammonia S, superphosphate P, muriate of potash K, agricultural salt N and dung D. The experiment was arranged in four blocks of eight plots, confounding the interactions SNP, PKD, and SKND with block effects. The 31 contrasts are listed in Table 3.1 in standard order.

Figures 3.1, 3.2, 3.3, 3.4 show the graphical displays of the methods BM, BP, LE and ZA respectively.

The four procedures agree in deeming three contrasts as being highly significant. These are the contrasts measuring the main effects of the factors S, D and N. The procedures BP and LE also find that two other contrasts are significant at the 0.05 significance level for testing of individual contrasts; however as it is discussed in the next chapter this rule might be too liberal for this size of experiment. It is interesting to note that the half-normal plot bends slightly upwards at about the 13th order statistic suggesting that the basic assumptions might not be properly satisfied.

TABLE 3.1
Contrasts for the Rothamsted mangolds example in standard order

No	Name	Value
1	S	5328
2	P	-312
3	SP	104
4	K	152
5	SK	1240
6	PK	-448
7	SPK	-96
8	N	2152
9	SN	728
10	PN	432
11	SPN (block)	-96
12	KN	992
13	SKN	-592
14	PKN	-40
15	SPKN	104
16	D	2896
17	SD	-96
18	PD	-56
19	SPD	760
20	KD	-168
21	SKD	776
22	PKD (block)	1248
23	SPKD	-656
24	ND	-664
25	SND	-88
26	PND	-176
27	SPND	-752
28	KND	160
29	SKND (block)	-352
30	PKND	408
31	SPKND	216

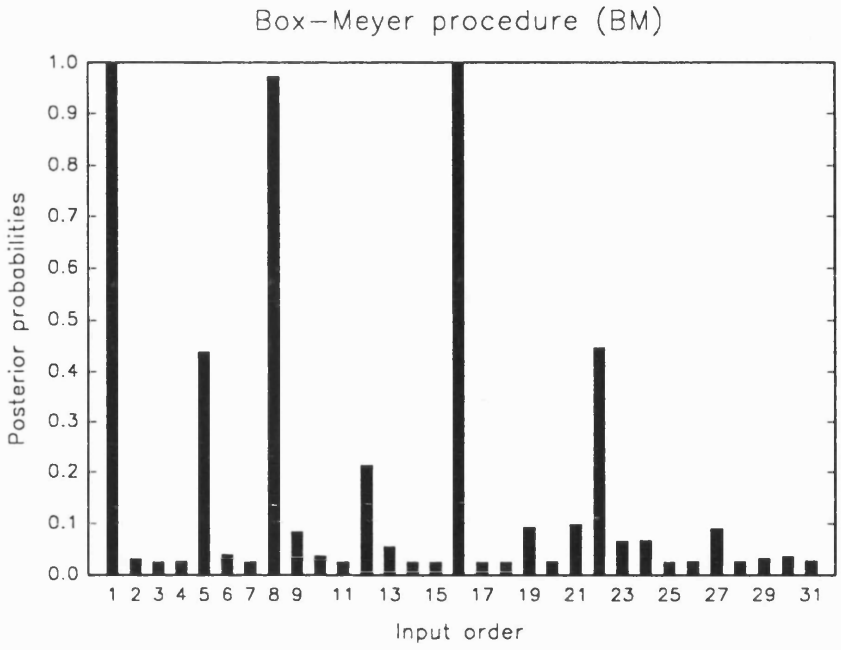


Figure 3.1: Posterior probabilities of being active for contrasts of the Rothamsted mangolds experiment

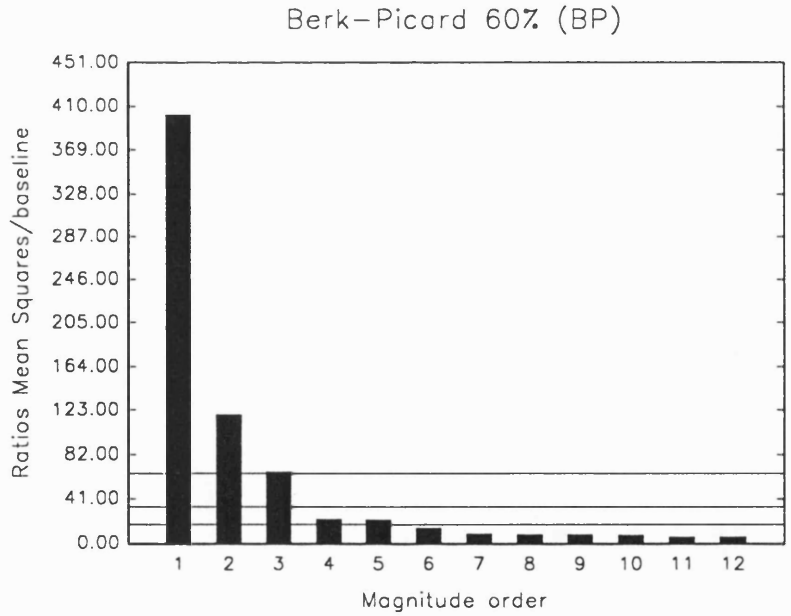


Figure 3.2: Ratios and significance lines at 0.05, 0.01 and 0.001 for the Rothamsted mangolds example

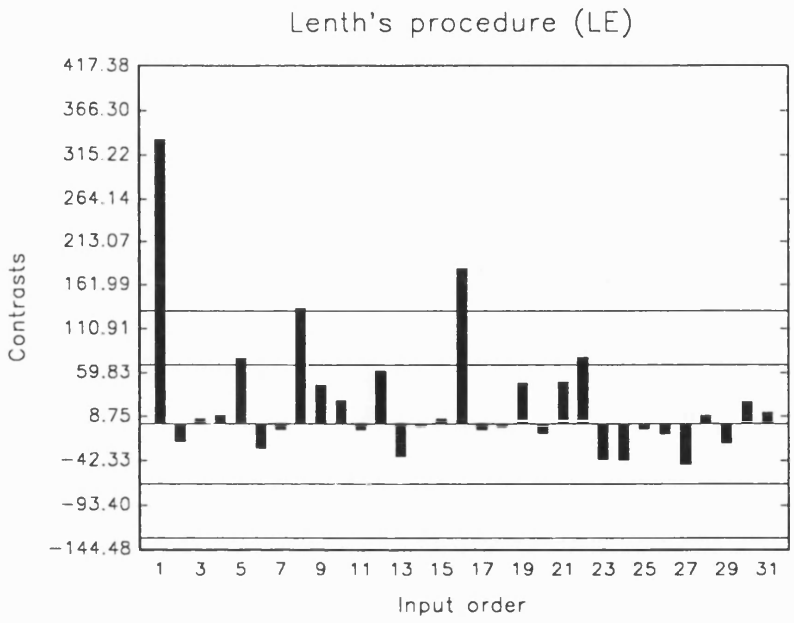


Figure 3.3: Margin of error and simultaneous margin of error for contrasts of the Rothamsted mangolds experiment

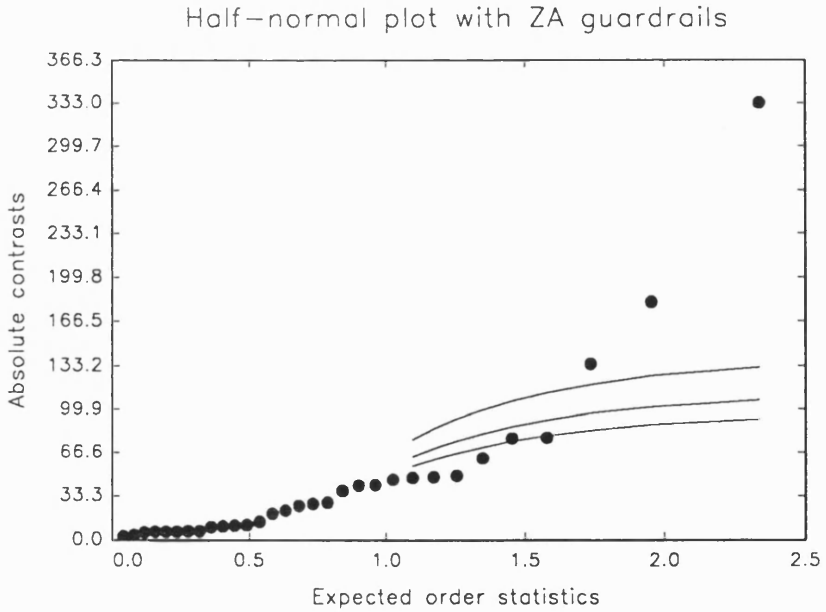


Figure 3.4: Half-normal plot for the Rothamsted mangolds example with ZA guardrails at PER=0.05 0.20 and 0.40

Chapter 4

Comparison of four selected methods

4.1 Error rates

As it was stated in Chapter 2, the relevant problem analyzing unreplicated factorials is to decide which, if any, of the m contrasts under study have nonzero mean. Using the notation introduced in section 2.1 this is equivalent to testing the following family of null hypotheses:

$$H_i : \mu_i = 0, \quad i = 1, \ldots, m. \tag{4.1}$$

The intersection of all the members of the family is called the *overall null hypothesis.* Classical approaches to simultaneous testing of a number of hypotheses are commonly based on extensions of the Neyman-Pearson paradigm for testing of a single null hypothesis. Critical values of certain tests statistic are obtained to control, under the overall null hypothesis, one or both of the following error rate criteria: (a) The expected proportion of the total number of tests that result in false rejections, and (b) the probability of one or more false rejections. The latter is the *probability of a nonzero family error rate* (PER), defined in chapter 2. As

with other concepts in simultaneous inference there is not a standard term for this idea. It is also called *experimentwise error rate.*

The terms to be used here are EPE for criterion (a) and PER for criterion (b).

In practice experimenters tend to test each contrast at some standard significance level δ (typically 0.05 or 0.01). The use of this criterion is equivalent to testing each member of the family at significance level δ. As, under the overall null hypothesis, the expected proportion of false rejections in the whole family is the average of the expected proportion of the members of the family, the use of this criterion controls EPE at level δ. However, the probability of rejecting the overall null hypothesis, i.e. the probability of at least one false rejection PER, can be then much greater than δ.

For a given value δ of EPE there corresponds a value, γ say, of PER and vice versa. The functional relationship between the two values, depends on the joint distribution of the test statistics involved in each procedure. When the test statistics are independent the following relationship holds:

$$\text{PER} = 1 - (1 - \text{EPE})^m. \tag{4.2}$$

The methods BP and LE offer critical values that allow for the control of EPE, while ZA and again LE present critical values that control PER. However in none of these methods are the test statistics independent, because of the denominator used in each case. Note also that because of the effect sparsity assumption, in each of these methods only a subfamily of (4.1) is actually tested. Nevertheless, the critical values in each case have been defined with regard to the whole family. For instance, the use of a significance level equal to δ say in BP means that under the null hypothesis a proportion δ out of the m are expected to be significant.

The Bayesian procedure BM, produces for each contrast its posterior probability of being active. In order to be able to obtain EPE and PER for this procedure we will use a frequentist interpretation and consider a contrast to be active if, using

the prior information suggested by Box and Meyer (1986), its posterior probability of being active is larger than a certain critical value, η say.

One way of looking for a fair basis for comparisons among the methods is to use the sets of critical values presented for each procedure and, via simulation, obtain empirically the values of both EPE and PER. These are then examined to see if there are critical values at which the four methods have similar error rates. If not, new sets of critical values may be obtained. This should be done trying to conserve the original idea of each method as much as possible.

Table 4.1 shows the EPE and PER characteristics for experiments with 7, 15, 31, and 63 contrasts, when the critical values used for LE and BP are supposed to control EPE at 0.05; the critical values used for ZA are supposed to control PER at 0.40; and the criterion used for BM is to deem a contrast to be significant if its posterior probably of being active is larger than 0.5. The values were obtained from 100,000 samples of standardized contrasts simulated for BP, LE and ZA, and 10,000 for BM which requires much more processing time.

There are some interesting features in this table. First, the relationships between the PER and EPE values reveal that the test statistics in each procedure are not independent. Take $m = 15$ for instance, according to (4.2) if the test statistics were independent a value of EPE $= 0.05$ should match to a value of PER $= 0.54$. Second, the fact that, when either of the values EPE or PER is similar in two procedures the other one is also similar for the same number of contrasts. This will facilitate the choosing of cut off points for comparisons.

Another discovery from Table 4.1 is the fact that the critical values for LE given by Lenth (1989) do not produce the empirical error rate they are supposed to. The critical values to construct the margin of error ME for several sizes of experiments should produce values of EPE of 0.05 approximately. However, the empirical EPEs obtained using those critical values are much smaller than 0.05, making the "margin of error" ME, too conservative. The inaccuracy is greater for small experiments. Recall that Lenth computed these critical values on the grounds that

the ratios U_i/PSE are distributed approximately as t with $m/3$ degrees of freedom. However these results indicate that for the sizes of experiments considered here, at least in the tails of the distributions, the approximations are rather inaccurate. This has been seen before by Berk and Picard (1991) but they did not seem to be aware of the implications. In their simulation study the "proportion judged *real*" for the null case is equivalent to our empirical EPE. The values obtained by Berk and Picard are 0.019 for $m = 7$ and 0.028 for $m = 15$ from 1,000 simulated data sets.

