Design of Experiments

If you are carrying out a survey, or monitoring a process using a control chart, the idea is to analyze the situation without changing anything. The essential feature of an experiment, on the other hand, is that the experimenter intervenes to see what happens. There are two main reasons for doing this: to investigate things which would not happen without the experimenter's intervention, or to disentangle cause and effect. Some of the areas where experiments are widely used include:

- The design of products and processes in manufacturing
- Evaluation of websites and internet tools of various types
- Drug trials to evaluating the effectiveness and risks of new pharmaceutical products

As an example, consider a plastic injection moulding process which may create parts that shrink too much. We cannot directly observe what is occurring to cause the shrinkage. Experience and reference works may tell us that several factors may be responsible, such as mould temperature, injection speed, type of plastic resin used, and so on. An experiment will determine which of these factors affect shrinkage the most.

There are also areas where experiments should perhaps be used more widely than they are at present – for example:

- Testing management and educational innovations
- Comparing different social policies

Ayres (2007, chapters 2 and 3) gives a non-technical, and very enthusiastic, account of several examples where designed experiments have proved their worth.

The key thing about an experiment is that we design different treatments, or sets of conditions or websites, or whatever, and then compare the effects of each. For example, we might compare several different versions of a website or manufactured product, or we might compare several different ways of managing a process. This helps to determine which factors (variables) are most important, how they interact, and what the optimum settings are for those factors (variables).

In this session I will start with the design of products in manufacturing, and then expand the discussion to the other areas of application. There is a well developed area of theory about both the design and analysis of experiments. The aims of this session are to introduce you to both of these. Please bear in mind that the theory in both areas is very extensive and this session is just an introduction to some of the basic ideas.

There is a very large literature on experimental design – see for example Ferguson and Dale (2007) on industrial experiments, Trochim (2006) on experiments in social research, Wood (2003, pages 168-172)

for a brief introduction to the logic and purposes of experiments, and Ayres (2007, chapters 2 and 3) for some interesting examples of the value of experiments.

Traditional One-Factor-at-a-time Approach to Experimentation

This is the simplest type of experiment. It involves varying one factor or variable, keeping all other factors (or variables) in the experiment fixed. For instance, consider two factors (say, A and B) each of which can be at one of two levels: level 1 and level 2 (a level is a value or setting of a factor – see Table 2 below for an example) as shown in Table 1. In the first experimental trial, it is obvious that we keep all factors at level 1. In the second trial, only the level of factor A has changed to level 2, keeping the level of factor B constant (i.e. at level 1).

Table 1 One-factor-at-a-time method

Experimental Trial or Run	*Factor A*	*Factor B*
1	1	1
2	2	1

The difference in the results between these two experimental runs in Table 1 provides an estimate of the effect of A. An effect here refers to the change in output (e.g., thickness, weight, efficiency, strength, etc.) which we measure during the experiment due to the change in factor levels (i.e., level 1 to level 2). This effect has been estimated when factor B was at level B (1), and therefore there is no guarantee whatsoever that A will have the same effect for other levels of B.

The one-factor-at-a-time approach of experimentation can be misleading and often leads to wrong conclusions. To obtain a more realistic answer, we need to find out the effect of each factor in conjunction with different levels of the other factors. We can achieve this by designing an experiment where we change the levels of factors simultaneously to study their effect on output. Sir Ronald A Fisher in the early 1920s developed some methods for effective experimentation which were a fundamental break from the old scientific tradition of varying only one-factor-at-a-time. His initial experiments were concerned with determining the effects of various fertilisers on plots of ground. Fisher used methods of statistical experimental design and analysis to draw conclusions about the effect of each fertiliser on the final condition of the crop. More recently, these experimental design techniques have been widely accepted in manufacturing organisations for improving product and process performance. We will explain the basic ideas using the scenario below as an illustration.

Factors, levels and response (performance) variables

The *response (performance) variable* is the outcome that we are interested in changing or controlling (yield in the example below).

The response variable used to assess quality levels will obviously be different in different situations. There are three possible types. It is very important to bear this in mind when designing and analysing an experiment:

a *Smaller-the-Better quality characteristics (e.g. number of defects or complaints)*

b *Larger-the Better quality characteristics (e.g. a customer rating)*

c *Target-is-the-Best quality characteristics (e.g. the dimension of a manufactured component)*

Factors are things (variables) which we think might influence the response variable (Pressure and Temperature in the example below).

Each factor is set at two or more *levels* – often coded as 1 and 2.

Let's see how this works in a particular example.

A chemical process

Suppose we are interested in finding the yield of a chemical process at two pressures, P(1) and P(2), and at two temperatures, T(1) and T(2).

Table 2 – List of factors and their levels

Factors	Level 1	Level 2
Pressure	1.5 bar*	2 bar
Temperature	60°C	80°C

* The bar is a unit of pressure.

We will first use the one-factor-at-a-time approach. The experiment was repeated twice and the average yields were calculated as in Table 3.

Table 3: One factor at a time experiment on the chemical process

Experimental Trial or Run	*Pressure*	*Temperature*	*Average yield (Kg)*
1	1	1	51
2	2	1	61
3	1	2	56

We can now use this table to estimate the effects of both Pressure and Temperature. Comparing the first two rows, the effect of changing Pressure from Level 1 to Level 2 is 10 Kg. Similarly comparing rows 1 and 3 suggests that the effect of changing Temperature from Level 1 to 2 is 5 Kg. In each case Level 2 is better (more yield) but the effect of Pressure is greater.

