1

Introduction to designs

1.1 Introduction

There are many statistical issues to consider in the design of an empirical study. Among these problems are the control of unwanted variation and the internal validity of a study. How can we be sure that a study is internally valid? In other words, how can we be sure that the treatment effect is attributed to the variables that are manipulated and not mainly influenced by unwanted variation? These questions and related problems are well documented and have been discussed in the social and biomedical literature by Cox (1958), Campbell and Stanley (1963), Cook and Campbell (1979), Cox and Reid (2000) and Shadish, Cook and Campbell (2002), among others. However, one important aspect that seems lacking in the discussion is the question whether the implemented design is the most efficient one for the objective or objectives of the study. We believe that researchers should always keep this question in mind when they design their studies. Doing so can invariably result in improved designs with higher statistical efficiency at minimal cost. Here and throughout, we use the term *statistical efficiency* or simply *efficiency* to indicate the accuracy of estimators of the model parameters in terms of the variances of the estimators.

Here is a simple illustration. Consider the 1987 Report of the Second Task Force on Blood Pressure Control in Children in the journal *Pediatrics* on the linear relation between systolic blood pressure and age. This conclusion was drawn from studies with samples of children with ages ranging from 12 to 18 years. Despite this informative finding, many subsequent studies continue to use inefficient designs to study the relationship between the two variables. Frequently, the design ignores existing information from the literature or simply comprises

An Introduction to Optimal Designs for Social and Biomedical Research M. P. F. Berger & W. K. Wong

equal number of children from different age groups equally distributed over the age range. For instance, the design may sample equal number of children at 12, 14, 16 and 18 years old (*Design 1*), or the design may have equal number of children at 12, 15 and 18 years old (*Design 2*). These are called *uniform designs* because they have an equal number of observations from each age group that are equally spaced over the age of interest. Uniform designs are popular because they are intuitive and simple to implement. However, these designs can be rather inefficient in part because (i) they do not incorporate existing information on the two variables in the study and (ii) no rationale is provided for in the choice of the age groups included in the study. For instance, why were the age groups of 12, 15 and 18 selected in the second uniform design? Would another design with age groups of 13, 15 and 17 be equally acceptable?

These two questions often lead to a host of other design questions, for example: (i) Is *Design 1* better than *Design 2*? (ii) Is it a good idea to sample only two age groups in the study? (iii) What about sampling cost–can I additionally include cost considerations? (iv) What happens if sampling cost is proportional to the age of the children? Children older than 18 years are excluded from the sample, but I want to make inference on the relationship between blood pressure and children older than 18 as well. Is *Design 1* better than *Design 2* for this purpose? Can I have a design that is good for studying the relationship between systolic blood pressure and age of children up to 18 years old and also for adults as well? Is there a 'best' design for the study? These are complicated questions and sometimes there are no clear answers for a practical design problem that can have many more constraints and objectives.

Optimal design theory offers a useful foundation answering these and other design questions. The beautiful framework allows us to find the best design for a given problem using computer algorithms. For simpler problems, the theory also enables us to determine the design analytically. This is useful because with a formula for describing the design, we can study properties of the design a lot more easily. Although optimal design theory has been available for many years, social, behavioural and biomedical scientists seem to have little exposure to its theory and potential applications in their fields. The aim of this book is to promote interest among researchers from these fields in the use of optimal design theory and enable them to design more efficient and less expensive studies.

Since the work of Fisher (1925, 1935), statisticians have worked on optimal design problems. A small sample of useful references over the years are Cox (1958), Kiefer (1959), Kiefer and Wolfowitz (1959), Fedorov (1972), Box, Hunter and Hunter (1978), Silvey (1980) and Atkinson and Donev (1992). There was a surge in research in this area in the early 1960s after the seminal papers by Kiefer and Wolfowitz (1959, 1960) showing that a design can be simultaneously optimal under two very different and useful objectives in the study. The monograph by Kiefer (1985) is a voluminous collection of pioneering work in this area by the author and provides a good account of the chronological developments of optimal design theory. However, despite the ubiquity of design issues in all studies, the number of statisticians working actively in this important area has always

been relatively small and continue to be so, compared to the larger statistical community working on data analysis issues. We believe there are several reasons for this.

Firstly, there has always been a strong focus on analytical solutions for optimal design problems, and the books published on optimal design theory require readers to have a good mathematical training to appreciate the theory. Although many social, behavioural and biomedical scientists receive training in statistical methods, this training does not include basic concepts of optimal design theory. Still another reason is that many statistics/biostatistics programmes at universities worldwide tend to give short shrift to design issues in their course curriculum. Another reason that optimal design ideas are underutilized in research is that its theory is generally perceived as too limiting. This is incorrect. For example, a popular criticism is that optimal design strategies are necessarily myopic. A common cited example is the optimal design for the homoscedastic simple linear model that requires an equal number of observations to be placed only at the two extreme ends of the design interval. This design cannot detect curvature and so why put all the eggs in one basket? Our response is that modern development in optimal design theory enables the researcher to construct more flexible designs that can capture the constraints and goals of the studies more realistically. In this particular example, one can construct an optimal design that balances the dual goals of estimating the two parameters in the linear model and at the same time estimates the curvilinear parameter with a user-specified level of efficiency. We will revisit this issue in Chapter 9.

However, there is increasing interest and realization that the theory can be more broadly applied to practical problems. Atkinson (1996) gave a compelling account of the usefulness of optimal designs and their potential applications to other fields. Some examples of use of optimal designs in social and behavioural research are McClelland (1997), Raudenbush (1997), Moerbeek, Van Breukelen and Berger (2000) and Berger (2005). The monograph by Berger and Wong (2005) is an attempt to show interesting applications of optimal design in different disciplines.

The use of optimal or highly efficient designs in social, behavioural and biomedical research has advantages. Fewer observations and therefore smaller sample sizes are required to find real effects, thus reducing the costs of the study. Our examples in this book show that optimal designs can reduce the number of observations from 20% to 40%, and even 50% in some cases when compared with the traditional or commonly used designs. This is especially beneficial in light of the ever-rising cost of conducting scientific studies. From an ethical viewpoint, a smaller sample is also highly desirable. For example, fewer patients may be required to undergo a controversial treatment or fewer animals need to be sacrificed in a toxicology study.

In this chapter, we review basic methodological concepts in the design of a study. We describe design terminology, types of different designs, requirements for a ‘good’ design and different kinds of validity issues that may arise in a study and how to control them.