TABLE 4.1

Some empirical error rate characteristics of the four procedures for experiments with 7, 15, 31 and 63 contrasts.

	Number of contrasts (m)							
	7		15		31		63	
Method	EPE	PER	EPE	PER	EPE	PER	EPE	PER
BM	0.045	0.21	0.027	0.26	0.019	0.38	0.016	0.58
BP	0.050	0.26	0.050	0.46	0.050	0.70	0.050	0.91
LE	0.020	0.10	0.029	0.25	0.037	0.53	0.044	0.84
ZA	0.088	0.40	0.052	0.40	0.026	0.40	0.011	0.40

The inaccuracy of Lenth's critical values for the "simultaneous margin of error" SME is expected to be larger since, besides the lack of precision of the approximations, they were obtained under the assumption (only approximately true) that the test statistics are independent.

In order to solve this problem, new sets of critical values controlling EPE and PER at 0.05 were obtained empirically as described below.

4.2 Corrected critical values for LE

Table 4.2 shows the critical values given by Lenth (1989) and the set of new values for experiments with m=7, 15, 31, and 63 contrasts. The new (empirical) critical

values were obtained by means of simulation. For each size of experiment, 100,000 sets of standardized contrasts were simulated and, for a range of possible critical values, EPE and PER were computed. Lenth's critical values are also exhibited in this table to show the difference between the two sets. This is large for the case m=7 and decreases gradually as the size of experiment increases. Note that if the values in the column "Empirical EPE = 0.05" are shifted one line down they are very close to Lenth's values in the column $t_{.975;d}$. Similarly with the columns "Empirical PER = 0.05" and Lenth's value for the SME $t_{\gamma;d}$, where in two occasions the corrected critical value is equal to one of Lenth's but for different m. This is mere coincidence but exhibits the magnitude of the inaccuracy.

From now onwards, LE with corrected critical values will be referred to as LE′.

TABLE 4.2

Quantiles of the t distribution with $d = m/3$ degrees of freedom (given by Lenth) and quantiles of the empirical distribution obtained from 100,000 samples simulated.

m	Lenth $t_{.975;d}$	Empirical EPE=0.05	Lenth $t_{\gamma;d}$	Empirical PER=0.05
7	3.76	2.30	9.01	4.86
15	2.57	2.15	5.22	4.22
31	2.22	2.07	4.22	3.91
63	2.08	2.01	3.91	3.81

Note: $\gamma = (1 + 0.95^{1/m})/2$.

4.3 Error rate characteristics for comparison criteria

A brief exploration of the empirical error rates of the four methods using several critical values revealed the following characteristics.

For experiments with number of contrasts m=7 and m=15, the use of the critical values controlling EPE at 0.05 for BP and LE′ give values of PER not

too far away from 0.20 and 0.40 respectively. Recall that ZA offers critical values controlling the PER at 0.05, 0.20 and 0.40. The use of critical values of ZA controlling the PER at 0.20 for m=7 and 0.40 for m=20 give values of EPE that are not far away from 0.05. BM gives similar values of EPE (not far from 0.05) and PER (close to 0.20 for m=7 and close to 0.40 for m=15) when the criterion used is to declare a contrast to be significant if its posterior probability of being active is greater than 0.5 for m=7 and 0.33 for m=15.

Something similar happens for experiments with m=31 and m=63, but using the critical values controlling EPE at 0.01 for BP and LE′. This meant obtaining critical values for LE′ controlling EPE at a value not originally considered by Lenth. The use of these critical values gives values of PER not far away from 0.20 ($m = 31$) and 0.40 ($m = 63$) which are levels suggested for ZA. The criteria that produce similar values for BM are to declare a contrast to be significant when its posterior probability of being active is greater than 0.65 for m=31 and 0.6 for m=15.

Table 4.3 shows the criteria adopted for the comparison of the four procedures on real data. The values of EPE and PER were obtained from 100,000 samples of standardized contrasts simulated for BP, LE and ZA and 10,000 for BM. Critical values for ZA, which uses several for each size of experiment, (one per step in the sequential procedure) are presented in Table 4.4.

There is a certain arbitrariness in all this. The procedures BP and LE′ are, by definition, strictly comparable in terms of EPE. The remaining criteria are a compromise between common usage and the desire of make the four procedures at least roughly comparable. The idea is to apply them to the sets of real data compiled and examine the results for features that may not be explained by the differences in error rate characteristics shown in Table 4.3.

TABLE 4.3

Critical values (c.v.) and empirical error rate characteristics of the four methods for the criteria to be used for comparisons on real data.

	Number of contrasts											
	7			15			31			63		
Method	c.v.	EPE	PER	c.v.	EPE	PER	c.v.	EPE	PER	c.v.	EPE	PER
BM	0.50	0.045	0.21	0.33	0.049	0.44	0.65	0.010	0.23	0.60	0.011	0.45
BP	23.76	0.050	0.26	18.93	0.050	0.46	33.62	0.010	0.22	32.76	0.010	0.40
LE′	2.30	0.050	0.23	2.15	0.050	0.40	3.03	0.010	0.19	2.74	0.011	0.38
ZA	†	0.041	0.20	†	0.052	0.40	†	0.010	0.20	†	0.011	0.40

†Critical values for ZA are given in Table 4.4.

TABLE 4.4

PER = 0.40, 0.20, and 0.05 level critical values for ZA, for experiments with m contrasts. The sequential procedure starts with the critical values at $k = m$.

$m = 7$	PER		
k	0.40	0.20	0.05
6	1.42	1.77	2.58
7	1.85	2.37	3.53

$m = 15$	PER		
k	0.40	0.20	0.05
12	1.48	1.73	2.22
13	1.72	2.04	2.65
14	1.94	2.32	3.04
15	2.17	2.60	3.41

$m = 31$	PER		
k	0.40	0.20	0.05
23	1.48	1.67	2.02
24	1.61	1.84	2.25
25	1.74	1.99	2.45
26	1.86	2.14	2.63
27	1.99	2.29	2.81
28	2.10	2.42	2.98
29	2.21	2.56	3.14
30	2.33	2.69	3.32
31	2.44	2.83	3.49

$m = 63$	PER		
k	0.40	0.20	0.05
45	1.53	1.69	1.98
46	1.61	1.79	2.10
47	1.68	1.88	2.22
48	1.75	1.96	2.31
49	1.82	2.04	2.41
50	1.89	2.12	2.50
51	1.95	2.19	2.58
52	2.02	2.27	2.67
53	2.08	2.34	2.75
54	2.14	2.41	2.84
55	2.21	2.48	2.92
56	2.27	2.54	3.00
57	2.33	2.61	3.09
58	2.39	2.68	3.17
59	2.45	2.75	3.25
60	2.51	2.82	3.33
61	2.56	2.88	3.40
62	2.62	2.95	3.48
63	2.68	3.01	3.57

4.4 Comparative performance on real data

In this section the main features of the results of applying the four methods to the 102 sets of real data described in Chapter 3 are presented. The methods were applied using the criteria exhibited in Table 4.3 and 4.4. The exposition is organized by the number m of contrasts.

4.4.1 Case $m = 7$

The results of applying the four methods to the 30 examples with 7 contrasts exhibited a complete agreement among the four methods in the contrasts deemed to be significant in 24 out of the 30 examples. The main features of the differences on the remaining 6 examples are analyzed below.

Table 4.5 shows the distributions of significant contrasts per experiment and total numbers of significant contrasts for the four methods. This suggests that the inability of ZA to examine more than 30% (in this case 2) of the contrasts is a serious drawback.

TABLE 4.5

Distribution of numbers of significant contrasts per experiment for four methods applied to 30 experiments with 7 contrasts

Number of significant contrasts	Method			
	BM	BP	LE	ZA
0	13	12	13	14
1	4	4	3	3
2	9	12	11	13
3	4	2	3	0
4+	0	0	0	0
Total no. of significant contrasts	34	34	34	29

Most other differences are a consequence of borderline contrasts, i.e. contrasts that were just significant using one method and nearly significant using another. However there is one example that deserves closer examination. It is a 2^3 presented by Box and Draper (1987) and labelled [1] in Appendix A. The aim of the experiment was to determine the effect of amounts of carbon (C), manganese (Mn) and nickel (Ni) on the temperature at which martensite starts being formed in a steel. The 7 contrasts are shown in Table 4.6.

TABLE 4.6

Contrasts of example [1] in standard order

Name	Value
C	-287.5
Mn	-45.0
Ni	-32.5
C.Mn	5.0
C.Ni	7.5
Mn.Ni	5.0
C.Mn.Ni	-5.0

The application of BP and LE′ resulted in the three main contrasts (C, Mn and Ni) being detected as significant while ZA deemed to be significant the 2 largest contrasts. The surprise was that the application of BM resulted in only the largest contrast being significant. The reason for the disagreement of BM with the other procedures in this example is not difficult to find. BM assumes that the contrasts come from one of two Normal populations, both with mean zero but one more highly variable than the other. As the variance is unknown, the particularly large contrast C makes it more likely that all the others come from the less variable population. On the other hand BP and LE′ measure the variability from the smallest 4 contrasts. This situation will be studied further in Chapter 6.

4.4.2 Case $m = 15$

As a result of applying the four methods to the 54 experiments with 15 contrasts using the criteria described above, the agreement in the contrasts considered to be significant was complete in 33 out of the 54 examples.

Table 4.7 shows the distributions of numbers of significant contrasts per experiment and the total numbers of significant contrasts per method. These results again suggest a problem with ZA: its inability to examine more than 4 contrasts appears, on this evidence, to be an undesirable limitation. It is also interesting to note that BP seems slightly less powerful than the other methods despite a slightly higher PER (Table 4.3).

TABLE 4.7

Distribution of numbers of significant contrasts per experiment for five methods applied to 54 experiments with 15 contrasts

Number of significant contrasts	Method			
	BM	BP	LE	ZA
0	6	6	6	6
1	13	12	11	9
2	15	18	20	19
3	9	9	6	7
4	6	5	5	13
5	4	4	4	0
6	0	0	2	0
7	1	0	0	0
8+	0	0	0	0
Total no. of significant contrasts	121	115	121	120

In most of the examples for which the methods disagree, the differences (leaving out ZA's limitation), are due to contrasts close to the critical values, however there are some odd examples. The most apparent is the set of dispersion contrasts in

an experiment on piston ring spacers presented by Grove and Davis (1992) and labelled [60] in Appendix A. The result of applying the 4 methods to this example exhibited the following numbers of contrasts deemed to be significant per method: BM, 7; BP, 4; LE′, 6; ZA, 4. However, the authors considered that none of the 15 dispersion contrasts was active and, in any case, the magnitude of the largest ones was small for any practical purposes. The reason for these disagreements is apparent in the half-normal plot presented by Grove and Davis; the smallest 9 contrasts, assumed to be null (the effect sparsity principle), do not form a straight line, in fact the ordered absolute contrasts seem to plot exponentially against the half-normal scores. This possible violation of a basic assumption seems to have affected the four methods in a similar way. An advantage of the half-normal plot is that violation of some basic assumptions as well as the presence of one faulty observation can be spotted by visually inspecting the graph (see Daniel, 1959).

4.4.3 Case $m = 31$

Table 4.8 shows the total numbers of significant contrasts and the maximum number of significant contrasts found in a single experiment as a result of applying BM, BP, LE and ZA on 9 examples with 31 contrasts. The methods were applied with the critical values that produced empirical values of EPE = 0.01 and of PER about 0.20 for the four methods as shown in Table 4.3.

There was a perfect agreement between BP and LE′ on one hand, and between BM and ZA on the other hand. The latter two methods identified two more contrasts as being active in the total. In this case ZA is able to examine up to 9 contrasts, which seems, for these examples and for PER = 0.20, to be adequate.

TABLE 4.8

Total numbers of significant contrasts and maximum per experiment for four methods applied to 9 experiments with 31 contrasts

	BM	BP	LE	ZA
Total number of significant contrasts	19	17	17	19
Maximum in a single experiment	5	5	5	5

4.4.4 Case $m = 63$

Table 4.9 shows the total numbers of significant contrasts and the maximum number of significant contrasts found in a single experiment as a result of applying the 4 methods to 9 examples with 63 contrasts. The methods were applied with the critical values exhibited in Table 4.3.

The results showed few surprises. There was complete agreement among the four methods in 6 out of the 9 experiments. For one of the remainder the application of ZA resulted in 10 contrasts being deemed significant while the application of BM, BP and LE′ resulted in 6 significant contrasts. However the 4 contrasts which made the difference are very similar and very close to be significant using these methods. The differences in the results for the other 2 examples are the consequence of similar situations, giving the total numbers of significant contrasts shown in Table 4.9.

The value of PER of about 0.40 used in this case corresponds to ZA's most liberal criterion. The maximum number of contrasts that this method is able to examine, which is 12, seems to be adequate.