Here we have the average yield values corresponding to only three combinations of temperature and pressure. The experimenter concluded from the above data that the maximum yield of chemical process will be attained from P(2) and T(1). But the question then arises as to what should be the average yield corresponding to P(2) and T(2)? If the effect of the higher level of Pressure is 10Kg extra Yield, and the effect of the higher level of Temperature is 5Kg extra Yield, then we might guess that the higher level of both would lead to 15Kg extra Yield. However, we cannot be sure because this combination is not tested in the experiment. The difficulty is that there may be an *interaction* between the two factors. To find out we need to extend the experiment so that it becomes a *full factorial* experiment (about which more below).

Interactions

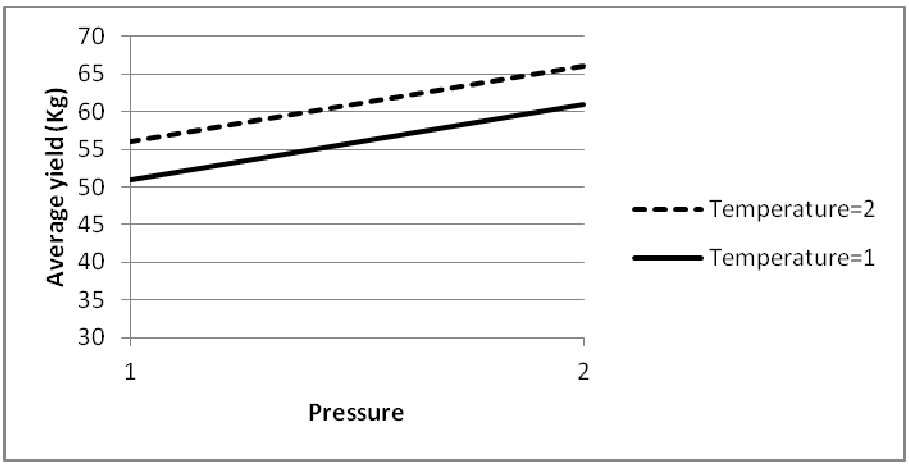
Two factors are said to interact with each other if the effect of one factor on the output depends on the level(s) of the other factor(s). I'll show you why this is important by considering two possible scenarios.

Table 4: Full factorial experiment on the chemical process – First possible scenario

Experimental Trial or Run	*Pressure*	*Temperature*	*Average yield (Kg)*
1	1	1	51
2	2	1	61
3	1	2	56
4	2	2	66

Here the two factors act independently of each other: there is no interaction and we can just add the effects of the two factors to find the difference between the first and the fourth row of Table 4. Interactions are conveniently analyzed by means of a graph:

Figure 1. Graph to show the interaction between Temperature and pressure – First possible scenario



The top line in this graph shows the effect of Pressure (i.e. the difference between the two levels) for the top level (2) of Temperature, T(2), and the bottom line shows the effect of Pressure for the bottom level (1) of Temperature. The fact that they are parallel indicates that the effect is the same (+10 Kg) for both levels of Temperature, so there is no interaction – the level of Temperature makes no difference to the effect of Pressure.

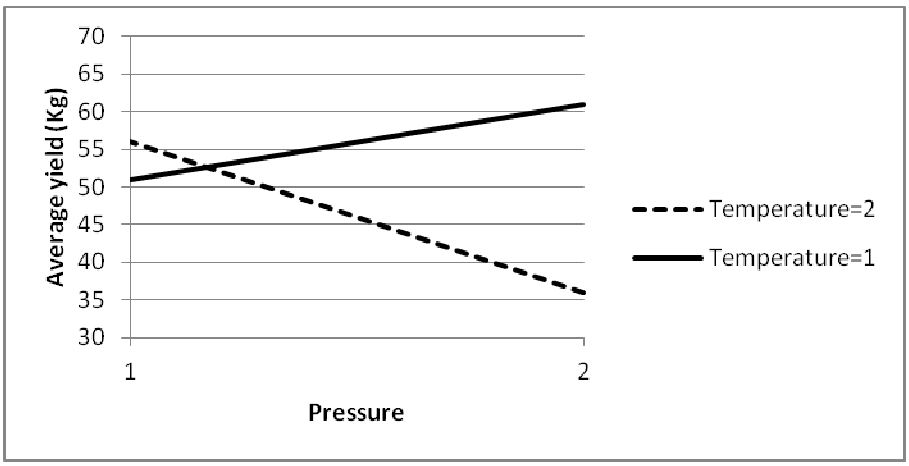
This graph is produced by the spreadsheet at http://woodm.myweb.port.ac.uk/interaction.xls . This is interactive and you should be able to change the response variable in the green cells and see what the interaction graph looks like.

The situation is very different for Table 5 and Figure 2. Here the lines are not parallel and the effect of Pressure is very different at the two levels of Temperature. The higher level of Pressure increases yield by 10Kg at the low level of Temperature, T(1), but at the high level of Temperature, T(2), the effect is the opposite – a *reduction* in the yield by 20Kg. The combination of high pressure and high temperature has a disastrous effect on the yield.