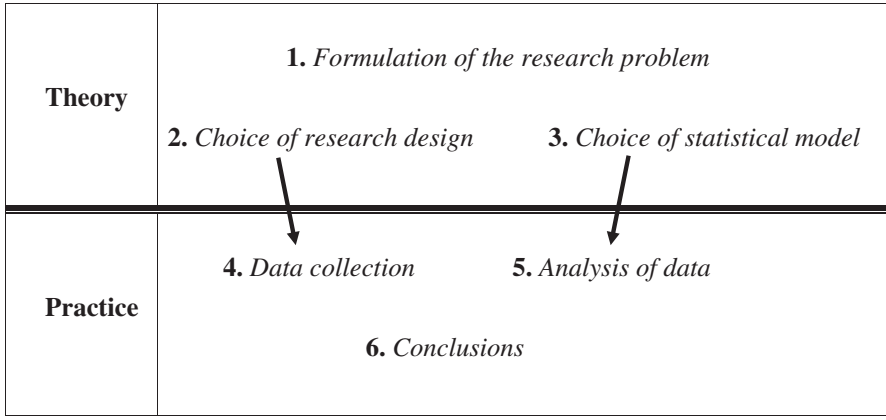
1.2 Stages of the research process

In social and biomedical research, there are six stages in the research process:

1. *Formulation of the research problem.* Many research problems may be formulated as a relationship between a set of variables X and an outcome variable Y. The main task is to identify the set of variables X to include in the study and identify the outcome variable Y of interest. For example, a psychologist wants to study the effect of a systematic desensitization therapy (X) on phobic reactions (Y) of patients or a health scientist wants to study the relation between sources of job-related stress (X) and burnout rate (Y) of professionals in medicine.
2. *Choice of the research design.* A research design is used to structure the research and data collection. A design choice not only includes the selection of the number of independent variables, a distinction between qualitative versus quantitative variables, random versus fixed variables and a crossed or nested relation among variables but also the selection of the number of measurements, time points and subjects within groups. At this stage, we decide on the sources and amount of unwanted variation to control for in the design.
3. *Choice of statistical model.* For design purposes, the statistical model must be chosen before the data are collected. The model is a mathematical relationship posited in Stage 1 and describes the outcome variable Y as a function of the variables X and the error term.
4. *Data collection.* In this stage, the data are collected based on the design chosen in Stage 2.
5. *Analysis of data.* The data are analysed based on the statistical model chosen in Stage 3. Regression diagnostics are used to check model assumptions and whether the model provides an adequate fit.
6. *Conclusions.* Conclusions are carefully inferred from the data analysis in Stage 5 and they may lead to reformulation of the research problem or to additional research problems.

These six stages are visualized in Table 1.1. The bold horizontal line in Table 1.1 separates theoretical considerations from practical issues. Since the choice of a design and the choice of a statistical model are very closely related, Stage 2 and Stage 3 may sometimes be interchanged in practice. In Stage 4, data are collected according to the design chosen in Stage 2 and in Stage 5 the data are analysed based on the statistical model selected in Stage 3. The arrows in Table 1.1 represent the connection between the activities in Stages 2 and 4, and between the activities in Stages 3 and 5, respectively. The problem is that in order to be able to choose a 'good' design, information about the true model or the data generating process is needed and this information is not available in Stage 2.

Table 1.1 Main stages of the research process.



1.2.1 Choice of a 'good' design

This book focuses on the selection of a research design (Stage 2 of Table 1.1). The typical aim in social or biomedical research is to evaluate the effects of one or more independent (treatment) variables on an outcome variable. For example, we may want to design a study to ascertain whether a new teaching method is more effective than the current method in raising the average score in mathematics among high school students in the Los Angeles County. As such, it is important to have a design capable of estimating the treatment effect precisely and also has maximal power for testing the hypothesis of interest. In practice, an efficient ('good') design is chosen under a set of constraints that usually adds further complications to the design problem. These restrictions can be cost constraints, feasibility constraints and ethical constraints. An example of a feasibility constraint could be a study aimed at controlling the increasing obesity rates among children in the United States. A new method may be potentially effective in getting children to lose weight, but if the method requires children to be segregated in classrooms, this may be problematic to implement. Parents are likely to object and such a design may not be feasible.

Optimal design theory offers a systematic way for finding an optimal or a highly efficient design using all current information for the problem at hand. Once the statistical model is specified and the objective or objectives of the study are clearly stated, along with the constraints, there are proven optimization methods that will generate the optimal design. In many situations, these numerical procedures can be easily automated. This means that there are computer programs that can generate tailor-made optimal designs sequentially. Many of these algorithms are guaranteed to converge to the optimal design; more recent development includes applying these algorithms to find multiple-objective optimal designs and include cost constraints as well. We postpone discussion of these algorithms to Chapter 11.

In the next subsections, we describe the basic elements of research designs, different types of designs, a brief overview of the methodological and ethical requirements of a 'good' design and give references in the literature.

1.3 Research design

The research design describes how data are to be collected to test whether the posited relations among variables hold or not. It is a plan for collecting and utilizing data so that information is generated to test hypotheses. A good or poor research design may be characterized by the amount of information it generates and its power for testing the hypotheses. Specifically, the research design determines conditions under which the study is to be carried out. Conditions here refer to selection of the combination of levels of all the *independent variables*, including how the *units of analysis* are allocated to each of the conditions and how many *replications* are planned. The choice of the design is based on a hypothetical relation between *independent* (*predictor, explanatory*) *variables* and *dependent* (*response*) *variables* posited in Stage 1 where it is also assumed that variation in the independent variables leads to changes in the dependent variable (effect).

1.3.1 Choice of independent variables and levels

A researcher must address the questions of how many independent variables to use and how many values (levels) of these independent variables are needed to investigate the relation between the independent variables and the dependent variable. If a whole range of levels of an independent variable is of interest and the study only includes a few of these levels, then it is best to randomly select these few levels from that range. Such independent variables are classified as *random factors*. On the other hand, the so-called *fixed factors* are independent variables with a fixed number of levels, such that inferences are only limited to these fixed levels. For example, the factor Age can be either fixed or random. Age is a fixed factor if persons from a fixed set of age groups are included in the study and the researcher is only interested in drawing conclusions for these fixed age groups. The factor Age is random if the age groups are randomly drawn from the complete range of all ages in the population so that conclusions apply to the whole range of ages. Fixed and random factors must be clearly identified in the statistical model because they affect our tests concerning model parameters. The actual number of selected levels of a factor also influences the quality of our statistical inferences. For example, we show in the next few chapters that the variances of the estimated parameters and the power for finding real effects depend on the number of selected levels of the independent variables.