TABLE 4.9

Total numbers of significant contrasts and maximum per experiment for four methods applied to 9 experiments with 31 contrasts

	BM	BP	LE	ZA
Total number of significant contrasts	28	26	27	30
Maximum in a single experiment	6	6	6	10

4.5 Concluding remarks

In order to make possible the comparison of the four methods it was decided to analyze empirically, for null experiments, the behaviour of the two main criteria used in simultaneous inference: the expected proportion of false rejections (EPE) and the probability of at least one false rejection (PER).

This led to the conclusion that the critical values in LE were quite inaccurate. The approximation of the t distribution used by Lenth (1989) is not adequate for small and moderate size of experiments, which are the most common. Another source of inaccuracy was Lenth's assumption of independence of the test statistics in the definition of the simultaneous margin of error (SME). New sets of critical values were obtained by means of simulation controlling the values of EPE at 0.05 for the margin of error (ME) and PER at 0.05 for the SME. Two extra critical values were obtained controlling EPE at about 0.01 for the cases of $m = 31,\ 63$ to match with the other procedures for the comparison on real data.

Eventually, it was possible to find sets of critical values matching the values of EPE and PER approximately for the four methods for each of the four sizes of experiments used.

On basis of critical values giving similar error rates, in most examples the

four procedures wholly agreed in the identification of the contrasts that should be deemed to be significant; the differences, in many cases, were small and unimportant. However, in some cases, it was possible to identify practical situations in which some of the methods will lead to a wrong interpretation.

For example, it was apparent that BM may overlook some real contrasts in situations where, besides these real contrasts, there is a very large contrast, which makes the others appear negligible. But the most serious problem was that the range of degrees of effect sparsity considered by ZA is, compared with the other procedures, unrealistic, at least for small and moderate sizes of experiment. This might be one of the reasons for the lack of popularity of ZA. An alternative, which considers a more appropriate number of contrasts assumed to be null, is presented in the next chapter.

Chapter 5

An alternative method of inference using half-normal plots

One of the features of the results of the previous chapter is that the maximum number of contrasts that can be assessed for significance using ZA seems to be inappropriate. In this chapter several modifications are analyzed and one which allows the examination of a more sensible number of contrasts is proposed.

5.1 Analysis of alternatives

Recall that ZA is a modification of the method of statistical testing with half-normal plots proposed by Daniel (1959). A more general statement of the detection process is set up here by considering the number of contrasts assumed to be null (effect sparsity), in an experiment with m contrasts, as a certain function $b(m)$. Then the information available will be used to find a compromise definition for $b(m)$.

Let $V_{1:m}, \ldots, V_{m:m}$ denote the ordered absolute contrasts of an experiment with m contrasts and let $W_{1:m}, \ldots, W_{m:m}$, denote the expected values of the order statis-

tics of a sample of size m from the standard ($\sigma = 1$) half-normal distribution. The procedure, in this more general way, uses sequentially the following test statistics:

$$T(k, b(m)) = \frac{V_{k:k}}{S(b(m), m)}, \quad k = m, m-1, \ldots, b(m)+1 \tag{5.1}$$

where

$$\mathrm{S}(b(m), m) = \frac{\sum_{i=1}^{b(m)} V_{i:m} W_{i:m}}{\sum_{i=1}^{b(m)} W_{i:m}^2}. \tag{5.2}$$

The detection process starts by comparing the ratio $T(m, b(m))$ with a critical value $c_{m,b(m)}$ which controls the PER at certain level γ, i.e., under the assumption that all the contrasts are null, $c_{m,b(m)}$ is such that

$$\Pr\{T(m, b(m)) < c_{m,b(m)}\} = 1 - \gamma. \tag{5.3}$$

If the statistic is smaller than the critical value, all the contrasts are deemed to be inert and the procedure stops; otherwise the largest contrast is declared to be significant and $T(m-1, b(m))$ is compared with the critical value $c_{m-1,b(m)}$ which, under the assumption that the remaining $m-1$ contrasts are null, is such that

$$\Pr\{T(m-1, b(m)) < c_{m-1,b(m)}\} = 1 - \gamma, \tag{5.4}$$

and so on. The procedure stops when either a test statistic is smaller than its corresponding critical value or the statistic $T(b(m), b(m))$ is encountered. Consequently the procedure is unable to assess the significance of more than $m - b(m)$ contrasts in a single experiment. The presence of more than $m - b(m)$ real contrasts would also contaminate the test statistic denominator (5.2) reducing the procedure's power.

A particular choice of $b(m)$ originates what will be referred to as a *variant* of the detection process. In this sense ZA is a variant which sets $b(m)$ as the integer number nearest to $0.683m + 0.5$. The results of the analyses made in Chapter 4, suggested that this is not appropriate. For example, for $m = 7$ and for $m = 15$, only up to 2 and 4 contrasts respectively may be examined using this variant, but about 10% of the examples for these sizes of experiment analyzed in Chapter 4 presented greater numbers of contrasts deemed to be significant as a result of applying BM, BP or LE′ at the standard significance level EPE = 0.05.

It is natural to look for a variant that uses a definition of $b(m)$ considering smaller numbers of contrasts assumed to be null in (5.2), as this will allow assessment of the significance of a larger number of contrasts. However, this also will increase the variation of the test statistics (5.1), because the sums in (5.2) will be over smaller numbers, and therefore increase the magnitude of the critical values at fixed PER, with negative effects on the power for situations covered by ZA. Hence, the value eventually adopted should be a result of a trade off between these two ideas.

The 0.60, 0.80 and 0.95 quantiles of the empirical cumulative distribution functions of the null distribution of the test statistic ratios $T(k, b(m))$ that would be involved in the procedure described above, for different choices of $b(m)$, were obtained via computer simulation for $m = 7$, 15, 31, 63. Then the variants were applied to the 102 experiments in Appendix A. The results are analyzed below.

5.1.1 Case $m = 7$

Figure 5.1 shows, for $m = 7$, the relevant segments of the the empirical cumulative distribution functions of the ratios $T(7, b(m))$, i.e. those involved in the first step of the procedure, for the variants $b(7) = 5$ (ZA), $b(7) = 4$, and $b(7) = 3$. As expected, the smaller $b(7)$ the larger the 0.60, 0.80 and 0.95 quantiles to be used as critical values. The differences between the $1 - \text{PER} = 0.95$ quantiles for two successive values of $b(7)$ are specially noticeable as they strongly affect the power. When $b(7)$ changes from 5 (ZA) to 4, the difference is 0.65 standard deviations and 1.15 standard deviations when $b(7)$ varies from 4 to 3.

The empirical cumulative distribution functions for the second largest ratio $T(6, b(7))$ for the same variants are exhibited Figure 5.2. The differences among them are similar to those for the largest ratio.

These three variants were applied to the 30 examples with 7 contrasts analyzed in Chapter 4. Table 5.1 shows the distributions of significant contrasts per

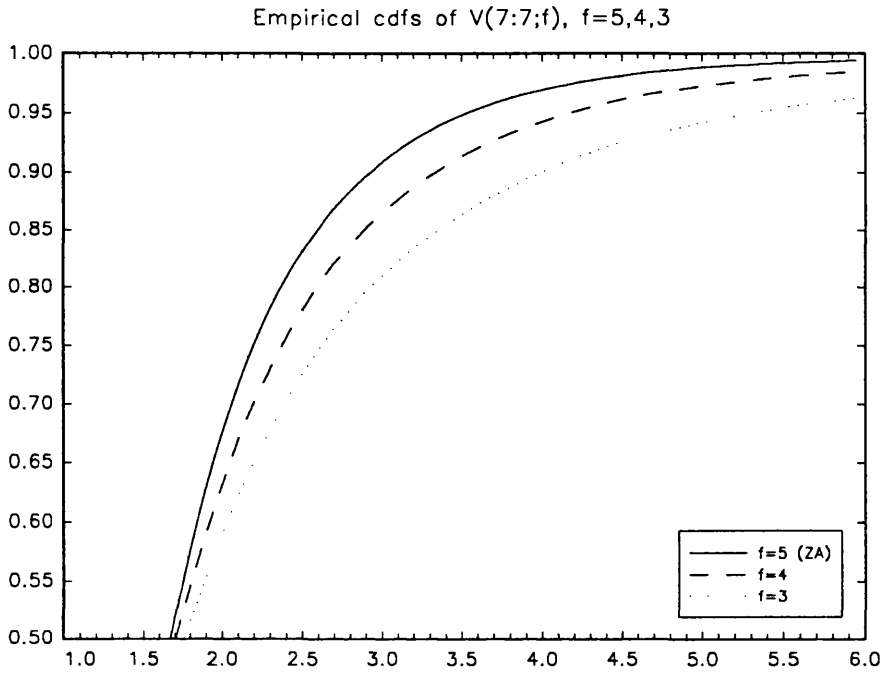


Figure 5.1: Empirical cumulative distribution function of the ratios $T(7, b(7))$ for b(7)=5,4,3

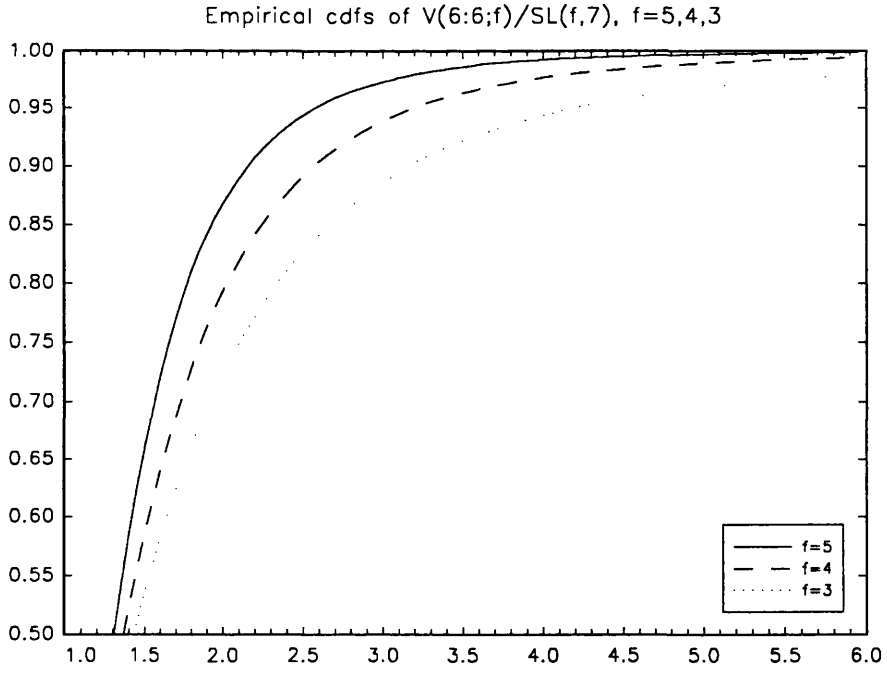


Figure 5.2: Empirical cumulative distribution function of the ratios $T(6, b(7))$ for b(7)=5,4,3

experiment and total numbers of significant contrasts as a result of applying the three variants to those examples using PER = 0.05.

TABLE 5.1

Distribution of numbers of significant contrasts per experiment for four variants of ZA applied to 30 experiments with 7 contrasts.

Number of significant contrasts	Variant		
	$b(7) = 5$ (ZA)	$b(7) = 4$	$b(7) = 3$
0	16	16	22
1	6	7	6
2	8	5	2
3	-	2	0
4+	-	-	0
Total no. of significant contrasts	22	23	10

This, together with the results in Chapter 4, suggests that taking $b(7) = 4$ is a good compromise. Despite having larger critical values the possibility of examining up to three contrasts made this variant appear slightly more powerful on these examples. When $b(7)$ is changed from 4 to 3 the total number of contrasts deemed to be significant falls sharply.

5.1.2 Case $m = 15$

For $m = 15$ the variants analyzed ranged from taking $b(15) = 11$ (ZA) down to $b(15) = 7$. The shapes of the empirical cumulative distribution functions of the test statistics were similar to those for the case $m = 7$, although the increments found between the PER = 0.05 critical values for consecutive values of $b(15)$ were smaller, varying from 0.14 standard deviations when $b(15)$ changed from 11 to 10, to 0.25 standard deviations when $b(15)$ was varied from 8 to 7.

The 5 variants were applied to 53 examples with 15 contrasts analyzed in Chapter 4 (the troublesome example [60] was left out). Table 5.2 shows the distributions of significant contrasts per experiment and total numbers of significant contrasts as a result of applying the 5 variants using PER = 0.05.