Table 5: Full factorial experiment on temperature and pressure – Second possible scenario

Experimental Trial or Run	*Pressure*	*Temperature*	*Average yield (Kg)*
1	1	1	51
2	2	1	61
3	1	2	56
4	2	2	36

Figure 2: Graph to show the interaction between Temperature and pressure – Second scenario



For complex manufacturing processes in today's industrial environment, interactions play an important role and therefore should be studied for achieving sound experimental conclusions. Therefore one may go for the *factorial experiments* recommended by Fisher so that both main factor effects (i.e., effect of temperature and pressure on the yield) and interaction effects can be studied. Factorial experiment can be of two types: *Full factorial* experiment and *Fractional factorial* experiment. We start with full factorial experiments.

Full factorial experiments

A full factorial experiment is an experiment which enables one to study all possible combinations of factor levels. For full factorial experiments, the experimenter must vary all factors simultaneously and therefore permit the evaluation of interaction effects. Two level experiments are the most widely used factorial experiments in industry – like the one above but often involving many more than two factors.

The full factorial experiment at two levels is generally represented by 2^k, where 2 stands for the number of levels and k, the number of factors to be studied. For example, if the number of factors to be studied is 3, then there are 8 different possible combinations of factor levels (2^3), so a full factorial experiment needs 8 runs or trials as in Table 6.

Table 6: Full factorial experiment with three factors at two levels

Experimental Trial or Run	Factor 1	Factor 2	Factor 3
1	1	1	1
2	1	1	2
3	1	2	1
4	1	2	2
5	2	1	1
6	2	1	2
7	2	2	1
8	2	2	2

Obviously four factors would need 2^4 or 16, and so on. The big advantage of full factorial experiments, of course, is that we can study interactions and see the effect of combinations of factor levels.

Main effects

The *main effect of a factor* is defined as

Main effect = Average response at high level of the factor – Average response at low level of the factor

For Table 4 above the main effect of Pressure for example is +10. The average (mean) of the two runs at the higher level (61 and 66) are 63.5, and for the lower level (51 and 56) is 53.5, so the effect of Pressure – in the sense of using level 2 instead of Level 1 – is +10 Kg of yield. Similarly the effect of Temperature is +5.

The main effect is only useful if the effect of each factor does not depend substantially on the level of the other factor: i.e. there is no substantial interaction between the factors. The way to check if this condition is reasonable, or whether there is an interaction between the factors, is to look at the interaction plot. This is Figure 1 above: the parallel lines show there is no interaction and the effect of one factor (Pressure) is the same at both levels of the other factor.

Table 5 and Figure 2 are different. Here the main effect of Pressure is -5, and Temperature is -10, indicating that the Level 2 yield is, on average, less than the Level 1 yield. However the strong interaction means that the main effects are not very useful because, as Figure 2 shows, the effect of Pressure is positive for the low level of Temperature, T(1), but negative for the high level, T(2). The main effect is the average, which is not very useful in this case.

Exercises

1 Table 7 gives another set of results for a another chemical process

Table 7: Two factor, two level full factorial experiment on another chemical process

Experimental Trial	T: Temperature	P: Pressure	Yield (Kg)
1	1	1	82
2	2	1	93
3	1	2	80
4	2	2	88

Draw an interaction graph and comment on the interaction.

Work out the main effects of the two factors. Are these useful in view of the interaction?

2 How many experimental runs would be needed for a full factorial experiment with 8 factors at 2 levels? What about 3 factors at 4 levels? Could you use these experiments to work out the main effects of all the factors?

3 Three women have entered a marathon. They decide to do an experiment to compare three approaches to training for the marathon. The first runner, Sue, runs for one hour, three times a week (Monday, Wednesday and Saturday). The second, Sharon, runs for an hour every day of the week. The third, Shirley, does a short run (15 minutes) three times a week. After their training, their times in the marathon were: Sue – 5 hours, Sharon – 4 hours 30 minutes, Shirley – 4 hours. (Shirley got the best result and Sue the worst.) They conclude that Shirley's training is the most effective, and decide that they will all follow Shirley's training regime when they prepare for their next marathon.

This is not a very good experiment! What are the problems and how would you improve it?

Replication and statistical significance

This is the process of repeating the experimental trials to improve the accuracy of experimentation. The analysis procedure is similar to experiments without replication. As an example, suppose that each of the runs in Table 7 were replicated an additional two times giving the results in Table 8. Note that the three figures in the final column correspond to the yield from three runs with the same factor levels.

Table 8: Two factor, two level full factorial experiment on a chemical process with three replications

Experimental Trial	T: Temperature	P: Pressure	Yield (Kg)
1	1	1	82, 84, 82
2	2	1	93, 92, 93
3	1	2	80, 80, 81
4	2	2	88, 86, 87

The main effects and the interactions can now be analysed in just the same way as before, except that

we use the average of the three yields at each combination of factor levels in place of a single value. For example, the average yield for the first combination of factor levels is 82.7.

One difficulty with experiments like this is that the effects and interactions observed may be due to essentially random factors. Are the results real? Or would the next set of trials yield very different results?

Intuitively the answers to these questions often seem fairly clear – although this is an area where intuition may be an unreliable guide. For example, in Table 8, each of the three experimental runs under each experimental condition are very similar, whereas the runs under different conditions are very different. This all suggests that further replications would lead to the same pattern.

