

1

Introduction to Design and Analysis of Experiments

1.1 Why Simultaneous Experiments?

Let us consider the problem of estimating the weight μ of a material W using four measurements by a balance. The statistical model for this experiment is written as

$$y_i = \mu + e_i, i = 1, 2, 3, 4,$$

where the e_i are uncorrelated with expectation zero (unbiasedness) and equal variance σ^2 . Then, a natural estimator

$$\hat{\mu} = \bar{y} = (y_1 + y_2 + y_3 + y_4)/4$$

is an unbiased estimator of μ with minimum variance $\sigma^2/4$ among all the linear unbiased estimators of μ . Further, if the normal distribution is assumed for the error e_i , then $\hat{\mu}$ is the minimum variance unbiased estimator of μ among all the unbiased estimators, not necessarily linear.

In contrast, when there are four unknown means $\mu_1, \mu_2, \mu_3, \mu_4$, we can estimate all the μ_i with variance $\sigma^2/4$ and unbiasedness simultaneously by the same four measurements. This is achieved by measuring the total weight and the differences among the μ_i 's according to the following design, where $\pm$ means putting the material on the right or left side of the balance:

$$\begin{aligned} y_1 &= \mu_1 + \mu_2 + \mu_3 + \mu_4 + e_1, \\ y_2 &= \mu_1 + \mu_2 - \mu_3 - \mu_4 + e_2, \\ y_3 &= \mu_1 - \mu_2 + \mu_3 - \mu_4 + e_3, \\ y_4 &= \mu_1 - \mu_2 - \mu_3 + \mu_4 + e_4. \end{aligned} \tag{1.1}$$

Advanced Analysis of Variance, First Edition. Chihiro Hirotsu.

© 2017 John Wiley & Sons, Inc. Published 2017 by John Wiley & Sons, Inc.

Then, the estimators

$$\begin{aligned}\hat{\mu}_1 &= (y_1 + y_2 + y_3 + y_4)/4, \\ \hat{\mu}_2 &= (y_1 + y_2 - y_3 - y_4)/4, \\ \hat{\mu}_3 &= (y_1 - y_2 + y_3 - y_4)/4, \\ \hat{\mu}_4 &= (y_1 - y_2 - y_3 + y_4)/4.\end{aligned}$$

are the best linear unbiased estimators (BLUE; see Section 2.1), each with variance $\sigma^2/4$. Therefore, a naïve method to replicate four measurements for each μ_i to achieve variance $\sigma^2/4$ is a considerable waste of time. More generally, when the number of measurements n is a multiple of 4, we can form the unbiased estimator of all n weights with variance σ^2/n . This is achieved by applying a Hadamard matrix for the coefficients of μ_i 's on the right-hand side of equation (1.1) (see Section 15.3 for details, as well as the definition of a Hadamard matrix).

1.2 Interaction Effects

Simultaneous experiments are not only necessary for the efficiency of the estimator, but also for detecting interaction effects. The data in Table 1.1 show the result of 16 experiments (with averages in parentheses) for improving a printing machine with an aluminum plate. The measurements are fixing time (s); the shorter, the better. The factor F is the amount of ink and G the drying temperature. The plots of averages are given in Fig. 1.1.

From Fig. 1.1, (F_2, G_1) is suggested as the best combination. On the contrary, if we compare the amount of ink first, fixing at the drying temperature 280°C (G_2), we shall erroneously choose F_1 . Then we may fix the ink level at F_1 and try to compare the drying temperature. We may reach the conclusion that (F_1, G_2) should be an optimal combination without trying the best combination, (F_2, G_1) . In this example the optimal level of ink is reversed according to the levels G_1 and G_2 of the other factor. If there is such an interaction effect between the two factors, then a one-factor-at-a-time

Table 1.1 Fixing time of special aluminum printing.

Ink supply	Temperature							
	$G_1 : 170^\circ\text{C}$				$G_2 : 280^\circ\text{C}$			
F_1 : large	5.9	3.7	4.6	4.4	5.7	5.0	4.9	2.1
		(4.65)				(4.43)		
F_2 : small	4.7	3.3	4.5	1.0	8.2	5.9	10.7	8.5
		(3.38)				(8.33)		