1.3.2 Units of analysis

The unit of analysis is the entity that is being analysed. *Units of analysis* are often referred to as *subjects*. Examples of the units of analysis are objects,

organizations, general practices, patients, workers, pupils, and nurses. These units may be nested hierarchically within other units. For instance, a general practice has its own sample of patients. This sample of patients is nested within this general practice and not in another general practice. In the same way, nurses are nested within hospitals where they work, and pupils are nested within their own schools. Such a nested structure will lead to specific designs and the analysis of the data will have to take the hierarchical data structure into account.

1.3.3 Variables

A variable is defined as an attribute, property or characteristic of a unit. Examples are the colour of objects, size of organizations or general practices, blood pressure of patients, income of workers, ability of pupils, and workload of nurses. Different classifications of variables are distinguished in the scientific literature. To select both a research design and the proper analysis technique, the types of variables involved must be known. Variables are considered to be *qualitative* or *quantitative*.

A *qualitative variable* is a variable in which units are grouped into a set of mutually exclusive categories. Examples are gender, race, treatments, types of cancer, and departments in a hospital. Qualitative variables are also referred to as *grouping* or *categorical* variables. The differences are expressed in different groups or categories, for example, male and female or colon cancer and lung cancer. Stevens (1951) referred to these variables as *nominal* variables.

A *quantitative variable* is one in which units differ in degree. A quantitative variable is also referred to as a *numerical variable* and its measurement may be continuous or discrete. A measurement of such a quantitative variable may be on an *interval scale* or on a *ratio scale* (Stevens, 1951). An example of an interval-scaled variable is a patient's temperature. Interval variables are invariant under linear transformations. For example, temperature expressed in degrees Celsius may be transformed to the Fahrenheit scale without loss of information. Ratio-scaled variables have an absolute zero point that is meaningful. Examples of ratio-scaled variables are length, weight, response time, and blood pressure of patients.

Finally, an *ordinal-scaled variable* (Stevens, 1951), which merely assigns numbers to units to reflect the ordering of these units on a variable, cannot easily be classified as quantitative or qualitative. It is quantitative in the sense that the direction of any association involving the ordinal variable can be expressed. On the other hand it may be considered to be qualitative, because the intervals of the scaled measurement merely represent an ordering.

A quantitative or qualitative variable is usually symbolized by a capital letter X or Y. The corresponding realization of the measurement is symbolized by a lower case letter x or y, respectively. The direction of a hypothesized relation usually coincides with the distinction between the *independent* (*predictor*) *variable* and *dependent* (*outcome*) *variable*. If, for example, one is interested in the effect of dosages of a drug on the systolic blood pressure of patients, then the

direction of the relation can be visualized as

$$X \text{ (dosage level)} \rightarrow Y \text{ (bloodpressure)}.$$

In practice an independent variable X can be measured differently in different studies. Dosages may be either represented by a rank ordering, that is, $x = 1, 2, 3, 4$, and so on, or by the actual values of each dosage level, for example, $x = 6, 9, 15, 20$ mg/kg, and so on. In the first case, the independent variable dosage level X is measured on an ordinal scale, while in the latter case dosage level X is measured on a ratio scale.

1.3.4 Replication

The concept of *replication* is important in every study, because replications are needed to estimate the variance of the errors in a statistical model. Typically, replications are assumed to be independent; if they are not, the error variance will be underestimated and a Type I error results. We recall Type I error is the probability of rejecting the null hypothesis when it is true. In micro-array experiments, a further distinction between technical (within samples) replications and biological (between samples) replications is often made.

Replications should be distinguished from *repeated measurements*. Repeated measurements can also be obtained under the same conditions (same combinations of factor levels), but are not independent and often display a trend over a period of time. Such repeated measurements should be treated as a separate factor in the design. An example of a set of repeated measurements is the blood pressure readings at successive occasions from patients taking the cholesterol-lowering medication Zocor.

1.4 Types of research designs

Different research designs are used in social and biomedical research. They range from purely experimental designs to observational types of designs.

An enormous amount of literature is available for conducting *experimental designs* in a controlled environment. Examples are factorial designs, cross-over designs, blocked and Latin square designs. These designs are referred to as *experimental designs* because the researcher has full control over the design of the study. Full control means that the experimenter can manipulate the treatment (or independent) variables and the experimental units are randomly assigned to the conditions. Further details are in Winer, Brown and Michels (1991), Kirk (1995) and Montgomery (2000), among others.

In *observational designs* the researcher does not have full control over the variation of the variables. Examples of such designs without complete control in biomedical and epidemiologic research are (prospective) cohort designs, cross-sectional designs and case–control designs, see Rothman and Greenland (1998) for details. Examples of observational studies in social

research are quasi-experimental designs, survey designs, and non-experimental designs. See Cook and Campbell (1979), Campbell and Stanley (1963) and Shadish, Cook, and Campbell (2002) for details on the problems connected with quasi-experimental designs. Quasi-experimental studies may contain a manipulation of the treatment variable, but they are carried out in existing environments with existing groups of subjects. In general, observational designs are more exposed to threats to internal validity mentioned in Section 1.5.2.

One could infer that observational designs have less control over all possible sources of unwanted variation than experimental designs and as such maintain less internal validity. On the other hand, observational studies are often chosen because the study cannot be easily carried out in a controlled environment. Concato, Shah and Horwitz (2000) and Benson and Hartz (2000) conducted meta-analyses to compare results from observational and experimental studies and concluded that well-designed observational studies may produce results that are more or less comparable to those obtained in purely experimental studies. However, it should always be kept in mind that observational studies are not experimental and therefore cannot guarantee that treatment effects are unbiased. Carefully designed randomized controlled studies are always preferable (Pocock and Elbourne, 2000).

1.5 Requirements for a 'good' design

The objectives of a 'good' design are to provide *interpretable*, *accurate* and *valid* conclusions. A number of methodological requirements for a good design have been described in the literature on the design of experiments and other studies, see, for example, Cox (1958, Section 1.2), Box and Draper (1987, Chapter 14) and Atkinson and Donev (1992, Chapter 6), among others. Although these requirements were originally only formulated for experimental designs, they also apply to other types of studies with less control over the independent variables.