As a contrast, if the three replications for the first experimental condition had been 83, 42, 123, and the remaining nine runs had been similarly varied, we would then have had much less confidence in conclusions based on the average yield from three runs. Table 9 below gives an example of this. The numbers are arranged so that means of the three replications for each experimental condition are identical to those for Table 8. This means that the main effects and the interaction will be just the same as for Table 8. However, the fact that the data are so much more variable means that the conclusions about the main effects and interaction are obviously far less certain.

Table 9: Response table with the same main effects and interactions as Table 8, but with more variable data

	P(1)	P(2)
T(1)	83	80
	42	90
	123	71
T(2)	100	128
	93	46
	84	87

The usual way this difference between Tables 8 and 9 is assessed statistically is by means of *an analysis of variance* – the abbreviation ANOVA is widely used for this. This is a method of testing the null hypothesis that the two factors have no effect on the response variable. The basic ideas are explained in the part of this course on hypothesis testing (http://woodm.myweb.port.ac.uk/q/StatNotes3.pdf). ANOVA is one of the many ways of working out significance levels or *p* values.

Table 9 is in a different format from Table 8 because Table 9 is in the format Excel needs to perform a *two factor ANOVA with replication* (which you will find under Data Analysis in the Data menu if the Analysis ToolPak is installed). The output from this procedure is detailed and technical, and will not be

covered here with the exception of the *p-values.* These are the main answers produced by an analysis of variance. They are:

Table 8:	p-value for main effect F =	0.000
	p-value for main effect M =	0.000
	p-value for interaction =	0.017

Table 9:	p-value for main effect F =	0.644
	p-value for main effect M =	0.827
	p-value for interaction =	0.931

These p-values are the probabilities of getting results "like" the actual results *if there were no effect and the pattern in the data is just due to chance*. So, for Table 8, if there really were no effect, and the results were just due to chance, the results obtained would be very unlikely indeed – and this is reflected in the low p-values. On the other hand, for Table 9, if there really were no effect and the results were due to chance, the results obtained are entirely plausible – which is reflected in the much higher p-values.

Notice that the interpretation of p-values is the opposite of what you might expect. The *lower* the p-value, the *stronger* the evidence that the effect is real (and not simply a matter of chance). Conventionally, 5% is often taken as the dividing line: p-values less than 5% are described as *significant* (ie signifying a genuine effect), and those more than 5% are *not significant* (ie the evidence is not strong enough). All three p-values for Table 8 are significant ($p<0.05$ for all three), but none of the p-values for Table 9 are significant.

Exercises

4 Suppose that the yields from the fourth experimental condition in Table 8 were 128, 126, 127 (instead of 88, 86, 87).

Calculate the main effects from this data, and also draw an interaction graph.

Do the results show an interaction? If so, describe it.

What do you think the p-values for the main effects and the interaction are? (It is not normally possible to work this out in your head, but in this example the situation is so clear that you should be able to come up with a good guess.)

5 Can you think of any problems in your home or work life where a full factorial experiment might be useful?

Fractional Factorial Experiments at Two Levels

The difficulty with full factorial experiments with a large number of factors is that the number of experimental runs may become too large. For example, if you want to study seven factors for a certain experiment and you choose a full factorial experiment, then you have to perform $2^7 = 128$ experimental runs.

The solution is to ignore some of the possible combinations of factors and study only a fraction of them. This is known as a *fractional factorial experiment*. For example if you study one sixteenth of the possible combinations you will only have to study 8 combinations of factor levels - this is a possibility illustrated by one of the case studies below. The difficulty, of course, with such highly fractional factorial designs is that you will not be able to investigate many of the interaction effects. The next example illustrates some of the issues.

A baking experiment

Baker's paradise is a newly established baking school. Despite continuous efforts, the bakery had failed to produce cakes which the customers liked. The management was looking for the combination of ingredients which would produce the nicest cakes. A project was initiated to study this problem. After a brainstorming session of 2 hours, it was decided that the experiment would include six factors. The factors which were considered for the experiment are shown in Table 6. Each factor was kept at two levels and the goal was to determine the factor-level combination yielding the nicest cakes.

Table 10 - List of factors for the Cake Baking Experiment

Factors/variables	Notation	Level 1	Level 2
Milk (cups)	M	¼	½
Sugar (cups)	S	½	¾
Eggs	E	2	3
Flour (cups)	F	¾	1
Oven Temperature (^{0}C)	O	200	225
Butter (cups)	B	¼	1/2

A full factorial experiment would require $2^6 = 64$ experimental runs. Because of limited time and resources, a *fractional factorial* experiment with eight runs was selected. The design of these experiments needs some care – this is discussed in the section on Orthogonal arrays below. These orthogonal arrays were used to decide on the idea of having eight runs, and on the factor levels used for each of them.

Each run was evaluated by asking a panel of customers to rate each set of cakes on a 1 (very nasty) to 10 (very nice) scale. The mean customer ratings – the response or performance variable – from each run

are shown in the Table 11.