TABLE 5.2

Distribution of numbers of significant contrasts per experiment of five variants applied to 53 experiments with 15 contrasts

Number of significant contrasts	Variant				
	$b(15) = 11$ (ZA)	$b(15) = 10$	$b(15) = 9$	$b(15) = 8$	$b(15) = 7$
0	21	21	24	26	26
1	14	15	13	10	11
2	8	7	6	7	5
3	5	5	5	5	5
4	5	2	2	2	4
5	–	3	3	3	1
6	–	–	0	0	0
7	–	–	–	0	0
8+	–	–	–	–	0
Total no. of significant contrasts	65	67	63	62	60

When the value of $b(15)$ was changed from 11 to 10, three examples appeared with 5 contrasts deemed to be significant and the total number of significant contrasts increased from 65 to 67. Subsequent reductions in the value of $b(15)$ resulted in a decrease in the total number of contrasts deemed to be significant. However, the use of $b(15) = 10$ may be too restrictive as, for the examples [59] and [69] the authors suspected that there might be as many as 6 real contrasts in each. With the use of the more permissive PER = 0.40, up to 3 examples appeared to have 6 significant contrasts detected with the variants using $b(15) \leq 9$, however no variant resulted in the detection of more than 6 significant contrasts, suggesting

that a good compromise is taking $b(15) = 9$. Note that the compromises of $b(7) = 4$ and $b(15) = 9$ are both about 0.6 times m. This coincides with the proportion of contrasts used by Berk and Picard (1991) for constructing the baseline for BP.

5.1.3 Cases $m = 31$ and $m = 63$

For $m = 31$ and $m = 63$ there are many possible variants and few examples to base a choice on so the criterion $b(m) = 0.60m$ was extended to these cases giving $b(31) = 19$ and $b(63) = 38$. The application of these variants to the examples with 31 and 63 contrasts analyzed in Chapter 4, as is shown below, gave virtually the same results as applying ZA and the other procedures compared in Chapter 4.

5.2 Method HP

A compromise value of $b(m)$ of about 60% of m seems to be quite safe in avoiding the contamination of $S(b(m), m)$ by large contrasts in small and moderate sizes of experiments ($m = 7, 15$), and does not seem to affect greatly the power in large experiments ($m = 31, 63$). The method with this definition of $b(m)$ will be called HP (after half-normal plot).

Although the analysis of the detection process that led to HP has been made without the half-normal plot, the idea is to use it with this plot (also called the Daniel plot). The beauty of the plot is that, besides aiding the interpretation, it may be used to examine the contrasts for departures from the basic assumptions (see Daniel, 1959). In order to facilitate the application of HP using the Daniel plot a macro in MINITAB code (macro HP) is presented in Appendix B.

Sets of critical values for the use of HP for different sizes of recommended designs up to $m = 31$ are presented in Table 5.5 at the end of this chapter. All sets of critical values were obtained with a minimum of 500,000 simulated samples. Although the expected values of order statistics of samples of the standard

half-normal distribution (HEOS) required for the computation of $S(b(m), m)$ are computed by the macro HP using the expression (3.19) obtained in Chapter 3, sets of these values for the same designs are presented in Table 5.6 next to the critical values.

5.3 Comparative performance of HP

In order to compare the performance of HP on real data with the four methods compared in Chapter 4, HP was applied to the same examples and using similar error rates, following the criteria presented in Table 4.3, i.e. PER = 0.20 for $m = 7$ and 31 and PER = 0.40 for $m = 15$ and 63. This gives EPE slightly greater than 0.05 for $m = 7$ and 15, and EPE slightly greater than 0.01 for $m = 31$ and 63.

A summary of the results of the application of the five methods to the examples analyzed in Chapter 4 using similar error rates is given in Table 5.3.

TABLE 5.3

Total numbers of contrasts deemed to be significant as a result of applying 5 methods to 102 examples using comparable error rates.

Number of contrasts	Number of examples	Method				
		BM	BP	LE′	ZA	HP
7	30	34	34	34	29	37
15	54	121	115	121	120	138
31	9	19	17	17	19	18
63	9	28	26	27	30	30

For experiments with 7 and 15 contrasts, HP appears, on the examples analyzed, to have greater power to detect active contrasts, while for experiments with 31 and 63, the performance of HP was comparable with other methods.

5.4 Example

To illustrate the use of HP consider the example labelled [56] in Appendix A from Grove and Davis (1992). The experimenters investigated the effects of 8 factors at two levels on the number of flow marks in the manufacture of plastic lids for car glove boxes. The factors were assigned to columns of the so called $L_{16}(2^{15})$ orthogonal array, to generate a 2^{8-4}_{IV} design.

Table 5.4 shows the ordered absolute contrasts, the expected values of order statistics from the standard half-normal distribution (HEOS), the factors and the alias structure (up to order 2). The one-letter codes (A–H) for the factors are used here for brevity.

Figure 5.3 shows the half-normal plot with guard rails using PER = 0.40, 0.20 and 0.05. Rather than using the ratios $T(k, b(m))$, the ordered absolute contrasts are plotted against the HEOS in Table 5.6. The regression through the origin using the 9 smallest pairs gives the coefficient $S(9, 15) = 0.635$. The products of this coefficient and the critical values given in Table 5.5 are plotted against the 6 largest order statistics and the points are joined by lines to improve the visual impact.

The procedure is to compare first the largest contrast with the guard rail corresponding to the desired PER. If it is plotted above the guard rail, this contrast is declared significant and the next largest contrast is examined. When a contrast plotted beneath the guard rail is met (or the guard rail has ended), that contrast and all the smaller ones are declared insignificant. The main effects of factors B (Melt temperature) and C (Mould temperature) are clearly active. Using the more liberal rule, PER = 0.40, there are three more contrasts that may be deemed to be active. The alias structure of Table 5.4 indicates that the aliased combination estimated by the $3rd$ largest absolute contrast involves the interaction BC, this makes it more likely to be active.

TABLE 5.4
The glove box lid example (Grove and Davis, 1992)

Order number	Absolute contrast	Expected order statistics (HEOS)‡	Factors and alias structure
1	0.038	0.079	Nozzle aperture (D)
2	0.088 †	0.158	AH + BD + CE + FG
3	0.088	0.239	AD + BH + CF + EG
4	0.263 †	0.322	AB + CG + DH + EF
5	0.338 †	0.407	Injection stroke (A)
6	0.338 †	0.496	Second-stage injection speed (E)
7	0.388	0.589	First-stage injection speed (F)
8	0.438	0.688	Change-over point (G)
9	0.438 †	0.794	AE + BF + CH + DG
10	0.563 †	0.910	AF + BE + CD + GH
11	0.988 †	1.040	Mould clamping force (H)
12	1.113	1.191	AC + BG + DF + EH
13	1.163 †	1.376	AG + BC + DE + FH
14	2.438 †	1.625	Mould temperature (C)
15	2.963 †	2.052	Melt temperature (B)

‡These values were obtained from Table 5.6.
†These values were negative before absolute values were taken.

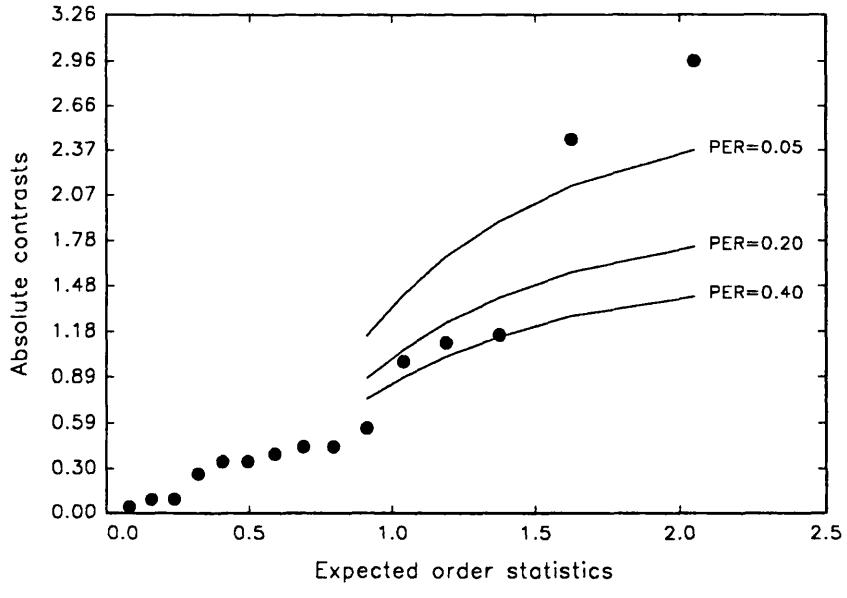


Figure 5.3: Half-normal plot of the glove box lid data with HP guard rails

5.5 A note about power

The main criterion followed for the specification of HP has been to make it appropriate for the degree of effect sparsity suggested by the experiments analyzed. However, in section 5.3, it was shown that for experiments with 7 and 15 contrasts HP identified larger numbers of contrasts as being active than the other methods when they were applied using similar error rates, and it was said that this fact suggested HP having greater power. It may be argued that perhaps the contrasts that made the difference could all be false positives.

Monte Carlo studies on power have shown that the detection rates exhibited by BP, BM and LE (Berk and Picard, 1991) as well as ZA (Zahn, 1975b) are rather low even for situations with active contrasts as large as 4σ. The explanation of these results can be found in the nature of the problem itself; for instance, as the expected value of the largest absolute contrast in a null experiment is about 2σ, any procedure would need much larger error rates than those recommended if it is to have a high detection rate for contrasts of that magnitude. Therefore, it would be sensible to expect that the larger set of contrasts deemed to be active by HP consists of a larger number of active contrasts correctly detected and a number of false positives that is also larger. However, since these experiments are typically used for the screening of factors, having some false positives does not matter as much as failing to identify active contrasts, as presumably all the contrasts deemed to be active will be exposed in later experiments.

5.6 Concluding remarks

The idea proposed by Daniel of using half-normal plots to assess the significance of orthogonal contrasts has been used mainly without any formal tests of significance. A simple modification of one of Zahn's versions of Daniel's procedure of statistical testing with half-normal plots has been described. Its utilization only requires a little effort beyond that usually involved in obtaining these plots with

the advantage that more objective decisions may be made.

The procedure suggested exhibited at least similar power for detecting active contrasts as other comparable methods when they were applied to a substantial number of unreplicated factorials published in the literature.

Analyzing unreplicated factorials involves examining the data, contrasts and residuals in different manners. Daniel (1976) has found it particularly useful to examine the signed contrasts in standard order and the full normal plots of the residuals. The method suggested here is intended to be an aid at the stage of deciding which contrasts should be considered to be active. Provided that the key assumptions are reasonably satisfied the experimenter may be confident in declaring to be active contrasts detected using the guard rail at PER = 0.05 as in the example given in section 5.4. In that example three other contrasts were significant using the more liberal rule of PER = 0.40. These need more careful consideration, and the experimenters should bear in mind the alias pattern as well as their own knowledge of the phenomenon being investigated before making their judgement.

TABLE 5.5

PER = 0.40, 0.20, and 0.05 level critical values for experiments with m contrasts. The entries each subtable are the 1−PER quantiles of the ratios $T(k; b(m))$ for the m given.

$m = 7$	PER		
k	0.40	0.20	0.05
5	1.13	1.45	2.23
6	1.53	2.03	3.21
7	1.92	2.59	4.18

$m = 8$	PER		
k	0.40	0.20	0.05
6	1.23	1.53	2.21
7	1.59	2.04	3.03
8	1.95	2.53	3.82

$m = 11$	PER		
k	0.40	0.20	0.05
8	1.24	1.50	2.04
9	1.53	1.88	2.61
10	1.80	2.25	3.15
11	2.08	2.61	3.69

$m = 15$	PER		
k	0.40	0.20	0.05
10	1.17	1.38	1.81
11	1.39	1.67	2.23
12	1.60	1.94	2.62
13	1.80	2.20	2.98
14	2.01	2.46	3.34
15	2.21	2.72	3.71

$m = 17$	PER		
k	0.40	0.20	0.05
11	1.15	1.35	1.74
12	1.34	1.61	2.11
13	1.53	1.85	2.45
14	1.72	2.08	2.77
15	1.90	2.31	3.09
16	2.08	2.53	3.39
17	2.26	2.76	3.71

$m = 19$	PER		
k	0.40	0.20	0.05
12	1.13	1.32	1.69
13	1.31	1.56	2.02
14	1.48	1.77	2.32
15	1.65	1.98	2.60
16	1.81	2.19	2.88
17	1.98	2.39	3.16
18	2.14	2.59	3.43
19	2.30	2.79	3.71

$m = 23$	PER		
k	0.40	0.20	0.05
15	1.20	1.39	1.74
16	1.36	1.59	2.02
17	1.52	1.78	2.27
18	1.66	1.96	2.51
19	1.81	2.13	2.73
20	1.94	2.30	2.95
21	2.08	2.48	3.18
22	2.22	2.64	3.40
23	2.36	2.81	3.63

$m = 26$	PER		
k	0.40	0.20	0.05
17	1.23	1.41	1.75
18	1.37	1.59	1.99
19	1.51	1.76	2.21
20	1.64	1.92	2.42
21	1.77	2.08	2.63
22	1.90	2.23	2.82
23	2.03	2.39	3.02
24	2.15	2.53	3.21
25	2.28	2.68	3.41
26	2.40	2.83	3.61

$m = 27$	PER		
k	0.40	0.20	0.05
17	1.18	1.35	1.67
18	1.32	1.53	1.91
19	1.45	1.69	2.12
20	1.58	1.85	2.32
21	1.70	2.00	2.53
22	1.82	2.14	2.71
23	1.94	2.29	2.90
24	2.07	2.44	3.09
25	2.19	2.57	3.27
26	2.30	2.72	3.46
27	2.42	2.86	3.64