A 'good' design should provide valid and reliable statistical inference. Campbell and Stanley (1963) were probably the first to distinguish internal validity from external validity and describe possible threats to both forms of validity. Cook and Campbell (1979) expanded this classification into four distinct types of validity, namely, statistical conclusion validity, construct validity, internal and external validity:

- *Statistical conclusion validity* is concerned with the validity of the conclusions based on the statistical methods employed. If the statistical analysis is inappropriate, the conclusions may not be valid.
- *Construct validity* is the correspondence between the measure and the construct that is being measured. Construct validity is usually investigated by empirical testing of hypothesized relations.
- *Internal validity* is concerned with the question whether the effect found in a study can be attributed to the variables that are manipulated.

- *External validity* is the generalizability of the conclusions. Can the conclusions of current study be generalized to other populations?

A 'good' design at the very minimum should provide control over the threats to *statistical conclusion validity* and *internal validity*. We focus on these two types of validity in the following subsection. For a detailed explanation of construct validity and external validity, we refer the readers to Cook and Campbell (1979).

1.5.1 Statistical conclusion validity

There are a number of threats to statistical conclusion validity and they can take on several forms. A good design should guard against these threats. Below is a short list of these threats and some advice on how one can minimize these threats.

1.5.1.1 Lack of statistical power for finding real effects

A 'good' design should have sufficient power for finding real effects. Low power may lead to the erroneous acceptance of the null hypothesis.

This problem may arise because of inappropriate choice of design, and parameters are poorly estimated. A carefully chosen design can improve the parameter estimates and power of the tests of interest. One way to minimize such a threat is to make sure that the ranges of the independent variables are properly selected. In practice, this means that each of the independent variables in the study should have large enough variance to fully capture the outcome variation as the value of the independent variable varies. For example, consider the outcome as the reaction time between application of the stimulus and a response to that stimulus. The reaction time of subjects decreases as young subjects grow older but then increases as older subjects become older. Consequently, if one limits the age range by selecting only young subjects in a study, one will not have enough information to know and test for the curvilinear relation between age and reaction time.

Another way to minimize the lack of power threat is to have a large enough sample size in the study. The cost of a study is always an important factor and so the researcher always has to find a balance between reigning the costs of the study and having enough power to detect treatment effects. Optimal design theory offers methods that can find optimal designs with sufficient power and efficiency at minimum cost.

A way to find out that there is lack of power for finding real effects is when we observe that the standard errors of the estimates are too large. One possible cause may be because of inaccurately measured variables. The error variance of the statistical model becomes large when the independent and dependent variables have large measurement errors. This may happen, for example, when the environment of the study is unstable or when the selected subjects have characteristics that cause the dependent variable to have relatively large variation. Such unreliable measurements can result in a high probability of incorrectly failing to reject the null hypothesis and loss of power for finding real effects.

Additional statistical issues on power and sample size calculation are discussed with examples in basic statistics monographs, such as Pocock (1983) and Chow, Shao and Wang (2007).

1.5.1.2 Violation of model assumptions

To prevent incorrect inferences, a design should allow the researcher to check model assumptions and to check for goodness of fit of the model. One way to infer that the standard errors are small is to check for model fit. This can be done before the actual study is performed in a so-called pilot study or this can be done by means of the data in the study itself.

A 'good' design should generate data that allow the researcher to perform a lack of fit test on the postulated model and enable the researcher to check specific aspects of the model. We emphasize that a design may be efficient for one statistical model but not for another. Consequently, it is important to keep in mind that goodness of a design assumes an underlying model. For example, if the relation between a predictor and a dependent variable can be adequately described by a linear model, then only two distinct measurement points for data collection are sufficient to estimate the model parameters. If, however, it is not certain whether the relation is linear, and one suspects it to be curvilinear, then more than two distinct measurement points are needed.

1.5.1.3 The model is unnecessarily complicated

In any study, the researcher should first make a concerted effort to decide on the appropriate level of complexity he or she is willing to entertain in the model and the design. The key is that the model should not be unnecessarily complicated and the design should be as simple as possible to meet the study objectives. In other words, the design should aim at striking a right balance between simplicity on one hand and statistical efficiency and practicality on the other. If a design is too simple, this may lead to invalid or weak conclusions, while a highly efficient design may disguise invalid conclusions. For example, consider a Latin square design with blocking and matching variables and controls for order effects at the same time. If the statistical model for the data analysis includes different covariates as well, then the interpretation of the results can become difficult. On the other hand, if we simplify the design by not controlling the effects of extraneous variables, we may end up with false conclusions.

1.5.1.4 Conclusions based on invalid extrapolation

The range of conditions under investigation is one of the requirements for a design to be externally valid. It is desirable that the study be undertaken in a broad range of conditions to increase the chance that the conclusions from the current study are valid for other ranges of values of the independent variable. A random effect design, where the levels of the independent variable (factor) are drawn randomly from a population of levels, has the advantage that the conclusions are

generalizable to other population levels of the factor. See Cook and Campbell (1979) for a list of threats to external validity. Threats to statistical conclusion validity arise when conclusions are inferred outside the experimental conditions.

For example, in the study of the relationship between systolic blood pressure and age of children, any inference drawn from the study for people aged 25 or higher is obtained from extrapolation. We have only data for children up to 18 years of age and yet we attempt to infer blood pressure levels for people aged 25 or higher. This is risky business because while there may be evidence that a linear relationship holds for the two variables between 2 and 18, there is no reason to believe that the same linearity assumption applies to people in the higher age groups.

1.5.2 Internal validity

Internal validity is concerned with the question of whether the effect can be attributed to the variables that are manipulated. A 'good' design for a study should always control the threats to internal validity. In this subsection, we provide a short list of the main threats to internal validity and refer the reader to Campbell and Stanley (1963) and Cook and Campbell (1979) for a more in-depth discussion. The main threats are *history*, *maturation*, *testing*, *selection*, *regression towards the mean*, and *mortality*.

1.5.2.1 History

Changes in outcome variable may not be caused by the treatment variable but by an event that took place during the study. For example, in assessing the effectiveness of the biology lessons given by a teacher, higher grades of pupils on a biology test may not be attributable to the biology lessons (treatment) taught by the teacher, but by an on-going television home programme that was seen by these pupils during the same period of time. The event (television programme) represents a history threat and can be separated from the treatment effect by including a control group in the study, which does not receive the treatment (biology lessons).