Table 11 - Response Table for the Cake Baking Experiment

Experimental run	M	S	E	F	O	B	Mean customer rating
1	1	1	1	2	2	1	5.5
2	2	1	1	1	2	2	5.8
3	1	2	1	2	1	2	6.5
4	2	2	1	1	1	1	6.0
5	1	1	2	1	1	2	6.2
6	2	1	2	2	1	1	7.2
7	1	2	2	1	2	1	6.2
8	2	2	2	2	2	2	7.3

This example is a simplified, artificial one to demonstrate how experimental design works. In a real situation, you would need to carry out research (involving, perhaps, a brainstorming session with the people involved with the process) to make sure that:

- *the list of factors is appropriate, and*
- *the levels are reasonable ones to try (obviously some levels may be quite obviously inappropriate – these do not need to be considered in the experiment), and*
- *the response variable – mean customer rating in this example – is a sensible one. It is possible, for example, that there may be different groups of customers with different tastes, and it may be more helpful to do a separate analysis for each of these groups.*

Calculation of Main effects

These are shown in Table 12. The calculation of the figure for M goes as follows. First we find the average of the mean customer rating for the high (2) level - i.e. the average of runs 2, 4, 6, 8 (since these have 2 in the first, M, column). Then we do the same the low (1) level runs. This comes to

$$0.25(5.8+6.0+7.2+7.3) - 0.25(5.5+6.5+6.2+6.2) = 6.575 - 6.1 = 0.475$$

This means that, on average, the higher level of M gets a rating 0.475 higher than the low level. The design has the advantage that the four high level runs include two high and two low level runs for each of the other factors, which means that the comparison should be fair with regard to the other factors. This is discussed in more detail in the section on orthogonal arrays below.

Table12 - Main Effects for baking experiment

Factor	*Main effect*
M	+0.475
S	+0.325
E	+0.775
F	+0.575
O	-0.275
B	+0.225

Table 12 suggests that the higher level produces the higher average rating for all factors except O. For O the lower temperature produces the best results. The biggest effect is achieved by the Eggs factor: including an extra egg makes a bigger difference than increasing the quantities of the other ingredients.

This suggests that the best results would be achieved by setting all the factors at the higher level, except for O, which should be set at the lower level. This combination, however, is not included in the experimental runs which have been performed. (Remember this is a fractional factorial experiment, so we have not tried all possible combinations.)

Despite this it is possible to make a crude prediction of the rating that would result from this optimum combination. The closest of the actual runs is Run 8 which produced a mean rating of 7.3. This run differs just in the level of O, so, as the effect of O is -0.275, a crude prediction for the optimum combination would be 7.3 + 0.275 = 7.575.

Confirmation run

It is possible that increasing the quantities of *all* ingredients may not be as effective as our results suggest. The obvious way of finding out is to do a *confirmation run* to verify that this combination does in fact produce a mean rating close to the prediction (or at least better than any of the combinations which have been tried in the experiment).

If this confirmation run were to be done, and the mean rating was 7.8, this would support the idea that this combination is the best. If, on the other hand, the mean rating were only 6.0, this would suggest that this conclusion is wrong and that we need to perform a more detailed experiment in which we try more combinations of factor levels.

Interactions and statistical significance

It is tempting to try to get some idea of two factor interactions from the results in Table 7. *However, this is not usually a good idea.* To see why not, consider the factors S and O.

There are four runs with low levels of S, and of these two have low levels of O, and two have high levels. The mean rating from the two with low levels of O (runs 5 and 6) is 6.7, and the mean from two with high levels of O (1 and 2) is 5.65. This suggests that if the level of S is low, O has a negative effect of -1.05 (i.e. the higher rating is obtained from the lower level).

Similar arithmetic with the four runs with the higher levels of S produces a positive effect of +0.5. This means that the effect of O seems to depend on the value of S. There seems to be an *interaction* between the two factors. A diagram like Figure 3 would show two lines which are not parallel; in fact one will slope upwards and the other downwards.

However, we should be very cautious about this result. The difficulty is that what appears to be an interaction could simply be the effect of Factor E. This factor appears to have a strong positive effect because the average of the last four experimental runs is higher than the average of the first four. However, if you check through the arithmetic necessary to work out the interaction between S and O, you will find this also involves the difference between the average of the first four experimental runs and the last four experimental runs[1]. This means we cannot distinguish between the main effect of E and an interaction between S and O. (If we did want to find out about the interaction we should avoid having a Factor E, and use this column for the interaction – see, for example, Mitra, 1993, p. 531.)

Fractional experiments like this are mainly for assessing the main effects; we must be very careful drawing conclusions about interactions from them. To do this we need a full factorial experiment.

However, it is possible that interactions like this do exist, so we must be cautious about drawing conclusions from the main effects, because these are an overall average which may not be very useful if the effect varies according the values of the other variables. This is one of the reasons why the confirmation run is so important.

The other reason why we should be cautious of the result and should do a confirmation run to check it is that the amount of data is very limited, and the differences we have found are fairly small, so there is a possibility that the patterns we found in the experiment are due to chance. Perhaps if we took another sample the results would be different? The formal way of checking whether this is plausible is to use an analysis of variance to estimate the statistical significance of the results – see the section on Replication and statistical significance above. A less formal method of checking would be to do one or more confirmation runs. If the results are a matter of chance they are unlikely to recur in the confirmation runs.

[1] This is a subtle point. To understand it, you should write down the arithmetic (in the sense of numbers and + and – signs) you would do to work out the difference between the effect of O at high levels of S and the effect of O at low levels of S: you will find that the expression you get involves the difference between the first four runs and the last four runs. This, of course, is the effect of E.