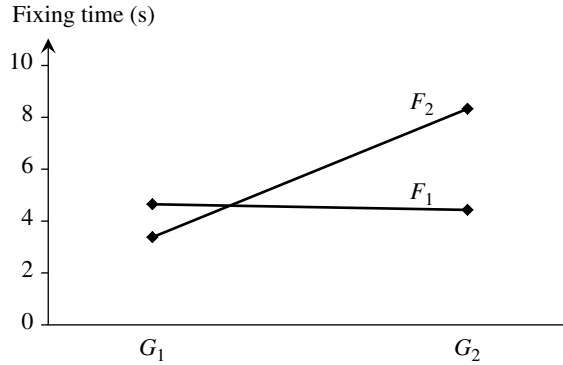
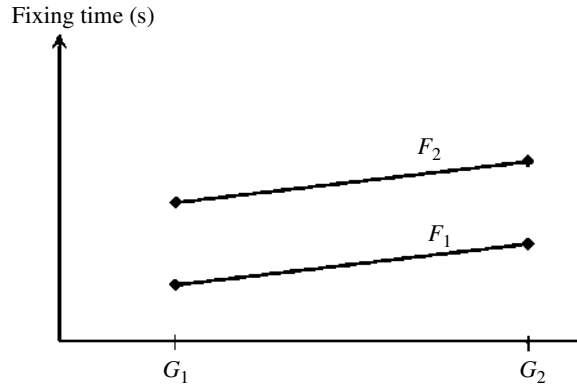
Figure 1.1 Average plots at (F_i, G_j) .

Figure 1.2 No interaction.

experiment will fail to find the optimal combination. In contrast, if there is no such interaction effect, then the effects of the two factors are called additive. In this case, denoting the mean for the combinations (F_i, G_j) by μ_{ij} , the equation

$$\mu_{ij} = \bar{\mu}_{i.} + \bar{\mu}_{.j} - \bar{\mu}_{..} \quad (1.2)$$

holds, where the dot and overbar denote the sum and average with respect to the suffix replaced by the dot throughout the book. Therefore, $\bar{\mu}_{..}$ implies the overall average (general mean), for example. If equation (1.2) holds, then the plot of the averages becomes like that in Fig. 1.2. Although in this case a one-factor-at-a-time experiment will also reach the correct decision, simultaneous experiments to detect the interaction effects are strongly recommended in the early stage of the experiment.

1.3 Choice of Factors and Their Levels

A cause affecting the target value is called a factor. Usually, there are assumed to be many affecting factors at the beginning of an experiment. To write down all those factors, a 'cause-and-effect diagram' like in Fig. 1.3 is useful. This uses the thick and thin bones of a fish to express the rough and detailed causes, arranged in order of operation. In drawing up the diagram it is necessary to collect as many opinions as possible from the various participants in the different areas. However, it is impossible to include all factors in the diagram at the very beginning of the experiment, so it is necessary to examine the past data or carry out some preliminary experiments. Further, it is essential to obtain as much information as possible on the interaction effects among those factors. For every factor employed in the experiment, several levels are set up – such as the place of origin of materials $A_1, A_2, \dots$ and the reaction temperature $170^\circ\text{C}, 280^\circ\text{C}, \dots$. The levels of the nominal variable are naturally determined by the environment of the experiment. However, choosing the levels of the quantitative factor is rather arbitrary. Therefore, sometimes sequential experiments are required first to outline the response surface roughly then design precise experiments near the suggested optimal points. In Fig. 1.1, for example, the optimal level of temperature G with respect to F_2 is unknown – either below G_1 or between G_1 and G_2 . Therefore, in the first stage of the experiment, it is desirable to design the experiment so as to obtain an outline of the response curve. The choice of factors and their levels are discussed in more detail in Cox (1958).

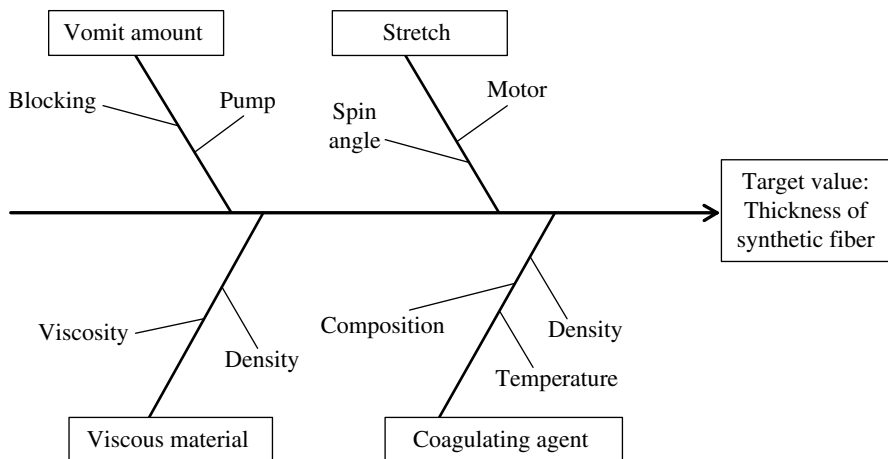


Figure 1.3 Cause-and-effect diagram.