$m = 31$	PER		
k	0.40	0.20	0.05
20	1.23	1.41	1.72
21	1.36	1.57	1.93
22	1.48	1.71	2.12
23	1.60	1.85	2.29
24	1.71	1.98	2.47
25	1.82	2.12	2.64
26	1.93	2.24	2.80
27	2.04	2.37	2.96
28	2.15	2.50	3.12
29	2.25	2.63	3.28
30	2.36	2.75	3.44
31	2.46	2.87	3.61

TABLE 5.6

Expected values of the order statistics of sample size m from the standard half-normal distribution

Order number k	m = 7	8	11	15	17	19	23	26	27	31
1	0.160	0.141	0.105	0.079	0.070	0.063	0.052	0.047	0.045	0.039
2	0.326	0.287	0.213	0.158	0.140	0.126	0.105	0.093	0.090	0.079
3	0.504	0.441	0.323	0.239	0.212	0.190	0.158	0.140	0.135	0.118
4	0.702	0.608	0.439	0.322	0.285	0.255	0.212	0.188	0.181	0.158
5	0.934	0.795	0.561	0.407	0.359	0.321	0.266	0.235	0.227	0.198
6	1.233	1.017	0.692	0.496	0.436	0.389	0.321	0.284	0.273	0.238
7	1.722	1.306	0.838	0.589	0.516	0.459	0.377	0.333	0.321	0.279
8		1.782	1.004	0.688	0.599	0.531	0.435	0.383	0.369	0.320
9			1.205	0.794	0.686	0.606	0.494	0.434	0.418	0.362
10			1.472	0.910	0.780	0.685	0.555	0.486	0.467	0.405
11			1.922	1.040	0.881	0.769	0.618	0.540	0.519	0.448
12				1.191	0.992	0.858	0.683	0.596	0.571	0.492
13				1.376	1.117	0.955	0.752	0.653	0.626	0.538
14				1.625	1.263	1.062	0.825	0.712	0.682	0.584
15				2.052	1.442	1.183	0.902	0.775	0.741	0.632
16					1.685	1.324	0.986	0.840	0.802	0.681
17					2.104	1.499	1.077	0.910	0.866	0.732
18						1.736	1.177	0.984	0.935	0.785
19						2.148	1.292	1.064	1.008	0.840
20							1.427	1.151	1.087	0.899
21							1.594	1.248	1.173	0.960
22							1.823	1.359	1.270	1.025
23							2.224	1.490	1.380	1.095
24								1.653	1.510	1.171
25								1.877	1.671	1.254
26								2.271	1.894	1.347
27									2.286	1.453
28										1.579
29										1.736
30										1.954
31										2.338

Chapter 6

Exploring extensions to the Bayesian approach BM

Recall from Chapter 4, that the application of BM to one of the examples with 7 contrasts resulted in only one contrast being deemed to be active while two other seemingly active contrasts were overlooked.

The example is a 2^3 design presented by Box and Draper (1987) and labelled [1] in Appendix A. The aim of the experiment was to determine the effects of amounts of carbon (C), manganese (Mn) and nickel (Ni) on the temperature at which martensite starts being formed in a steel. The 7 contrasts are shown in Table 6.1.

The application of BP and LE′ resulted in the three main contrasts (C, Mn and Ni) being detected as significant. HP, the alternative for ZA is in agreement with this. However the application of BM resulted in only the largest contrast being deemed active. Indeed, as it is shown in Figure 6.1, only C has posterior probability of being active larger than 0.5, the critical value for BM according to the comparison criteria defined in Chapter 4.

TABLE 6.1

Contrasts of example

[1] in standard order

Name	Value
C	-287.5
Mn	-45.0
Ni	-32.5
C.Mn	5.0
C.Ni	7.5
Mn.Ni	5.0
C.Mn.Ni	-5.0

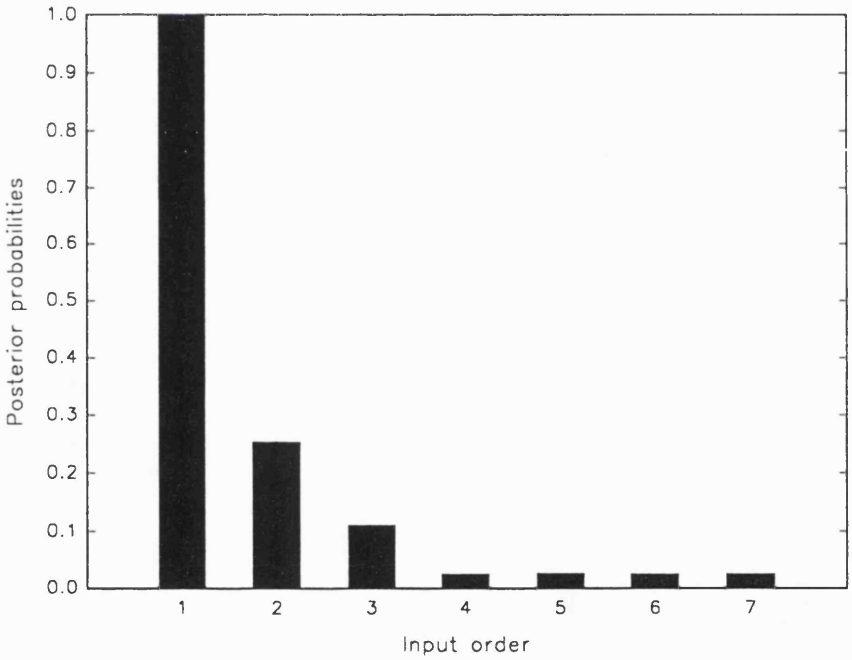


Figure 6.1: Posterior probabilities of being active for contrasts of the Martensite example, when $\alpha = 0.20$ and $k = 10$.

BM assumes that the contrasts are independent realizations of a random variable whose distribution is a mixture of the form $(1 - \alpha)F_1 + \alpha F_2$ where F_1 is $N(0, \tau^2)$ and F_2 is $N(0, (k\tau)^2)$ with $k > 1$. If, as suggested by the other methods, the three largest contrasts in this example are active, the values of $\alpha = 0.20$ and $k = 10$ proposed by Box and Meyer (1986) are far from covering this particular

situation. The specially large contrast C makes it more likely that all the others come from the less variable population F_1.

If the experimenters had an idea a-priori that a particularly large contrast would be present and suspected that one or two other contrasts would also be active but of more moderate size then they would have reason to choose a much larger value for k (e.g. $k = 30$ or $k = 40$) and, as they would expect more than 20% of contrasts to be active, they also should choose a larger value for α (e.g. $\alpha = 0.25$ or $\alpha = 0.30$). Figure 6.2 shows the posterior probabilities of being active for the seven contrasts when $\alpha = 0.25$ and $k = 30$. The contrasts Mn and Ni as well as C have posterior probabilities of being active larger than 0.5 thus, according to our criterion for this size of experiment defined in Chapter 4, these contrasts are deemed to be active. However, when the method is applied with these values of α and k to the other 29 of the 30 examples with 7 contrasts listed in Appendix A, the results are different to those obtained when the recommended values $\alpha = 0.20$ and $k = 10$ were used. Since we are dealing with real data it is not possible to know which contrasts are actually real and which are not. However in 27 of those 29 examples the number of contrasts deemed to be significant when BM is applied using the recommended values $\alpha = 0.20$ and $k = 10$ is in complete agreement with at least two other of the methods BP, LE′, and HP when they were applied using comparable error rates. These examples (identified by the labels given in Appendix A) along with the numbers of contrasts deemed to be active are exhibited in Table 6.2.

It will be assumed here that the three largest contrasts in example [1] are observations of real effects and that the inferences for the 27 examples shown in Table 6.2 are sensible. The idea, then, is to explore modifications to BM in order to make it robust enough to deal with situations like the one suggested by example [1], but giving, for the 27 examples mentioned, basically the same results that were obtained using the recommended values. The criterion to follow is that of Chapter 4 for the case $m = 7$ contrasts, i.e. contrasts with posterior probabilities larger than 0.5 are declared to be active, and the error rates EPE and PER are to be

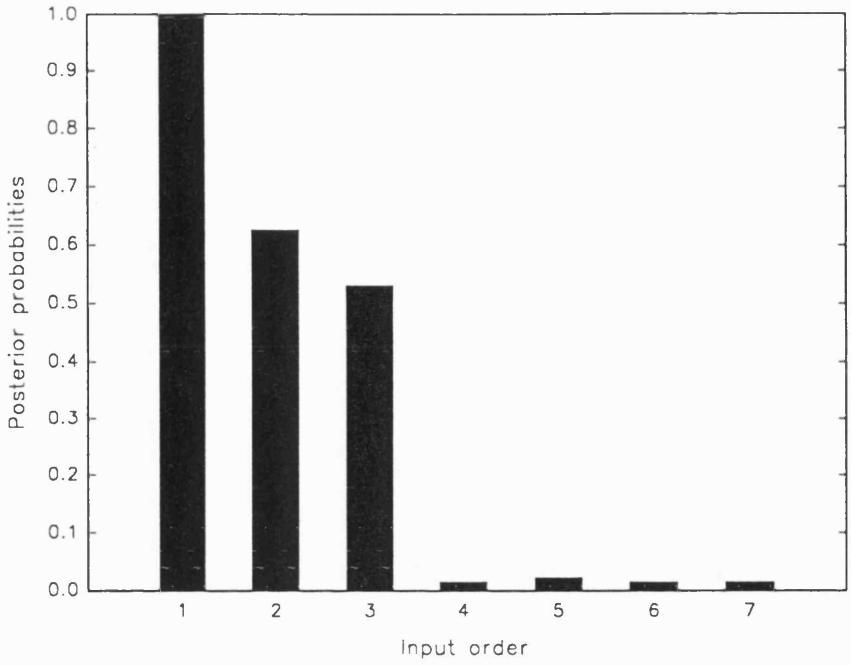


Figure 6.2: Posterior probabilities of being active for contrasts of the Martensite starts temperature example, when $\alpha = 0.25$ and $k = 30$.

kept about 0.05 and 0.20 respectively.

TABLE 6.2

Examples with 7 contrasts for which the number of contrasts deemed to be significant by BM coincided with at least two other methods.

No. of significant contrasts	Examples
0	[4] [5] [6] [15] [17] [19] [22] [23] [24] [26] [27] [28] [30]
1	[10] [12] [16]
2	[2] [7] [8] [11] [13] [20] [21] [25] [29]
3	[14] [18]

6.1 Considering prior distributions for the parameters α and k

In order to tackle situations like the one suggested in example [1] without biasing the inferences too much for the more common situations, the experimenters may wish to be able to express their prior beliefs about specific values of α and k by assigning certain prior probabilities to the values suggested by Box and Meyer and certain prior probabilities to other (including larger) values. In general, the procedure will be more flexible if the experimenters have the option of stipulating prior distributions for α and k.

In order not to make the numerical integration too heavy a burden in the computation of the posterior probability that a contrast is active, discrete distributions are assumed for these parameters. Let $f(\alpha_j)$, $j = 1, \ldots, r$ and $f(k_l)$, $l = 1, \ldots, s$ denote the independent prior probability mass functions of α and k respectively. Continuing with the notation of section 3.2 of Chapter 3, for a particular value of $k = k_l$ the estimated contrasts $U_1, \ldots, U_m$ are considered as a sample of independent random variables with density

$$\begin{aligned} f(U_i|\xi_i, k_l, \tau) &= \xi_i \left[\frac{1}{k_l \tau\sqrt{2\pi}} \exp\left(\frac{-U_i^2}{2k_l^2\tau^2}\right)\right] \\ &+ (1-\xi_i)\left[\frac{1}{\tau\sqrt{2\pi}} \exp\left(\frac{-U_i^2}{2\tau^2}\right)\right], \ (i = 1, \ldots, m) \end{aligned} \tag{6.1}$$

where ξ_i is a random variable that takes the value 1 when the contrast i is active and 0 when it is inert, and whose prior probability distribution is, for a particular value of $\alpha = \alpha_j$

$$f(\xi_i|\alpha_j) = \alpha_j^{\xi_i}(1-\alpha_j)^{(1-\xi_i)}, \quad \xi_i = 0, 1. \tag{6.2}$$

By direct application of Bayes' theorem, the posterior probability that the ith contrast is active given U_i, τ, and the particular values α_j and k_l is

$$\Pr[\xi_i = 1|U_i, \tau, \alpha_j, k_l]$$

$$= \frac{\alpha_j \left[\frac{1}{k_l \tau\sqrt{2\pi}} \exp\left(\frac{-U_i^2}{2k_l^2\tau^2}\right)\right]}{\alpha_j \left[\frac{1}{k_l \tau\sqrt{2\pi}} \exp\left(\frac{-U_i^2}{2k_l^2\tau^2}\right)\right] + (1-\alpha_j)\left[\frac{1}{\tau\sqrt{2\pi}} \exp\left(\frac{-U_i^2}{2\tau^2}\right)\right]}. \tag{6.3}$$