Experimental design and orthogonal arrays

Unless the number of factors is very small, the use of fractional factorial designs instead of full factorial designs can save a lot of time and money. *Orthogonal arrays* are useful to plan such experiments.

An orthogonal array is a matrix of numbers arranged in columns and rows. Each column represents a specific factor or condition that can be changed from experiment to experiment. Each row represents the state of the factors in a given experimental run. So called *orthogonal arrays* have the property that the levels of the various factors are arranged in such a way that the effect of one factor can be separated from the effects of the other factors (assuming no interactions). Table 13 shows one of these orthogonal arrays: L_8 (2^7). 8 refers to the number of experimental conditions tried out, 2 is the number of factor levels, and 7 is the number of factors. A full factorial experiment would involve 2^7 or 128 experimental conditions, so this design is far more economical.

Table 13: The L_8 (2^7) orthogonal array.

Experimental run	***Factors***						
	1	*2*	*3*	*4*	*5*	*6*	*7*
1	1	1	1	1	1	1	1
2	1	1	1	2	2	2	2
3	1	2	2	1	1	2	2
4	1	2	2	2	2	1	1
5	2	1	2	1	2	1	2
6	2	1	2	2	1	2	1
7	2	2	1	1	2	2	1
8	2	2	1	2	1	1	2

The main advantage of orthogonal arrays is that they allow us to compare the effect of low and high levels of (for example) Factor 1 because the comparison is “fair” or “balanced” with respect to the other factors. For example the high level (of Factor 1) is combined with low levels of Factor 2 in rows (experiments) 5 and 6, and with high levels of Factor 2 in rows 7 and 8. Exactly the same is true of the low levels of Factor 1. This means the comparison is fair in the sense that it cannot be attributed to any of the other factors. The same is true of any other pair of factors.

It is possible to delete one or more of the columns of an orthogonal array and still have an orthogonal array. For example, if we delete the seventh column in Table 13, we get an array for investigating the effects of six factors. (The array in Table 11 is orthogonal: if you swap the 1’s and 2’s in Table 13, and delete one of the Columns you should be able to get the array in Table 11.)

Concluding comments

Most of the examples above have been from production of one sort or another. In this area, experiments can be useful for finding the best levels of various factors with a view to controlling the level of the response variable.

Another key aim is often to control the variability of the response variable. If one is manufacturing a component it is very important to ensure that the results are consistent as possible. This can be achieved by taking the standard deviation of the results (replications) of a given run as a response variable: e.g. in Tables 8 and 9. This principle is also used in Ferguson and Dale (2007: 410-414 especially the table on p. 413).

Designed experiments are also used in the social sciences, management, education, medicine and many other fields. Often these experiments involve human beings, and, typically, there are fewer factors that can be controlled and the "noise factors" are more important. (Noise factors are factors that are not controlled by the experimenter: these are the factors which mean that replications of an experiment do not always produce the same results – see Tables 8 and 9.) There may also be practical and ethical problems in manipulating human beings! Some of these issues are explored in the exercises below.

Exercises

6 A company selling children's toys on the web has done an experiment to determine the effectiveness of two layouts of website -called simple (1) and complex (2), and the presence (1) or absence (0) of music when visitors arrive. These two factors defined four versions of the website: over a period of a week visitors were randomly sent to one of these four versions. The resulting sales are given in the table.

What can you conclude about the effect of the two factors?

It would have been easier to have sent all visitors arriving on one day to one version, all visitors the next day to the second version, and so on. Would this design have been as good?

Experimental condition	Complexity	Music?	Sales
1	1	1	82
2	2	1	93
3	1	0	75
4	2	0	75

7 An experiment was designed to investigate five two level factors; A, B, C, D and E. The results are shown below.

Run	A	B	C	D	E	Response
1	1	1	1	1	1	42
2	1	1	1	2	2	50
3	1	2	2	1	1	36
4	1	2	2	2	2	45
5	2	1	2	1	2	35
6	2	1	2	2	1	55
7	2	2	1	1	2	30
8	2	2	1	2	1	54

Is this design based on an orthogonal array?

Suggest the optimum design parameters (control factors) based on the assumption that the experimenter wanted to *minimise* the response. Which variable has the largest effect?

Suppose the experimenter also wants to minimise the variation of the response variable due to noise factors. How does the experiment need to be extended to achieve this?

8 **Surveys versus controlled experiments**

A medical researcher wants to assess the effects of aspirin, diet and smoking on the incidence of heart disease. She decides that an experiment is impractical and that she will have to get her data by monitoring a large number of people over several years and getting data on how much aspirin they take each week, what sort of diet they have, how much they smoke, and any signs of heart disease. She then analyses each of the first three variables (aspirin, diet and smoking) to see if it is related to the incidence of heart disease.

What do you think of this approach to the research? What are the problems?

A second researcher working on the effects of aspirin, diet and smoking on the incidence of heart disease recognises the difficulties of survey research, and decides to do a controlled experiment over a period of five years. She chose a full factorial design with each factor at three levels:

Aspirin: 0, 60 or 300 mg per day
Diet: high fish diet, vegetarian, junk food
Smoking: 0, 10, 100 cigarettes per day

2700 volunteers were randomly assigned to each of the 27 different combinations of factor levels.