1.4 Classification of Factors

This topic is discussed more in Japan than in other countries, and we follow here the definition of Takeuchi (1984).

(1) **Controllable factor.** The level of the controllable factor can be determined by the experimenter and is reproducible. The purpose of the experiment is often to find the optimal level of this factor.

(2) **Indicative factor.** This factor is reproducible but not controllable by the experimenter. The region in the international adaptability test of rice varieties is a typical example, while the variety is a controllable factor. In this case the region is not the purpose of the optimal choice, and the purpose is to choose an optimal variety for each region – so that an interaction analysis between the controllable and indicative factors is of major interest.

(3) **Covariate factor.** This factor is reproducible but impossible to define before the experiment. It is known only after the experiment, and used to enhance the precision of the estimate of the main effects by adjusting its effect. The covariate in the analysis of covariance is a typical example.

(4) **Variation (noise) factor.** This factor is reproducible and possible to specify only in laboratory experiments. In the real world it is not reproducible and acts as if it were noise. In the real world it is quite common for users to not follow exactly the specifications of the producer. For example, a drug for an infectious disease may be used before identifying the causal germ intended by the producer, or administered to a subject with some kidney difficulty who has been excluded in the trial. Such a factor is called a noise factor in the Taguchi method.

(5) **Block factor.** This factor is not reproducible but can be introduced to eliminate the systematic error in fertility of land or temperature change with passage of time, for example.

(6) **Response factor.** This factor appears typically as a categorical response to a contingency table and there are two important cases: nominal and ordinal. The response is usually not called a factor, but mathematically it can be regarded and dealt with as a factor, with categories just like levels.

One should also refer to Cox (1958) for a classification of the factors from another viewpoint.

1.5 Fixed or Random Effects Model?

Among the factors introduced in Section 1.4, the controllable, indicative and covariate factors are regarded as fixed effects. The variation factor is dealt with as fixed in the laboratory but dealt with as random in extending laboratory results to the real world.

Therefore, the levels specified in the laboratory should be wide enough to cover the wide range of real applications. The block is premised to have no interaction with other factors, so that the treatment either as fixed or random does not affect the result. However, it is necessarily random in the recovery of inter-block information in the incomplete block design (see Section 9.2).

The definition of fixed and random effects models was first introduced by Eisenhart (1947), but there is also the comment that these are mathematically equivalent and the definitions are rather misleading. Although it is a little controversial, the distinction of fixed and random still seems to be useful for the interpretation and application of experimental results, and is discussed in detail in Chapters 12 and 13.

1.6 Fisher's Three Principles of Experiments vs. Noise Factor

To compare the treatments in experiments, Fisher (1960) introduced three principles: (1) randomization, (2) replication and (3) local control.

To explain randomization, Fisher introduced the sensory test of tasting a cup of tea made with milk. The problem then is to know whether it is true or not that a lady can declare correctly whether the milk or the tea infusion was added to the cup first. The experiment consists of mixing eight cups of tea, four in one way and four in the other, and presenting them to the subject for judgment. There are, however, numerous uncontrollable causes which may influence the result: the requirement that all the cups are exactly alike is impossible; the strength of the tea infusion may change between pouring the first and last cup; and the temperature at which the tea is tasted will change in the course of the experiment. One procedure that is used to escape from such systematic noise is to randomize the order of the eight cups for tasting. This process converts the systematic noise to random error, giving the basis of statistical inference.

Secondly, it is necessary to replicate the experiments to raise the sensitivity of comparison. It is also necessary to separate and evaluate the noise from treatment effects, since the outcomes of experiments under the same experimental conditions can vary due to unknown noise. The treatment effects of interest should be beyond such random fluctuations, and to ensure this several replications of experiments are necessary to evaluate the effects of noise.

Local control is a technique to ensure homogeneity within a small area for comparing treatments by splitting the total area with large deviations of noise. In field experiments for comparing a plant varieties, the whole area is partitioned into n blocks so that the fertility becomes homogeneous within each block. Then, the precision of comparisons is improved compared with randomized experiments of all an treatments.