1.5.2.2 Maturation

Change of outcome may be caused by changes that people undergo during the period of a study. This so-called maturation effect can also be controlled by inclusion of a separate control group of subjects with the same maturation effect.

1.5.2.3 *Testing*

Change in outcome scores may be induced by the effect of practice or memory. An example from education is that students being tested twice, may perform better the second time because they remember what kind of questions were asked the first time. Such effects are likely to arise in designs with repeated measurements

and can be controlled by counterbalancing the order of presentation or by the inclusion of a suitable control group.

1.5.2.4 Selection

In observational studies, subjects are not randomly assigned to groups. Controlling extraneous variables is usually difficult to achieve. This problem is referred to as the *selection effect* and it may account for variation of the outcome variable that is not caused by the effect of the treatment that one is interested in. Some control can be obtained by matching and blocking procedures. The idea behind blocking is that units are grouped in such a way that all units within the same block have the same characteristics and are likely to respond in the same way. Matching is grouping units into pairs, so that in the absence of treatment effects the pairs of units will produce the same responses.

1.5.2.5 Regression towards the mean

Regression towards the mean is related to the selection effect. This threat to internal validity can best be explained as follows. Any athlete knows that it is extremely difficult to beat one's own record a second time. How can this be explained? Assume that any measurement y consists of two elements: the true measurement y_T and measurement error e, that is, $y = y_T + e$. The true measurement is assumed to remain the same, while the measurement error changes due to chance. The record of an athlete is a high score and such a high score can be caused by a high true score y_T, and/or by a high value of the error e. The second time that the athlete runs under similar conditions, the high true measurement y_T will remain high, while the high error will be expected to become smaller due to chance. Overall, the measurements y are likely to become smaller the second time. Control of this effect in a design can be established by random assignment to groups or by inclusion of a control group.

1.5.2.6 Mortality

Incidental loss of data or dropout and loss of subjects during the process of the study can differentially affect the dependent variable measurements over different conditions and treatments. The dropout pattern can influence the internal validity. In selecting a 'good' design, one has to take this threat into account. For example, in the design of a clinical trial, it should not be longer than it is necessary to observe a meaningful effect because a longer trial is almost certain to have more dropouts and missing data than a shorter trial.

1.5.3 Control of (unwanted) variation

It is usually difficult to identify extraneous variables that distort the specific effects under study. These extraneous variables are often referred to as *confounders* and may interact with the independent variables. In general these

variables can be controlled by *inclusion*, *exclusion*, *statistical control* or by *randomization*.

1.5.3.1 Inclusion

Inclusion of an extraneous variable means that the variable is included in the study and that its specific effects on the dependent and other variables are taken into account. In fact the extraneous variable is added as an extra factor in an experimental design. For example, in a medical study, if one assumes that gender is a confounder and that it could distort the relation between dosage level and blood pressure, it can be 'controlled' by adding it as an extra factor in the design.

1.5.3.2 Exclusion

Exclusion or elimination means that the impact of the extraneous variable is eliminated from the design. In this way, the variable is held constant for all the units in the study. A potential distorting covariate may be gender. Holding a variable constant means that, for example, only male or female patients are included in a study. Of course, exclusion may affect the generalizability of the results negatively.

1.5.3.3 Statistical control

Control of extraneous variables is also possible via statistical manipulation. Examples of such a statistical control are inclusions of extraneous variables as covariates in a regression model or in an analysis of covariance. Schematically, the idea behind statistical control can be explained by Figure 1.1a and b.

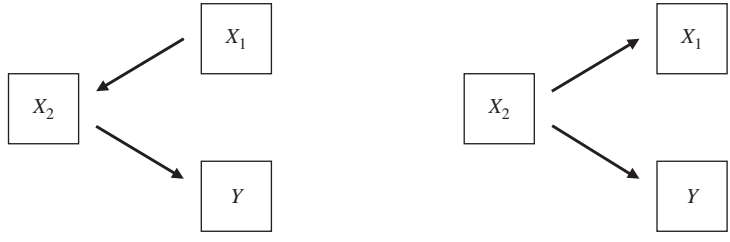


Figure 1.1 Two diagrams for the relation between variable X_1 and variable Y influenced by the effect of a third variable X_2.

The arrows in Figure 1.1a indicate that the variable X_1 affects X_2, which in turn affects Y. The relation between X_1 and Y is fully explained by the successive effects of X_1 on X_2 and of X_2 on Y. In Figure 1.1b, the variables X_1 and Y are both affected by X_2. One often refers to the relation between X_1 and Y in Figure 1.1b as *spurious*, because both have a common cause X_2. In both

diagrams, the variable X_2 plays a role in the relation between X_1 and Y. Any correlation between X_1 and Y would disappear if the variable X_2 is excluded or held constant. Proper handling of the variable X_2 in a regression analysis or analysis of covariance model would also lead to the same results.

1.5.3.4 Randomization as method of control

Randomization was introduced by Fisher (1926) for agricultural experiments comparing the effect of different fertilizers on the growth of wheat. Randomization plays a role in scientific studies in two ways. Firstly, randomization ensures generalizability of the conclusions. This is because random samples are often drawn from a finite population to generalize conclusions, and in a simple random sampling procedure, theoretically every unit in the population is equally likely to be included in the sample. Secondly, randomization can be applied to control the effects of extraneous variables. For example, random assignment of subjects to a treatment group and a control group can be established by flipping a fair coin for each subject: heads means assigning the subject to the treatment group and tails means assigning the subject to the control group. This process guarantees a 50–50% chance that each patient is assigned to either group and helps to ensure that all extraneous variables are equally distributed between the two groups. Of course, such a process can only be applied in purely experimental studies.

Randomization, however, should not be automatically done in all studies. Certain combinations of factor levels may not be possible and in some situations random assignment may not be practical or feasible. For example, random assignment in quasi- and non-experimental studies is usually not possible and control must be established by other procedures, such as restricted randomization, matching and blocking. See, for example, Box (1990) for a discussion on randomization.