Discuss whether this experiment is possible, useful and ethical. What do you think is the best approach to this problem?

Notes on answers to exercises

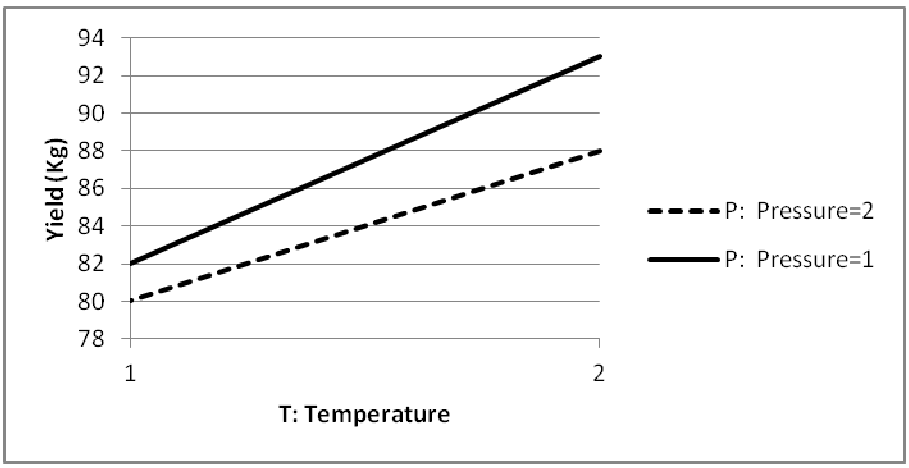
1 Average response (i.e. yield) at low level of T = (82 + 80)/2 = 81

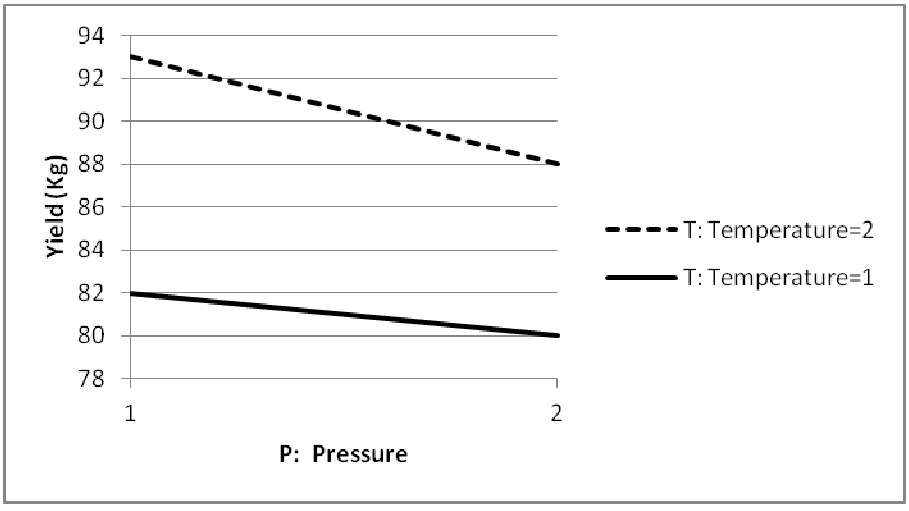
Average response at high level of T = (93 + 88)/2 = 90.5

Therefore effect of factor T = 90.5 – 81 = 9.5

Similarly, effect of factor P = 84 – 87.5 = -3.5 (negative effect implies that the average response is higher at the lower level of the factor).

There are two possible interaction graphs you could draw. It doesn't matter which you drew; the conclusions from each are the same.





As the lines on both graphs are almost parallel, there is not a large interaction between the factors. The effect of Factor F is similar at both levels of Factor M, so the main effects are useful.

2 Eight factors at 2 levels needs 2^8 or 2x2x2x2x2x2x2x2=256 experimental runs. Two factors at 4

levels requires 4^3=64 experimental runs. If in doubt about these answers, if you start to write them down systematically (like Table 6 above) you should see the pattern.

3 It may be useful to set out the data as a table:

Experimental Trial (runner / training regime)	*Duration of training runs*	*Frequency of training runs*	*Marathon time*
Sue	1 hour	3 times a week	5 hours
Sharon	1 hour	7 times a week	4 hours 30 minutes
Shirley	15 minutes	3 times a week	4 hours

There are many obvious things wrong with this experiment. Most marathon training regimes are a lot harder than this. It would have been a good idea to do some research to ensure that the training regimes in the experiment are likely to be good ones. Then a systematic experiment could be used to check the effect of factors like the distance run on training, the frequency of training, and probably other factors. It is likely that many of these factors will interact – e.g. if you are running 30 km then running every day is unlikely to be a good idea (imagine it!), but if you decided to go for 1 km runs then running every day is more likely to be a good idea.

There three other obvious problems. First this is a one-factor-at –a –time experiment. The combination of 15 minutes 7 times a week is not tested, and there is no way of finding out about the interaction between the two factors.

Second, no account is taken of any differences in ability of the three runners. A better response variable, which would take some account of this, would be the difference between the run in the marathon they are training for, and their time in the last marathon they ran (if they've run one before) So, for example, if Sue's last time was 6 hours, her *improvement* would be 1 hour. If Shirley's last time was 3 hours 30 minutes, her improvement would be negative (–30 minutes). From the point of view of this response variable, Sue did better than Shirley.