Fisher's idea to enhance the precision of comparisons is useful in laboratory experiments in the first stage of research development. However, in a clinical trial for comparing antibiotics, for example, too rigid a definition of the target population and the

causal germs may not coincide with real clinical treatment. This is because, in the real world, antibiotics may be used by patients with some kidney trouble who might be excluded from the trial, by older patients beyond the range of the trial, before identifying the causal germ exactly, or with poor compliance of the taking interval. Therefore, in the final stage of research development it is required to introduce purposely variations in users and environments in the experiments to achieve a robust product in the real world. It should be noted here that the purpose of experiments is not to know all about the sample, but to know all about the background population from which the sample is taken – so the experiment should be designed to simulate or represent well the target population.

1.7 Generalized Interaction

A central topic of data science is the analysis of interaction in a generalized sense. In a narrow sense, it is the departure from the additive effects of two factors. If the effect of one factor differs according to the levels of the other factor, then the departure becomes large (as in the example of Section 1.2).

In the one-way layout also, the main effects of a treatment become the interaction between the treatment and the response if the response is given by a categorical response instead of quantitative measurements. In this case, the data y_{ij} are the frequency of the (i, j) cell for the i th treatment and the j th categorical response. If we denote the probability of cell (i, j) by p_{ij} , then the treatment effect is a change in the profile $(p_{i1}, p_{i2}, \dots, p_{ib})$ of the i th treatment and the interaction effects in terms of p_{ij} are concerned. In this case, however, a naïve additive model like (1.2) is often inappropriate, and the log linear model

$$\log p_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

is assumed. Then, the factor $(\alpha\beta)_{ij}$ denotes the i th treatment effect. In this sense, the regression analysis is also a sort of interaction analysis between the explanation and the response variables. Further, the logit model, probit model, independence test of a contingency table, and canonical correlation analysis are all regarded as a sort of interaction analysis. One should also refer to Section 7.1 regarding this idea.

1.8 Immanent Problems in the Analysis of Interaction Effects

In spite of its importance, the analysis of interaction is paid much less attention than it deserves, and often in textbooks only an overall F - or χ^2 -test is described. However, the degrees of freedom for interaction are usually large, and such an overall test cannot tell any detail of the data – even if the test result is highly significant. The degrees of freedom are explained in detail in Section 2.5.5. In contrast, the multiple comparison procedure

based on one degree of freedom statistics is far less powerful and the interpretation of the result is usually unclear. Usually in the text books it is recommended to estimate the combination effect μ_{ij} by the cell mean $\bar{y}_{ij.}$, if the interaction exists. However, it often occurs that only a few degrees of freedom can explain the interaction very well, and in this case we can recover information for μ_{ij} from other cells and improve the naïve estimate $\bar{y}_{ij.}$ of μ_{ij} . This also implies that it is possible to separate the essential interaction from the noisy part without replicated experiments. Further, the purpose of the interaction analysis has many aspects – although the textbooks usually only describe how to find an optimal combination of the controllable factors. In this regard the classification of factors plays an essential role (see Chapters 10, 11, 13, and 14).

1.9 Classification of Factors in the Analysis of Interaction Effects

In case of a two-factor experiment, one factor should be controllable since otherwise the experiment cannot result in any action. In case of controllable vs. controllable, the purpose of the experiment will be to specify the optimal combination of the levels of those two factors for the best productivity. Most of the textbooks describe this situation. However, the usual F test is not useful in practice, and the simple interaction model derived from the multiple comparison approach would be more useful.

In case of controllable vs. indicative, the indicative factor is not the object of optimization but the purpose is to specify the optimal level of the controllable factor for each level of the indicative factor. In the international adaptability test of rice varieties, for example, the purpose is obviously not to select an overall best combination but to specify an optimal variety (controllable) for each region (indicative). Then, it should be inconvenient to hold an optimal variety for each of a lot of regions in the world, and the multiple comparison procedure for grouping regions with similar response profiles is required.

The case of controllable vs. variation is most controversial. If the purpose is to maximize the characteristic value, then the interaction is a sort of noise in extending the laboratory result to the real world, where the variation factor cannot be specified rigidly and may take diverse levels. Therefore, it is necessary to search for a robust level of the controllable factor to give a large and stable output beyond the random fluctuations of the variation factor. Testing main effects by interaction effects in the mixed effects model of controllable vs. variation factors is one method in this line (see Section 12.3.5).

1.10 Pseudo Interaction Effects (Simpson's Paradox) in Categorical Data

In case of categorical responses, the data are presented as the number of subjects satisfying a specified attribute. Binary (1, 0) data with or without the specified attribute are a typical example. In such cases it is controversial how to define the interaction

Table 1.2 Simpson's paradox.