The aim is to compute the unconditional posterior probability that $\xi_i = 1$, $i = 1, \ldots, m$ given $\mathbf{U} = (U_1, \ldots, U_m)$ i.e.

$$\Pr[\xi_i = 1|\mathbf{U}] = \sum_{j=1}^{r}\sum_{l=1}^{s}\int_0^\infty f(\xi_i = 1, \alpha_j, k_l, \tau|\mathbf{U})d\tau$$

$$= \frac{1}{f(\mathbf{U})}\sum_{j=1}^{r}\sum_{l=1}^{s}\int_0^\infty \Pr[\xi_i = 1|\alpha_j, k_l, \tau, \mathbf{U}) f(\mathbf{U}|\alpha_j, k_l, \tau) f(\alpha_j, k_l, \tau)d\tau \tag{6.4}$$

where

$$f(\mathbf{U}|\tau, \alpha_j, k_l) = \prod_{i=1}^{m}\left\{\frac{1}{\tau\sqrt{2\pi}}\left[\frac{\alpha_j}{k_l}\exp\left(\frac{-U_j^2}{2k_l^2\tau^2}\right)\right.\right.$$

$$\left.\left. + (1-\alpha_j)\exp\left(\frac{-U_j^2}{2\tau^2}\right)\right]\right\}, \tag{6.5}$$

$$f(\tau, \alpha_j, k_l) = \frac{1}{\tau} f(\alpha_j) f(k_l) \tag{6.6}$$

and

$$f(\mathbf{U}) = \sum_{j=1}^{r}\sum_{l=1}^{s} f(\alpha_j) f(k_l) \int_0^\infty f(\tau) f(\mathbf{U})|\tau, \alpha_j, k_l)d\tau. \tag{6.7}$$

The evaluation of the integrals can be made with the software written for the standard case. It was found that the desired results were almost obtained when the following prior distributions for α and k were used.

$$f(\alpha) = \begin{cases} 1/3, & \text{for } \alpha = 0.05, 0.25, 0.45 \\ 0, & \text{elsewhere,} \end{cases} \tag{6.8}$$

$$f(k) = \begin{cases} 1/3, & \text{for } k = 5, 25, 45 \\ 0, & \text{elsewhere.} \end{cases} \tag{6.9}$$

The application of the procedure using these priors to example [1] resulted in the three main contrasts being declared active while the other 4 were not. The results of the application to the 27 examples differed from those exhibited in Table 6.2 only for examples [12], for which this version of the procedure resulted in 2 contrasts deemed to be active and [14], for which 2 contrasts were deemed to be active; although in the latter the posterior probability of being active for the third largest contrast is 0.499 so, the result almost coincides with the one presented in Table 6.2

Note that the values for α and k in (6.8) and (6.9) are equally spaced within reasonable intervals and with uniform probabilities. The empirical error rates of this version of the procedure, obtained from 10,000 simulated samples are EPE = 0.052 and PER = 0.18, pretty close to those of the comparison criterion used in Chapter 4.

6.2 Assuming a heavier-tailed distribution for the active contrasts

Another idea motivated by example [1] is to consider that the active contrasts come from a heavier-tailed distribution than the normal, which could better explain the appearance of the largest contrast in that example. Such a distribution would be the Student's t with a small number ν of degrees of freedom and scale parameter τ. Incorporating this assumption, and still considering prior distributions $f(\alpha_j)$, $j = 1,\ldots,r$ for α and $f(k_l)$, $l = 1,\ldots,s$ for k, the estimated contrasts $U_1,\ldots,U_m$ are independent random variables with density

$$\begin{aligned} f(U_i|\xi_i,\tau,k_l) &= \xi_i\left[\frac{\Gamma(\frac{\nu+1}{2})}{k_l\tau\sqrt{2\pi}\Gamma(\frac{\nu}{2})}\left(1+\frac{U_i^2}{\nu k_l^2\tau^2}\right)^{-\frac{\nu+1}{2}}\right] \\ &+ (1-\xi_i)\left[\frac{1}{\tau\sqrt{2\pi}}\exp\left(\frac{-U_i^2}{2\tau^2}\right)\right],\ (i=1,\ldots,m) \qquad (6.10) \end{aligned}$$

where

$$f(\xi_i|\alpha_j) = \alpha_j^{\xi_i}(1-\alpha_j)^{(1-\xi_i)}, \quad \xi_i = 0, 1. \tag{6.11}$$

Proceeding in the same way as in the previous section, the unconditional posterior probability that the ith contrast is active is

$$\Pr[\xi_i = 1|\mathbf{U}] = \sum_{j=1}^{r}\sum_{l=1}^{s}\int_0^\infty f(\xi_i = 1, \alpha_j, k_l, \tau|\mathbf{U})d\tau$$

$$= \frac{1}{f(\mathbf{U})}\sum_{j=1}^{r}\sum_{l=1}^{s}\int_0^\infty \Pr[\xi_i = 1|\alpha_j, k_l, \tau, \mathbf{U}) f(\mathbf{U}|\alpha_j, k_l, \tau) f(\alpha_j, k_l, \tau)d\tau \tag{6.12}$$

where

$$\Pr[\xi_i = 1|U_i, \tau, \alpha_j, k_l]$$

$$= \frac{\alpha_j\left[\frac{\Gamma(\frac{\nu+1}{2})}{k_l\tau\sqrt{2\pi}\Gamma(\frac{\nu}{2})}\left(1+\frac{U_i^2}{\nu k_l^2\tau^2}\right)^{-\frac{\nu+1}{2}}\right]}{\alpha_j\left[\frac{\Gamma(\frac{\nu+1}{2})}{k_l\tau\sqrt{2\pi}\Gamma(\frac{\nu}{2})}\left(1+\frac{U_i^2}{\nu k_l^2\tau^2}\right)^{-\frac{\nu+1}{2}}\right] + (1-\alpha_j)\left[\frac{1}{\tau\sqrt{2\pi}}\exp\left(\frac{-U_i^2}{2\tau^2}\right)\right]}, \tag{6.13}$$

$$f(\mathbf{U}|\tau, \alpha_j, k_l) = \prod_{i=1}^{m}\left\{\frac{\alpha_j}{\tau\sqrt{2\pi}}\left[\frac{\Gamma(\frac{\nu+1}{2})}{k_l\Gamma(\frac{\nu}{2})}\left(1+\frac{U_i^2}{\nu k_l^2\tau^2}\right)^{-\frac{\nu+1}{2}}\right]\right.$$

$$\left. + \frac{(1-\alpha_j)}{\tau\sqrt{2\pi}}\exp\left(\frac{-U_j^2}{2\tau^2}\right)\right\}, \tag{6.14}$$

$$f(\tau, \alpha_j, k_l) = \frac{1}{\tau}f(\alpha_j)f(k_l), \tag{6.15}$$

and

$$f(\mathbf{U}) = \sum_{j=1}^{r}\sum_{l=1}^{s} f(\alpha_j)f(k_l)\int_0^\infty f(\tau)f(\mathbf{U})|\tau, \alpha_j, k_l)d\tau. \tag{6.16}$$

The computation of the posterior probabilities (6.12) can be made by simple modifications to the software written for the standard case, but one has to be careful in specifying approximately the relevant interval for the variable of integration when using the GAUSS routines for numerical integration. In this case

the relevant interval for τ is approximately (0,25) while in the normal case it is approximately (0,10).

If we fix $\alpha = 0.20$ and $k = 10$ (Box and Meyer's recommended values) and use $\nu = 3$, the application of this version of the procedure to the 30 examples, reproduces exactly the results obtained with Box and Meyer's version for the examples [2],...,[30]. The application to example [1] resulted in the two largest contrasts presenting posterior probabilities of being active larger than 0.5, while for the third largest contrast the posterior probability of being active was 0.46, so the desired results were almost reached.

Considering prior densities for α and k it was found that the desired results were obtained for $\nu = 3$ and the following prior distributions for α and k:

$$f(\alpha) = \begin{cases} 1/3, & \text{for } \alpha = 0.10, 0.20, 0.30 \\ 0, & \text{elsewhere,} \end{cases} \tag{6.17}$$

$$f(k) = \begin{cases} 1/3, & \text{for } k = 5, 10, 15 \\ 0, & \text{elsewhere.} \end{cases} \tag{6.18}$$

In fact, the results differed only for example [12], for which the application of this version of the procedure resulted in two contrasts instead of one deemed to be active. However the posterior probability of being active for the second largest contrast was just over 0.5.

6.3 Concluding remarks

One of the 30 examples with seven contrasts suggested that there may be situations in which an especially large contrast may be present along with some other real contrasts of more moderate size. The application of BM to experiments with this feature may result in the smaller real contrasts being overlooked. Some ideas to make the procedure more general and robust in situations like the one suggested

have been explored in this chapter. It has been shown that stipulating prior distributions for the parameters α and k, as well as considering a heavier-tailed distribution for the active contrasts, can effectively help to solve the problem. It was found that the use of the t distribution for the active contrasts alone did make the procedure more robust, but not enough for the extreme situation suggested by example [1].

Chapter 7

Discussion and Conclusions

Factorial experimentation, and in particular unreplicated factorial experiments, plays an important role in today's approach to quality. For this reason, over the last few years there has been an increasing interest in developing procedures to decide which factors have active effects in these experiments in which the usual estimate of the standard error cannot be used.

Most methods have been developed to analyze two-level factor designs, however they can be applied to orthogonal designs involving factors with more than two levels if all the effects are standardized. In the statement of the problem in section 2.1, this idea was formalized. For the standardized effects the more general term *contrasts* has been used throughout this dissertation.

The aim of the research was to find the procedure(s) that can most effectively help experimenters in the industrial field to analyze unreplicated factorials. We decided to study and compare four methods that we thought, after a literature review, were the most promising.

It was found that most of the work in this area has involved simulation studies in which the operating characteristics of the methods are investigated for situations (number and magnitude of real contrasts in an experiment) that have been

arbitrarily chosen by the analysts. However it is not difficult to find examples of real data that suggest situations quite different of those usually included in simulation studies.

Consequently it was decided to analyze the performance of the procedures in the light of the results of applying them to a substantial number of real examples. We were aware of the difficulties that this involves. Using real data we do not know the true numbers and sizes of the real effects. However, each set of data is a reflection of the underlying situation. In this sense, we know that at least we are being realistic.

Most of the data sets found in an extensive search in the literature were from experiments involving only two-level factors, the most common being those with $m = 7$, 15, 31, and 63 contrasts, so the four methods were implemented for these sizes of experiments.

At this point one of the main problems was to obtain critical values for ZA (Zahn's version of Daniel's procedure for half-normal plots) for $m = 7$, 31 and 63. The Daniel plot is constructed by plotting the ordered absolute contrasts $x_{(i)}, \ldots, x_{(m)}$ against their approximate expected values under the assumption of a null experiment. It is common to approximate these expected values by $G^{-1}(p_i)$, where G is the standard half-normal cumulative distribution function, and p_i are conventional probability levels such as $p_i = (i - 1/2)/m$ or $p_i = i/(m + 1)$. When decisions are made on the basis of a visual inspection of the plot, it does not make much difference which of these (or any other sensible choice of the p_i) are used.

For the more formal method ZA, Zahn (1975) used the exact expected values to obtain critical values for $m = 15$. Applying general results by David and Johnson (1954) we derived an expression to obtain, to the desired precision, the expected values of the order statistics of samples from the half-normal distribution. An advantage of having such an expression is that those expected values can be obtained with any statistical package able to provide values of the inverse Normal cumulative distribution function. This facilitates the implementation of ZA (and

HP, the method proposed in Chapter 5).

The method BP proposed by Berk and Picard (1989) was also supplemented with critical values for $m = 63$.

For the four methods, the behaviour of the proportion of false rejections (EPE) and the probability of at least one false rejection (PER) were analyzed empirically for m null contrasts. From the relation between these two values in the four methods, it was seen that, for $m = 31$ and $m = 63$, the use of EPE= 0.05, which is often used, is too permissive, as this is equivalent to use values of PER about 0.70 for $m = 31$, and about 0.90 for $m = 63$, which means that most analyses of null experiments will result in at least one false positive.

It was found that the values tabulated by Lenth (1989) for the method named LE were miscalibrated making this procedure too conservative. This is important since software for the use of this method has been published e.g. by Stephenson (1991), and examples have been analyzed using this method e.g. by Berk and Picard (1991) who seemed to be unaware of the errors in Lenth's values.

New sets of values which effectively control EPE = 0.05 for the margin of error (ME) and PER = 0.05 for the simultaneous margin of error (SME), were obtained empirically and presented in Table 4.2.

Eventually it was possible to set up a comparison criteria matching the values EPE and PER approximately. With this purpose, extra sets of critical values were obtained for some methods.

The results of applying the four methods, using comparable levels, to the sets of data for the four sizes of experiments, showed a healthy harmony. The percentages of experiments for which the four methods were in complete agreement about the contrasts deemed to be significant were, by experiment size, as follows: for $m = 7$, 80%; for $m = 15$, 61%; for $m = 31$, 56%; and for $m = 63$, 67%.