1.5.3.5 Matching, blocking, restricted randomization and balancing as methods of control

There are also more or less *ad hoc* procedures to control for unwanted variation. For example in a learning study, students with the same age can be grouped into three treatment groups. The variable age is then *matched* to ensure that in all three treatment groups there are students with exactly the same age. In general, *blocking* procedures isolate or partition out variation that is attributable to an extraneous variable, so that it does not influence treatment effects and estimates of error variance. Blocks of units are homogenous with respect to the extraneous variable. *Balancing* is a procedure to obtain groups of subjects with, for example, the same number of students in the same age category. *Restricted randomization* is used to control the randomization in small samples to achieve a balance in group sizes or a balance between groups on other characteristics. An extension of balancing is *counterbalancing*, which is used in designs to control for possible order effects by changing the order of presentation of treatments. Examples of

such designs are the so-called cross-over design and the Latin square design. Cross-over designs are discussed in Chapter 8.

1.6 Ethical aspects of design choice

A 'good' research design should also meet a number of ethical requirements. Although ethical requirements may conflict with the methodological requirements, in an ideal situation ethical requirements can be reinforced by the methodological requirements of a study. A 'good' design in social, behavioural and biomedical studies should be both ethically and methodologically sound. A discussion on the benefits of both approaches is given by Palmer (2002).

Ethics codes for social, psychological and biomedical research have been developed via international consensus over the past 60 years. Starting with the Nuremberg Code (Nuremburg Code, 1947; Trials of War Criminals, 1949), research has been ethically guided by the various updates of the Declaration of Helsinki (2000).

Nowadays professional societies, such as the *American Medical* and *Psychological Associations*, have adopted these codes for the protection of the rights and interests of humans and animals (American Psychological Association, 2002). These codes specify what is required and what is forbidden and they are used by the various human subject research committees at universities and research institutes as guidelines in assessing research proposals. Apart from the commonly known principle of *informed consent*, these codes have quite firm requirements for the design of a study. Emanuel, Wendler and Grady (2000) proposed seven ethical requirements in designing clinical studies, which are also applicable to the design of studies in the social and biomedical field. Without discussing the ethical requirements in depth, we will briefly list the requirements proposed by Emanuel Wendler and Grady (2000).

Social and scientific value. The study design should enable evaluation of a treatment intervention to improve health and increase knowledge.

Scientific validity. A study should use accepted scientific methods, including statistical techniques to provide valid and reliable data. This requirement is closely related to the statistical conclusion validity of a study.

Fair subject selection. A study should not specifically select vulnerable individuals for risky research and reserve the socially strong subjects for the more beneficial research projects.

Favourable risk-benefit ratio. Potential risks to the subjects should be proportionate to the benefits to the subject and society.

Independent review. Reviewers of the proposed design of a study should be independent and not affiliated with the research project itself.

Informed consent. Subjects should be informed about the purpose of the research, its procedures, potential risks and benefits, so that the individual understands it and can make a voluntary decision whether to enrol, continue participation or stop.

Respect for potential and enrolled subjects. The research protocol should include respect for subjects by permitting withdrawal from research, by protecting privacy, by informing subjects of newly discovered risks or benefits and by maintaining the welfare of subjects.

It should be emphasized that the ethical requirements that focus on the well being of the subjects are important and can easily be overlooked or forgotten when a study is designed. Designers of research studies sometimes have the tendency to concentrate on the more methodological requirements of a good design than on ethical aspects of designing a study.

To summarize, a 'good' design for any type of study should ensure sufficient information and power for finding real effects in the region of interest, be robust against violations of model assumptions, enable adequate check of model fit, enable adequate check of model assumptions, allow for adequate control of extraneous variables, ensure simplicity of data patterns and computations, require minimum costs and sample size and enable valid extrapolation of the conclusions. In addition to these methodological requirements, ethical aspects should be taken into consideration as well, especially those that focus on the well being of the subjects and their informed consent.

1.7 Exact versus approximate designs

This book focuses on the use of optimal or highly efficient designs in social and biomedical research. Requirements for such optimal and highly efficient designs have been discussed in the previous sections. In this subsection, we make clear the distinction between two types of designs: *exact* and *approximate designs*. This distinction is important for understanding the material in the rest of this book and reasons for our choice to work with approximate designs instead of exact designs.

To understand the distinction between *exact* and *approximate designs*, let us consider a typical design problem for a dose–response study. The researcher has to decide in advance how to select from a given dose interval, *the number of dose levels* to use, the *dose levels*, and the *number of* subjects to assign to each of these dose levels. Suppose that the available resources allow us to take a fixed number of subjects, say N, in this study. An *exact design* tells us how many subjects to assign to each dose. If $N = 60$ subjects are available for the study, an exact design may require two different doses with 20 subjects given to the first dose level and 40 subjects given to the other second dose. Alternatively, if resources only allow $N = 57$ subjects, an exact design may allocate 22 subjects at a first dose level and 35 subjects at the second dose. Of course, these doses must all be selected from the dose interval of interest that was specified in advance.

Approximate design is another way to allocate a given number of N subjects. Such a design may specify that one-third of the subjects be given to a first dose level and two-third of the subjects be given to the other dose level. If $N = 60$, this results in a design that has 20 subjects at the first dose and 40 subjects at

the second dose. The defining characteristic of an approximate design is that it is specified by the number of dose levels, the dose levels and the *proportion* of subjects or units assigned to each dose. Approximate designs may appear to be exact designs but there is a key difference. Unlike exact designs, approximate designs can be specified regardless of the value of N. For instance, if $N = 180$ in the same example, the allocation scheme is the same, that is, one-third to the first dose and two-third to the second dose, resulting in the design that has 60 subjects at the first dose and 120 at the second dose, or vice versa. So the approximate design coincides with the exact design in this specific instance.

What happens if $N = 100$? The above approximate design then requires $100/3 = 33.3333$ subjects at the first dose and 66.6667 subjects at the second dose. Silly? Well, this is precisely the explanation behind its name. Approximation is the name of the game. Sometimes these designs are also called *continuous designs* suggesting that the allocation scheme has a continuum connotation. In practice, an approximate design is rounded in some natural way so that it becomes an exact design before implementation. For the approximate design just discussed, the implemented (rounded) design can take one of the following forms:

1. 33 subjects at dose 1 and 67 subjects at dose 2 or
2. 34 subjects at dose 1 and 66 subjects at dose 2.

So the design that we use in practice may not be unique, but, nevertheless, should be very similar to any one of the candidate designs listed above.