Third, and perhaps most obviously, there is only one runner trying each training regime. There are obviously big differences between individual runners, so to make a realistic comparison, we need a larger sample of runners trying each training regime. Then we can compare the averages. It is important that the runners are chosen for each training regime *at random* because if the runners were allowed to choose there may be a tendency for the better runners to choose the harder training regimes, so the results would not provide a fair comparison. These issues are dealt with in the next section on Replication and statistical significance.

4 The main effects are now 28.3 for T, and 16 for P. The lines on the graph are *not* parallel, so there is a definite interaction. The effect of one factor depends on the level of the other.

If we start from the first experimental condition, changing the level of T will increase the yield by 10. Starting again from this condition, if we change the level of P, the yield goes *down* by 2.4. This means we might expect doing both changes together would lead to an increase of 10-2.4 or 7.6. However, this is

not so: changing the levels of both factors together leads to a massive increase of 44.3. The two factors interact, so we cannot calculate their combined effects by adding the effects of each factor.

All three p-values are 0.0000 (i.e. zero to four decimal places). The reason is that the main effects and interaction are large compared with the variation between the three replications in each experiment condition (the difference between biggest and smallest is only 1 or 2 for all four experimental conditions). This means that the results would have been very, very unlikely to be the result of chance, so the evidence for the main effects and interaction is statistically significant.

5 There are many possible uses—e.g. experiments to find the best recipe for cooking something, or to find the factors that have an impact on blood pressure (e.g. exercise, salt consumption, stress level).

6 The main effect of music is +12.5 and of complexity is +5.5. Both music and complexity seem to help. However, the factors interact: complexity does *not* help in the site without music. Sending visitors to one site one day and another site the next would not have been a good idea because it would have been difficult to separate differences between days (e.g. traffic may be better on a Saturday) from the effect of the different sites. In practice this sort of experiment is fairly easy and they are widely performed. Google, for example, may send visitors to slightly different versions of their search engine to help adjust some of the parameters (factors) they use.

7 It is an orthogonal array. It is simply the first five columns of the array $L_8(2^7)$ in the Appendix.

The analysis of the main effects suggests that the best factor levels for minimising the response are:

A(1), B(2), C(2), D(1), E(2)

D has the largest effect on the response (15.25), so this is the most important factor to control. It would obviously be a good idea to do a confirmation run of this combination of levels.

To analyse any noise factors it would be necessary to measure the response variable several times with each combination of control factors. Then the standard deviation, or another measure of variation, could be used as a response variable.

8 There are two main problems with this type of survey. The first is that variables like diet are extremely complex and not easily summarised in a helpful form to analyse the statistics. The second, even more important, problem is that many variables are not *controlled*. To see why this matters, consider the aspirin analysis. The idea here would be to compare people on different doses of aspirin to see if their propensity to heart disease varies in any systematic way. However, to be useful, this comparison needs to be “fair” in the sense that the groups being compared are similar apart from their consumption of aspirin - i.e. all the other important variables need to be controlled. In a survey, this is unlikely to be so, perhaps because the group taking a lot of aspirin may be doing so because they are less healthy than average. Or, with the diet comparison, it is likely that people on supposedly healthy diets may also have other healthy habits like taking a lot of exercise. It is impossible to be sure from a

survey of this kind, and so we can never be sure that the variable we are looking at is really the one having the effect. There are always far too many other possibilities to check!

In theory, an experiment like the one suggested has the advantage that, because people are put in groups *at random*, the groups should be similar except for the variables which are controlled by the experiment. This avoids the second problem of the survey, and means that any comparisons should be fair.

In practice, of course, the experiment is not possible. It would obviously not be possible to persuade 2700 volunteers to take part, and if it was possible it would not be ethical - especially telling people to smoke for the purposes of the experiment. In addition, the idea of just three types of diet is clearly so unrealistic that the results would be of little value.

Despite this, on occasions, there is good justification for doing experiments like this with people - exampls are drug trials, and experiments to analyse the effectiveness of various aspects of a marketing campaign on the behaviour of potential customers. In such experiments the main noise factors are likely to be due to the people involved. Experiments with people tend to be rather more difficult to design and organise than industrial experiments!

In practice, research on the factors causing heart disease has to use a mix of survey and experiment. Experiments are possible for factors (like new drugs) which can be controlled, and where there are not ethical bars to one of the treatments (e.g. where there is no strong reason to believe that one treatment is better than the others). And in surveys, it may be possible to use a statistical analysis (multiple regression in particular – as discussed in another part of this course) to make allowances for some of the interfering variables.

References

Ayres, I. (2007). *Super crunchers - how anything can be predicted.* London: John Murray.

Ferguson, I., & Dale, B. G. (2007). Design of experiments. In B. G. Dale, T. van der Wiele, & J. van Iwaarden, *Managing quality (5th edition)* (pp. 402-424). Oxford: Blackwell.

Mitra, A. (1993). *Fundamentals of quality control and improvement.* New York: Macmillan.

Trochim, William M. (2006). The Research Methods Knowledge Base: Experimental Design, at http://www.socialresearchmethods.net/kb/desexper.php (version current as of October 20, 2006).

Wood, M. (2003). *Making sense of statistics: a non-mathematical approach.* Basingstoke: Palgrave Macmillan.