The differences for many of the remaining examples were small and unimportant. However, it was possible to identify practical situations in which some of the

methods will lead to a wrong interpretation.

The most serious problem suggested by the analyses of these examples was the inability of ZA to examine more than 30% of the contrasts in a single experiment. This was evident for the cases with $m = 7$ and $m = 15$. An investigation of various alternatives, in Chapter 5, led to the suggestion of HP as a sensible alternative which, besides overcoming ZA's drawback, seems to be, on the examples analyzed, slightly more powerful than the other procedures for experiments of small and moderate sizes. However this should be confirmed with further research.

Sets of critical values for HP were obtained to make possible the application of this method to ten sizes of recommended designs. In order to facilitate the application of this method a macro in MINITAB code was written. The macro computes the expected values of order statistics of samples from the half-normal distribution with the expression derived in Chapter 3. Those values are used as plotting positions. The half-normal plot is produced with the HP guardrails using PER = 0.05, 0.20 and 0.40 following Daniel's recommendation. The macro HP along with an output is presented in Appendix B.

One of the 30 examples with 7 contrasts suggested that BM, the Bayesian approach proposed by Box and Meyer (1986) may overlook some active contrasts in situations where, there is a very large contrast which makes the others appear to be inert. Some extensions of this method that resolve this particular problem were obtained in Chapter 6. Further investigations using simulations based on situations like the one suggested by example [1], may lead to a more general and flexible procedure.

A final assessment of the methods analyzed is as follows:

1. The method LE proposed by Lenth (1989) is too conservative. It may be recommended only when used with the corrected values (i.e. LE′) presented in Chapter 4.

2. ZA (Zahn's version of Daniel's procedure for half-normal plots) is not recom-

mended for experiments with less than 31 contrasts, since an important proportion of experiments with 7 and 15 contrasts may present more active effects than the number that ZA examines.

3. The methods BP, LE′, BM and HP should produce identical results in most experiments, with HP seemingly slightly more powerful, followed by BM, LE and BP. Applying BM one should be alert to extreme situations like the one suggested by example [1]. HP has the additional advantage that the half-normal plot can be used to examine the contrasts for departures from the basic assumptions as proposed by Daniel (1959).

Discussing the effect sparsity assumption is difficult as this is mainly based on practical experience throughout the years and it does not seem to be possible at this stage, to make a precise definition. The methods analyzed here all assume the sparsity principle, and the investigation which led to HP in Chapter 5 suggests that assuming that about 60% of the smallest contrasts are inert is quite safe. This coincides with the judgement made by Berk and Picard (1991) who reserved about the 60% smallest contrasts for the construction of the baseline for BP. Violation of the effect sparsity assumption however, except in extreme situations, may be spotted by visually inspecting the half-normal plot.

It is important to point out that the procedures analyzed here are intended to be an aid at the stage of deciding which contrasts should be deemed to be active. Careful planning and checking on the assumptions as well as the experience, knowledge and good judgement of the experimenters are essential.

Appendix A

List of examples

Examples with 7 contrasts

[1]. Box and Draper (1987), p.137, 2^3 "martensite start temperature".

[2]. Bennet and Franklin (1954), p.499, 2^3 "spinning band laboratory fractionating column".

[3]. Box, Hunter and Hunter (1978), p.307, 2^3, "pilot plant investigation";

[4]. —, p.354, 2^3, "development of screening facility", response: solids removed, phase (a).

[5]. —, —, response: solids removed, phase (b).

[6]. —, —, response: flow retreated, phase (b).

[7]. —, p.422, 2^{4-1}, "stability of new product".

[8]. —, p.424, 2^{7-4}, "bottleneck at the filtration stage of an industrial plant".

[9]. Daniel (1976), p.54, 2^3, "thickening time of cement".

[10]. Davies (1956), p.258, 2^3, "investigation of a nitration process".

[11]. —, p.454, 2^{4-1}, "preparation of a dyestuff".

[12]. —, p.457, 2^{5-2}_{III} "yield of a medicinal product".

[13]. —, p.511, 2^{5-2}_{III} "chemical reaction".

[14]. —, —, "chemical reaction stage 2".

[15]. —, p.541, 2^3 "yield surface".

[16]. Grove and Davis (1992), p.4, 2^{7-4}_{III} "carburettor assembly".

[17]. —, p.29, 2^3 "sled test".

[18]. —, p.198, 2^{6-3}_{III} "tee clamps, stage 2", location contrasts.

[19]. —, —, dispersion contrasts.

[20]. Johnson and Leone (1977), p.797, 2^3, "strength of steel".

[21]. John (1990), p.307, 2^3 "mullet fish".

[22]. Pignatello and Ramberg (1985), 2^{5-2}_{III} "leaf springs" (response Z).

[23]. Taguchi (1986), p.126 2^{5-2}_{III} "wear in a pump".

[24]. Taguchi (1987), p.166, 2^{7-4}_{III} "Ina Tile", number of defectives.

[25]. Winterbottom (1992), 2^{7-4}_{III} "Ina Tile", size, location contrasts.

[26]. —, —, dispersion contrasts.

[27]. Fearn (1993), 2^{4-1}IV "puff pastry", response: pastry height.

[28]. —, —, response: pastry score.

[29]. —, —, response: pastry length.

[30]. —, —, response: pastry moisture content.

Examples with 15 contrasts

[31]. Box and Draper (1987), p.140, 2^4 "Chang-Konomenko-Franklin", response 1: PDA.

[32]. —, —, response 2: DMP.

[33]. —, —, response 3: PD.

[34]. —, —, response 4: R.

[35]. —, p.172, 2^{6-2}_{IV}, "business at Ozzie's bar".

[36]. —, p.174, 2^{6-2}_{IV} "plywood adhesives".

[37]. —, p.180, 2^{7-3}_{IV} "packing times".

[38]. —, p.184, 2^{5-1}_{V} "process in drug manufacture".

[39]. Bendell, et al. (1989), p.101, 2^{8-4} "epitaxial-process", location contrasts.

[40]. —, —, dispersion contrasts.

[41]. —, p.231, 2^{8-4}_{IV} "refractory properties", green density.

[42]. —, —, fired density.

[43]. —, —, modulus of rupture.

[44]. —, p.257, 2^{15-11}_{III} "Quinlan on shrinkage" (log data), mean.

[45]. —, —, log(S).

[46]. Bennett and Franklin (1954), p.257, 2^4 "moisture content".

[47]. Box, Hunter and Hunter (1978), p.324, 2^4 "process development study".

[48]. —, p.376, 2^{5-1} "reactor example".

[49]. —, p.398, 2^{8-4} "injection moulding".

[50]. —, p.424, 2^{7-3} "bottleneck at the filtration stage".

[51]. —, p.429, 2^{8-4} "model-controller-aircraft system".

[52]. Daniel (1976), p.71, 2^4 "stone drill".

[53]. Davies (1956), p.274, 2^4 "preparation of an isatin".

[54]. —, p.462, 2^{5-1} "quality of a basic dyestuff".

[55]. —, p.466, 2^{5-1} "yield of penicillin".

[56]. Grove and Davis (1992), p.56, 2^{8-4} "moulding of glove box lids".

[57]. —, p.247, 2^{6-2} "welding process for fuel tanks" (untransformed data), response: strength (location contrasts).

[58]. —, —, response: strength (dispersion contrasts).

[59]. —, p.267, 2^{7-3} "oil control piston ring spacers", location contrasts.

[60]. —, —, dispersion contrasts.

[61]. Holms and Berrettoni (1969), 2^{5-1} "stress rupture times".

[62]. Johnson and Leone (1964), p.196, 2^4 "rifle performance".

[63]. John (1990), p.293, 2^4 "manufacture of tires"

[64]. —, p.307, 2^4 "thickness of film",response: average thickness.

[65]. —, —, response: uniformity.

[66]. Montgomery (1992), p.291, 2^4 "filtration rate of a chemical product".

[67]. —, p.301, 2^4 "manufacture of panels".

[68]. —, p.345, 2^{5-1} "manufacture process for an integrated circuit".

[69]. —, p.353, 2^{6-2} "injection moulding process".

[70]. —, p.371, 2^{7-3} "eye focus time".

[71]. —, p.385, "wine testing" (mean only).

[72]. Pignatello and Ramberg (1985), 2^{6-2}, "leaf springs" (response Z_1).

[73]. Stowe and Mayer (1966), Plackett and Burman "quality of a catalyst".

[74]. Taguchi (1987), p.189, 2^{9-5} "electric welding", response: tensile strength.

[75]. —, —, response: elongation.

[76]. —, p.217, 2^{9-5} "sodium silicate".

[77]. —, p.417, 2^{8-4} "Ina Tile" (monetary yield).

[78]. —, p.431, 2^{10-6} "dewaxing apparatus", response: increase of yield.

[79]. —, —, response: coagulation point.

[80]. Taguchi and Wu (1980), 2^{9-5} "tensile strength".

[81]. Fearn (1993), 2^4 "puff pastry", response: pastry specific height.

[82]. —, —, response: pastry length.

[83]. —, —, response: pastry shrinkage.

[84]. —, —, response: pastry eccentricity.

Examples with 31 contrasts

[85]. Bennett and Franklin (1954), p.585, 2^{6-1} "yield of alfalfa".

[86]. Box, Hunter and Hunter (1978), p.376, 2^5 "reactor example".

[87]. Cochran and Cox (1957), p. 2^5 "texture on icing in cakes".

[88]. Daniel (1976), p.129, "Yates' 2^5 on beans".

[89]. —, p.136, "Davies' 2^5 on penicillin".

[90]. Grove and Davis (1992), p.224, 2^{9-4} "seat belt".

[91]. Johnson and Leone (1977), p.801, 2^5 "residual acidity in a purification process".

[92]. Kempthorne (1952), p.269, 2^5 "Rothamsted's on mangolds".

[93]. Montgomery (1992), p.363, 2^{8-3} "machine's blade profile".

Examples with 63 contrasts

[94]. Box and Draper (1987), p.115, 2^6 "study of dyestuffs manufacture", response: strength.

[95]. —, —, response: hue.

[96]. —, —, response: brightness.

[97]. Taguchi (1987), p.444, 2^{16-10} "wool washing and carding".

[98]. Fearn (1993), 2^{8-2} "puff pastry", response: pastry height.

[99]. —, —, response: pastry length.

[100]. —, —, response: pastry width.

[101]. —, —, response: pastry score.

[102]. —, —, response: pastry moisture content.

Appendix B

MINITAB macro HP

This macro has been design for MINITAB, Release 8. It produces a half-normal plot with HP guard rails for m contrasts using the appropriate sets of critical values given in Table 5.5.

Figure B.1 shows the output produced by the macro for the glove box lid data analyzed in Chapter 5.

The macro uses 12 consecutive columns of the MINITAB worksheet. The first 4 of these columns are for the user to set the critical values and data, the rest are used by HP for computations. The macro also uses constants k47 to k60 as well as the constant k100 where MINITAB (Release 8) stores the constant π.

The macro is used as follows:

1. Store the three columns of critical values from the appropriate subtable in Table 5.5 as the first three columns (e.g. c1, c2, c3).
2. Store the contrasts (in any order) in the fourth column (e.g. c4).
3. Set the constant k47 equal to the number of the column where the contrasts have been stored (e.g. let k47=4).
4. Execute macro HP.

```
note Macro begins
brief 0
note Prepares column numbers to be used (k47-k58)
let k47=k50-3
let k58=k50+8
set ck58
k47:k58
end
copy ck58 k47-k58
note Computes the expected values of order statistics (HEOS)
let k60=count(ck50)
set ck51
1:k60
end
let ck51=ck51/(k60+1)
let ck52=(1+ck51)/2
invcdf ck52 ck52
let ck53=(1-ck51)
let ck54=ck51*ck53
let k59=(k60+2)
let ck55=sqrt(2/k100)*expo(-ck52**2/2)
let ck56=ck52+ck54*ck52/(2*k59*ck55**2)
let ck56=ck56+ck54*(ck53-ck51)*(1+2*ck52**2)/(3*k59**2*ck55**3)
let ck56=ck56+ck54**2*ck52*(7+6*ck52**2)/(8*k59**2*ck55**4)
let ck51=(1/1000)*round(1000*ck56)
erase ck52-ck56
note Makes half-normal plot with HP guardrails
let k61=round(0.6*k60)
let k62=k61+1
let ck50=sort(abs(ck50))
copy ck50 ck52;
use 1:k61.
copy ck51 ck53;
use 1:k61.
copy ck51 ck58;
use k62:k60.
noconstant.
regres ck52 1 ck53;
coeficients ck54.
let ck55=ck47*ck54(1)
let ck56=ck48*ck54(1)
let ck57=ck49*ck54(1)
GPlot ck50 ck51;
title 'Half-normal plot with HP guard rails';
XLabel 'expected order statistics';
YLabel 'absolute contrasts';
Line 0 2 ck55 ck58;
```

```
Line 0 7 ck56 ck58;
Line 0 3 ck57 ck58.
let k50=ck54(1)
erase ck52-ck58
brief 2
note end of macro
```