Kiefer pioneered the approximate design approach and his extensive work in this area is well documented in Kiefer (1985). There was criticism of this approach initially, but it is now widely accepted as a practical way of finding optimal approximate designs. Kiefer gave three powerful reasons for working with approximate designs instead of exact designs.

The first reason is that optimal exact designs are very difficult to find and they depend on the specific value of N in addition to being dependent on the model and optimality criterion. This means that for each model and each criterion, we need to have literally an endless list of optimal designs for the practitioners because each different value of the sample N results in a different optimal design. In contrast, the optimal approximate design is independent of the value of N and consequently they are easier to describe and be listed. Second, there is rounding involved when we implement an approximate design. One can show that the implemented optimal approximate design is always close to the (unknown) optimal exact design and the difference between the exact optimal design and the implemented design vanishes if N gets large. The third reason is that optimal exact designs usually require complicated mathematical theory and many optimal exact designs still cannot be found for relatively simple problems. In contrast, optimal approximate designs are available either in analytical form or they are found using iterative methods via computer algorithms. In summary, there are compelling reasons to work with approximate designs in practice and much of the rest of this book will focus on approximate optimal designs.

1.8 Examples

In this section, we first illustrate the research process using a simple hypothetical radiation dosage example. We then present studies taken from the social, behavioural and biomedical literature to illustrate how different types of design issues can arise in practice. For each study, we describe the problem and design issues, but defer how one may improve the designs to a later chapter.

1.8.1 Radiation dosage example

Suppose that a radiologist is interested in the linear effect of radiation dosage (X) on tumour shrinkage (Y) and assumes that the relation between radiation dosage levels and tumour shrinkage can be adequately described by the simple linear regression model. The mathematical relationship can be written as $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where β_0 is the intercept parameter and β_1 is the slope parameter. The tumour shrinkage for the ith patient is y_i and the radiation dose is x_i. We assume each error term ε_i has mean 0 and constant variance, and all observations are independent. In clinical practice, the range of dosage levels must be restricted, because an overdose can harm the experimental units, which are patients in this case. On the other hand, too small a dosage is likely to be ineffective. We assume for illustrative purposes here that the radiologist can only use eight equally spaced dosage levels and that sufficient funding is available to include only $N = 16$ patients. If these dosage levels are indicated by the numbers 1, 2, 3, ..., 8, a simple and intuitive design is to allocate n patients to each dose levels. This implies we must have $n = N/8 = 2$ patients per dosage level. Table 1.2 schematically shows this design for the study. The numbers 1 through 8 indicate the dosage levels and these dosage levels are equally spaced between the minimum and maximum dosage levels.

Table 1.2 A balanced radiation dosage design.

Dosage levels							
1	2	3	4	5	6	7	8
n	n	n	n	n	n	n	n

Following Section 1.2, we have now covered Stages 1, 2 and 3 of the research process. We identified radiation dosage as the only variable X to study its effects on the outcome variable Y, which is tumour shrinkage. That fulfilled Stage 1 of the process. The posited linear regression model between the tumour shrinkage and the radiation dose along with the choice of the design in Table 1.2 fulfilled Stages 2 and 3. In Stage 4, we collect data to test the scientific hypothesis of interest. In this case, the null hypothesis is that there is no linear relation between dosage level and tumour shrinkage, that is, $H_0 : \beta_1 = 0$. An important design question is whether this design is able to estimate β_1 efficiently and whether there

is sufficient power for testing the null hypothesis $H_0 : \beta_1 = 0$ to detect if the data support a linear relationship between tumour shrinkage and radiation dosage level.

The design issue for this problem is particularly pressing because we know that we have only a small sample of $N = 16$ patients and that measurement errors in radiation studies are usually quite large. This means it is absolutely crucial to choose the design carefully to minimize cost and maximize efficiency. What design would that be? Would it be more efficient to assign patients to a smaller number of dosage levels, such as dosage levels 1, 5 and 8? Or would it be more efficient to assign patients to the most extreme dosage levels 1 and 8? We will show in Chapter 2 that the design with one-half of the $N = 16$ patients assigned to dosage level 1 and the other one-half to dosage level 8 is the most efficient for estimating the slope parameter β_1. When the linear model assumption holds, this design that assigns equal proportions of patients to the extreme two dose levels is also optimal for several other purposes, but we defer further discussion to Chapter 2. Of course, an optimal design would still have to comply with ethical requirements. How efficient is the design in Table 1.2 compared to the optimal design for estimating β_1? It can be shown that this design is not efficient; in fact, its efficiency is only approximately 40%. We explain later on how this percentage is computed and what the efficiency means in practice.

1.8.2 Designs for the Poggendorff and Ponzo illusion experiments

Over the years, psychologists and cognitive scientists have been interested in the so-called Poggendoff and Ponzo illusions. These illusions are displayed in Figure 1.2. The Poggendorf illusion is that the ends of a straight line segment passing behind an obscuring rectangle seem to be offset when in fact they are aligned. The railroad or Ponzo illusion is that the upper horizontal line appears much longer than the lower horizontal line. In the Ponzo illusion, the subject is required to adjust the length of the lower horizontal line to match that of the upper horizontal line. In the Poggendorff illusion, the subject positions the

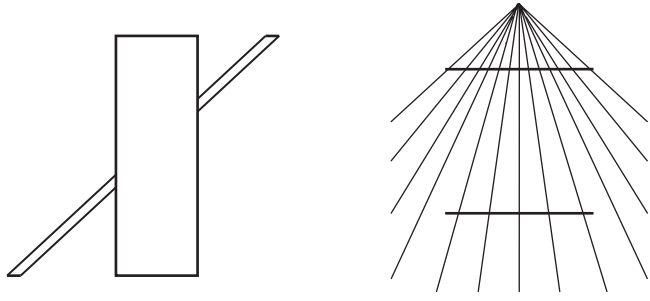


Figure 1.2 Graphs of the Poggendorf and Ponzo illusions.

right-hand bar as a continuation of the bar on the left. The subject's error in these adjustments (in inches) is measured and recorded as the magnitude of the illusion.

Bock (1975) reported two experiments conducted by Leibowitz and Gwozdicki (1967) and Leibowitz and Judisch (1967), which studied the magnitude of pictorial illusions as a function of age. The original design for each of the two studies was the same and had 16 children in each of the following age groups 5.0, 6.0, 7.0, 8.5, 10.5, 12.5, 14.5, 16.5 and 19.5. This design with a total of 144 children is summarized in Table 1.3. Bock (1975, Chapter 4) showed that the magnitude of the Poggendorff and Ponzo illusions as a function of age could be adequately described by a quadratic and cubic regression model, respectively.