	Young ($j = 1$)		Old ($j = 2$)		Young + old	
	($k = 1$)	($k = 2$)	($k = 1$)	($k = 2$)	($k = 1$)	($k = 2$)
Drug ($i = 1$)	120	40	10	30	130	70
Drug ($i = 2$)	30	10	40	120	70	130

effects, see Darroch (1974). In most cases an additive model is inappropriate, and is replaced by a multiplicative model. The numerical example in Table 1.2 will explain well how the additive model is inappropriate, where ($k = 1$) denotes useful and ($k = 2$) useless. In Table 1.2 it is obvious that drug 1 and drug 2 are equivalent in usefulness for each of the young and old patients, respectively. Therefore, it seems that the two drugs should be equivalent for (young + old) patients. However, the collapsed sub-table for all the subjects apparently suggests that drug 1 is better than drug 2. This contradiction is known as Simpson's paradox (Simpson, 1951), and occurs by additive operation according to the additive model of the drug and age effects. The correct interpretation of the data is that both drugs are equally useful for young patients and equally useless for old patients. Drug 1 is employed more frequently for young patients (where the useful cases are easily obtained) than old patients, and as a result the useful cases are seen more in drug 1 than drug 2. By applying the multiplicative model we can escape from this erroneous conclusion (Fienberg, 1980) – see Section 14.3.2 (1).

1.11 Upper Bias by Statistical Optimization

As a simple example, suppose we have random samples $y_{11}, \dots, y_{1n}$ and $y_{21}, \dots, y_{2n}$ from the normal population $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$, respectively, where $\mu_1 = \mu_2 = \mu$. Then, if we select the population corresponding to the maximum of $\bar{y}_1$ and $\bar{y}_2$, and estimate the population mean by the maximal sample, an easy calculation leads to

$$E\{\max(\bar{y}_1, \bar{y}_2)\} = \mu + \sigma/\sqrt{n\pi}$$

showing the upper bias as an estimate of the population mean μ . The bias is induced by treating the sample employed for selection (optimization) as if it were a random sample for estimation; this is called selection bias.

A similar problem inevitably occurs in variable selection in the linear regression model, see Copas (1983), for example. It should be noted here again that the purpose of the data analysis is not to explain well the current data, but to predict what will happen in the future based on the current data. The estimation based on the data employed for optimization is too optimistic to predict the future. Thus, the Akaike's

information criterion (AIC) approach or penalized likelihood is justified. One should also refer to Efron and Tibshirani (1993) for the bootstrap as a non-parametric method of model validation.

1.12 Stage of Experiments: Exploratory, Explanatory or Confirmatory?

Finally, most important in designing experiments is to define the target of the experiments clearly. For this purpose it is useful to define the three stages of experiments. The first stage is exploratory, whose purpose is to discover a promising hypothesis in the actual science – such as industry and clinical medicine. At this stage the exploring data analysis, analysis of variance, regression analysis, and many other statistical methods are applied. Data dredging is allowed to some extent, but it is most inappropriate to take the result as a conclusion. This stage only proposes some interesting hypotheses, which should be confirmed in the following stages. The second stage is explanatory, whose purpose is to clarify the hypothesis under rigid experimental conditions. The design and analysis of experiments following Fisher's principle will be successfully applied. The third stage is confirmatory, whose purpose is to confirm that the result of laboratory experiments is robust enough in the actual world. The robust design of Taguchi is useful here. It should be noted that in these stages of experiments, the essence of the statistical method for summarizing and analyzing data does not change; the change is in the interpretation and degree of confidence of the analytical results. Finally, follow-up analysis of the post-market data is inevitable, since it is impossible to predict all that will happen in the future by pre-market research, even if the most precise and detailed experiments were performed.

References

- Copas, J. B. (1983) Regression, prediction, shrinkage. *J. Roy. Statist. Soc.* **B45**, 311–354.
- Cox, D. R. (1958) *Planning of experiments*. Wiley, New York.
- Darroch, J. N. (1974) Multiplicative and additive interaction in contingency tables. *Biometrika* **61**, 207–214.
- Efron, B. and Tibshirani, R. (1993) *An introduction to bootstrap*. Chapman & Hall, New York.
- Eisenhart, C. (1947) The assumptions underlying the analysis of variance. *Biometrics* **3**, 1–21.
- Fienberg, S. E. (1980) *The analysis of cross classified data*. MIT Press, Boston, MA.
- Fisher, R. A. (1960) *The design of experiments*, 7th edn. Oliver & Boyd, Edinburgh.
- Simpson, E. H. (1951) The interpretation of interaction in contingency tables. *J. Roy. Statist. Soc.* **B13**, 238–241.
- Takeuchi, K. (1984) Classification of factors and their analysis in the factorial experiments. *Kyoto University Research Information Repository* **526**, 1–12 (in Japanese).