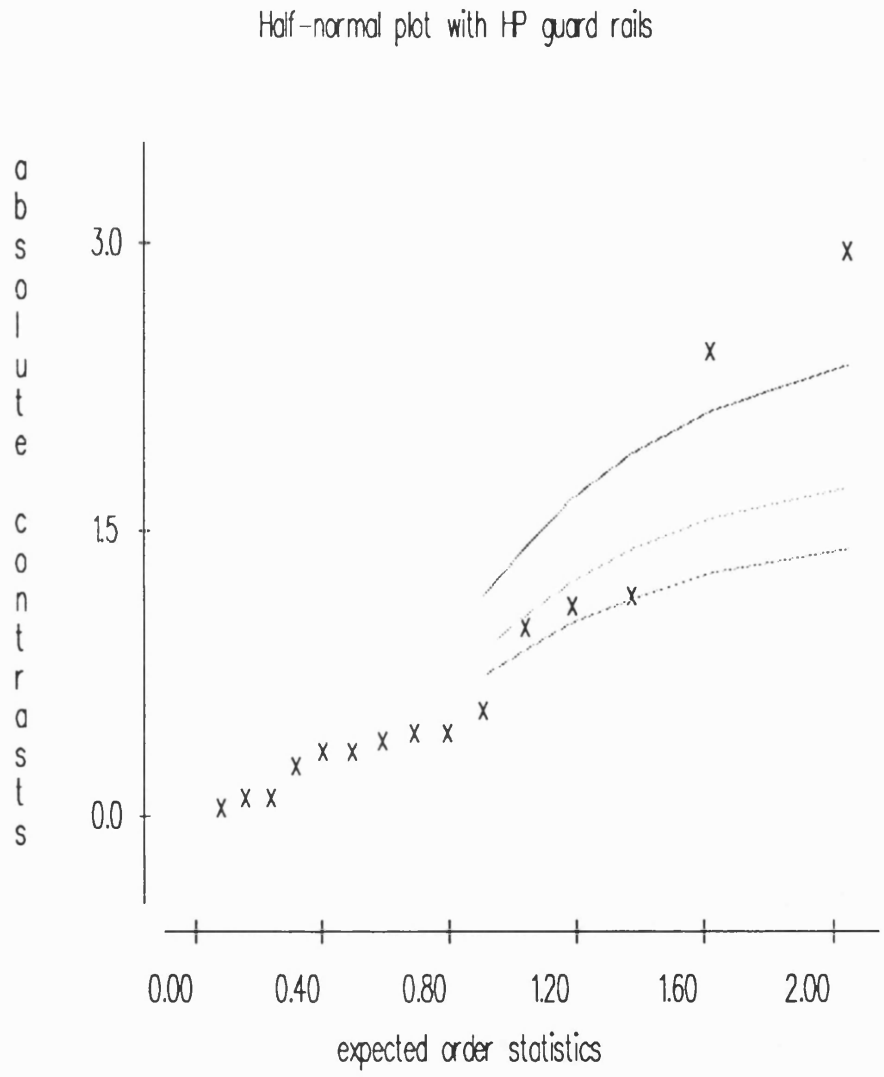


Figure B.1: Output of macro HP for the example [56] analyzed in Chapter 5

References

Bartlett, M. S. (1937). Properties of Sufficiency and Statistical Tests, *Proceedings of the Royal Society, Series A* **160**, 268–282.

Bendell, A., Disney, J. and Pridmore, W. A. (Eds.) (1989). *Taguchi Methods: Applications in World Industry.* IFS Publications, Bedford, U.K.

Bennett, C. A. and Franklin, N. L. (1954). *Statistical Analysis in Chemistry and the Chemical Industry.* John Wiley, New York.

Benski, H. C. (1989). Use of a Normality Test to Identify Significant Effects in Factorial Designs, *Journal of Quality Technology*, **21**, 174–178.

Berk, K. and Picard, R. (1991). Significance Test for Saturated Orthogonal Arrays, *Journal of Quality Technology*, **23**, 79–89.

Bissell, A. F. (1989). Interpreting Mean Squares in Saturated Fractional Designs, *Journal of Applied Statistics* **16**, 7–18.

Box, G. E. P. (1949). The General Distribution Theory of a Class of Likelihood Criteria, *Biometrika* **35**, 318–335.

Box, G. E. P. (1988). Signal-to-Noise ratios, Performance Criteria, and Transformations (with discussion). *Technometrics*, **30**, 1–17.

Box, G. E. P. and Bisgaard, S. (1987). The Scientific Context of Quality Improvement. *Quality Progress*, **20**, No. 6, 54–61.

Box, G. E. P., Bisgaard, S. and Fung, C. (1988). An Explanation and Critique of Taguchi's Contributions to Quality Engineering. Report No. 28, Center for Quality and Productivity Improvement, University of Wisconsin, Madison, Wisconsin.

Box, G. E. P. and Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces.* John Wiley & Sons, Inc., New York.

Box, G. E. P., Hunter, W. G. and Hunter, J. S. (1978). *Statistics for Experimenters.* John Wiley, New York.

Box, G. E. P. and Jones, S. P. (1986). An Investigation of the Method of Accumulation Analysis. Technical Report 19, Center for Quality and Productivity Improvement, University of Wisconsin, Madison, Wisconsin.

Box, G. E. P. and Meyer, R. D. (1986). An Analysis for Unreplicated Fractional Factorials. *Technometrics*, **28**, 11–18.

Byrne, D. M. and Taguchi, S. (1989). The Taguchi Approach to Parameter Design. In *Taguchi Methods: Applications in World Industry*, (A. Bendell, J. Disney, and W. A. Pridmore, Eds.), pp. 57–76. IFS Publications. Bedford, U.K.

Cochran, W. G. (1941). The Distribution of the Largest of a Set of Estimated Variances as a Fraction of Their Total. *Annals of Eugenics* **11**, 47–52.

Cochran, W. G. and Cox, G. M. (1957). *Experimental Designs.* Wiley, New York.

Cox, D. R. (1990). Quality and Reliability: Some Recent Developments and a Historical Perspective. *Journal of the Operational Research Society.* **41**, 95–101.

Daniel, C. (1959). Use of Half Normal Plots in Interpreting Factorial Two-Level Experiments. *Technometrics*, **1**, 311–341.

Daniel, C. (1976). *Applications of Statistics to Industrial Experimentation.* Wiley, New York.

David, H. A. (1970). *Order Statistics.* Wiley, New York

David, F. N. and Johnson, N. L. (1954). Statistical Treatment of Censored Data. Part I. Fundamental Formulae. *Biometrika*, **41**, 228–240.

Davies, O. L. (Ed.) (1956). *The Design and Analysis of Industrial Experiments*, (2nd. ed.), Oliver and Boyd, Edinburgh.

Deming, W. E. (1982). *Quality, productivity and competitive position.* Massachusetts Institute of Technology Center for Advanced Engineering Study, Cambridge, Massachusets.

Dueker, M. S. (1988). Saturated Designs. In *Encyclopedia of Statistical Sciences*, (S. Kotz and N. L. Johnson, Eds.), Vol. 8, Wiley, New York.

Ealey, L. A. (1988). *Quality by Design. Taguchi Methods and U. S. Industry.* ASI Press, Dearborn, Michigan.

Fearn, T. (1993) Personal communication.

Finney, D. J. (1945). The Fractional Replication of Factorial Arrangements. *Annals of Eugenics*, **12**, 291–301.

Fisher, R. A. (1942). The Theory of Confounding in Factorial Experiments in Relation to the Theory of Groups. *Annals of Eugenics*, **11**, 341–353.

Garvin, D. A. (1988). *Managing Quality.* The Free Press, New York.

Grove, D. M. and Davis, T. P. (1992) *Engineering, Quality and Experimental Design.* Longman, Essex, UK.

Gunter, B. (1987). A Perspective on the Taguchi Methods *Quality Progress*, **20**, No. 6, 44-52.

Hamada, M. and Wu, C. F. J. (1991). Analysis of Censored Data From Highly Fractionated Experiments. *Technometrics*, **33**, 25–38.

Hoaglin, D. C. (1983). Letter Values: a Set of Selected Order Statistics. In *Understanding Robust and Exploratory Data Analysis*, (D. C. Hoaglin and J. W. Tukey, Eds.), John Wiley & Sons, New York, N.Y.

Holms, A. G. (1966). *Multiple-Decision Procedures for the ANOVA of Two-Level Factorial Replication-Free Experiments.* PhD Thesis, Western Reserve University, June 1966.

Holms, A. G. and Berrettoni, J. N. (1969). Chain-Pooling ANOVA for Two-Level Factorial Replication-Free Experiments, *Technometrics* **11**, 725–746.

Hunter, J. S. (1987). Signal to Noise Ratio Depicted. *Quality Progress*, **20**, No. 5, 7–8.

John, P. W. M. (1990) *Statistical Methods in Engineering and Quality assurance.* Macmillan, New York.

Johnson, N. L. and Leone, F. C. (1964). *Statistics and Experimental Design*, Vol. 2. Wiley, New York.

Johnson, N. L. and Leone, F. C. (1977). *Statistics and Experimental Design in Engineering and the Physical Sciences*, Vol. 2, 2nd ed., Wiley, New York.

Kempthorne, O. (1952). *The Design and Analysis of Experiments.* John Wiley & Sons, Inc., New York.

Lenth R. P. (1989). Quick and Easy Analysis of Unreplicated Factorials, *Technometrics* **31**, 469–473.

Logothetis, N. and Wynn, H. P. (1989). *Quality Through Design*, Oxford Science Publications, Oxford.

Miller, R. G. (1966). *Simultaneous Statistical Inference.* McGraw-Hill Book Company, New York.

Montgomery, D. C. (1992) *Design and Analysis of experiments*, 3rd edn. Wiley, New York.

Nair, V. N. (1986). Testing in Industrial Experiments With Ordered Categorical Data (with discussion). *Technometrics*, **28**, 283–291.

Nair, V. N. (Ed.) (1992). Taguchi's Parameter Design: A Panel Discussion. *Technometrics*, **34**, 127–161.

Olson, D. M. (1979). A Small-Sample Test for Non-Normality, *Journal of Quality Technology* **11**, 95–99.

Pignatello, J. J. Jr. and Ramberg, J. S. (1985) Discussion of "Off-Line Quality Control, Parameter Design, and the Taguchi Method." *Journal of Quality Technology*, **17**, 198–206.

Plackett, R. L. and Burman, J. P. (1946). The Design of Optimum Multifactorial Experiments. *Biometrika*, **33**, 305–325.

Rao, C. R. (1946). On Hypercubes of Strength d and a System of Confounding in Factorial Experiments. Bulletin of the Cal. Mathematical Society, **38**, 67.

Rao, C. R. (1947). Fractional Experiments Derivable From Combinatorial Arrangements of Arrays. *Journal of the Royal Statistical Society (Supplement).* **9**, 128–139.

Shoemaker, A. C. Tsiu, K. L. and Wu. C. F. J. (1990). Economical Experimentation Methods for Robust Design. *Technometrics*, **33**, 415–427.

Stephens, M. A. (1986). Tests for the Uniform Distribution. In *Goodness-of-Fit Techniques.* (R. B. D'Agostino and M. A. Stephens, Eds.), Vol. 68, Marcel Dekker Statistics Textbooks and Monographs, New York, NY.

Stephenson, W. R. (1991). A Computer Program for the Quick and Easy Analysis of Unreplicated Factorials. *Journal of Quality Technology*, **23**, 63–67.

Stowe, R. A. and Mayer, R. P. (1966) Efficient Screening of Process Variables. *Industrial and Engineering Chemistry*, **59**, 36–40.

Taguchi, G. (1986). *Introduction to Quality Engineering.* Asian Productivity Organization and UNIPUB, White Plains, N.Y.

Taguchi, G. (1987). *System of Experimental Design*, Vol 1. Unipub, White Plains, NY.

Taguchi, G. and Wu, Y. (1980) *Introduction to Off-Line Quality Control.* Central Japan Quality Control Association, Nagoya, Japan.

Tippett, L. H. C. (1935). Some Applications of Statistical Methods to the Study of Variation of Quality in the Producction of Cotton Yarn (with discussion). *Journal of the Royal Statistical Society (Supplement).* **II**, 27–55.

Voss, D. T. (1988). Generalized Modulus-Ratio Tests for Analysis of Factorial Designs with Zero Degrees of Freedom for Error, *Communications in Statistics–Theory and Methods* **17**, 3345–3359.

Winterbottom, A. (1992). The Use of a Generalized Signal-to-Noise Ratio to Identify Adjustment and Dispersion Factors in Taguchi Experiments. *Quality and Reliability Engineering International*, **8**, 45–56.

Yates, F. (1935). Complex Experiments. *Journal of the Royal Statistical Society (Supplement).* **II**, 181–247.

Zahn, D. A. (1975a). Modifications and Revised Critical Values for the Half-Normal Plot, *Technometrics* **17**, 184–200.

Zahn, D. A. (1975b). An Empirical Study of the Half-Normal Plot. *Technometrics* **17**, 201–211.