Table 1.3 Design of Ponzo and Poggendorff studies.

	Age groups								
	1	2	3	4	5	6	7	8	9
Age	5	6	7	8.5	10.5	12.5	14.5	16.5	19.5
n	16	16	16	16	16	16	16	16	16

Figure 1.3 displays the fitted polynomials. It clearly shows the curvilinear relationship between age and the magnitude of the illusion. For the Poggendorff illusion, the magnitude of the illusion decreases as the age of the children increases and the magnitude increases again for older children (Figure 1.3a), whereas for the Ponzo illusion, the magnitude of the illusion first increases with the age of the children, then decreases and finally increases again (Figure 1.3b).

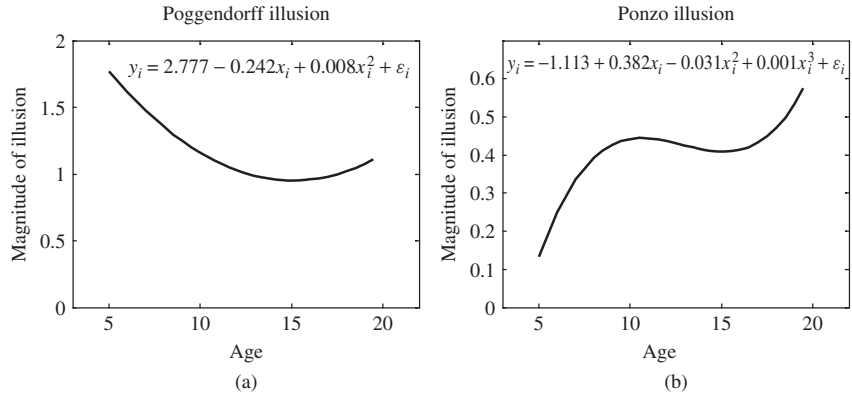


Figure 1.3 Magnitude of the Poggendorff and Ponzo illusions as function of age (data from Bock, 1975, Chapter 4).

The *design* for such studies requires selection of the age groups and the number of children in each of the age groups. One design question can be raised here is whether the nine age groups used in the original design are all needed to

estimate the polynomial functions that describe these two magnitudes of illusions. More specifically, how many distinct age groups are needed to estimate the quadratic and cubic function for these two illusions as efficiently as possible? Another question is whether these 144 children all contribute the same amount of information to estimate the parameters in the two functions. More importantly, is it possible to choose the number of age groups and the number of children to include in each of these age groups to obtain the most precise estimate of the polynomial relationship at minimal cost? In Chapter 3, we provide some answers to these questions using optimal design theory.

1.8.3 Uncertainty about best fitting regression models

Suppose that an epidemiologist studies the relationship between body weight and the joint effects of height and age of nutritionally deficient children. This hypothetical example is described in Kleinbaum *et al*. (1998, Chapter 8, Example 8.1) using data from 12 nutritionally deficient children with ages lying between 6 and 12 years. The heights of the children were roughly between 50 and 60 cm, and their weights roughly ranged from 50 to 80 kg.

The design question for this example is as follows: Which design will give us the most efficient estimate for the relationship between weight and the joint effects of height and age? This question is difficult and really quite impossible to answer unless further assumptions are made. For instance, what is the posited relationship between weight and the two predictors, height and age? Is the epidemiologist aware of postulated models in the literature or able to perform a pilot study to obtain a preliminary estimate of this relationship? Clearly, the answer to the design question depends on the assumed statistical model. What are plausible models? In the absence of good prior information, one considers the simplest possible models, and hope that the data will tell us which model seems most appropriate later on. From the design perspective, one can determine an efficient design for each plausible model and use these designs to come up with a reasonable model. In practice, researchers usually begin by considering simple linear models. Here are some common regression models to consider for two predictors X_1 and X_2 and a single outcome Y. For the ith child, we denote the value of the dependent variable, body weight, by y_i and the values of the independent variables, height and age, by x_{1i} and x_{2i}, respectively:

$$\text{Model 1: } y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i. \tag{1.1}$$

This model considers only the main effects for age and height on weight. If, in addition, we entertain an interaction between age and height, the model becomes

$$\text{Model 2: } y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \varepsilon_i. \tag{1.2}$$

The epidemiologist may also be interested in a second order effect for height (X_1). In this case, the model is

$$\text{Model 3: } y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \beta_4 x_{1i}^2 + \varepsilon_i. \qquad (1.3)$$

If a second order effect for age (X_2) is also entertained, the model becomes

$$\text{Model 4: } y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \beta_4 x_{1i}^2 + \beta_5 x_{2i}^2 + \varepsilon_i. \qquad (1.4)$$

Model 4 contains linear and quadratic terms for both the independent variables and also an interaction between these two variables. To complete the list of possible models for two variables, Model 5 below contains a second order term for both independent variables, that is, it is the same as Model 4, but without the interaction effect:

$$\text{Model 5: } y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_4 x_{1i}^2 + \beta_5 x_{2i}^2 + \varepsilon_i. \qquad (1.5)$$

Details on the actual analysis of such regression models can be found in Kleinbaum *et al*. (1998, Chapter 8), among others.

From the design viewpoint, we know that the choice of an efficient design depends on the model specification. The design problem confronting the epidemiologist is that data needs to be collected to validate the assumed model, but it is impossible to specify how data should be collected without a design. For this study, the design problem consists of the question what combination of levels of the independent variables (height and age) will provide the most efficient estimators of the parameters and how many children should be selected for each of these combinations. This design problem is explained in further detail in Chapter 3.

1.8.4 Designs for a priori contrasts among composite faces

Galton (1878) and Stoddard (1886) were probably the first to compose portraits of photographic exposures of faces. Langlois and Roggeman (1990) provided empirical evidence that composite faces seem to be 'better looking' than the original individual pictures. In their study, Langlois and Roggeman found a strong (curvi)linear relation between attractiveness and the number of faces entering the mixed or 'averaged' composite face. Illustrations of such a morphing effect can be found on the web site: http://www.beautycheck.de/.

The experiment performed by Langlois and Roggeman (1990) consisted of five levels of composite faces that were obtained by 'averaging', respectively, 2, 4, 8, 16, and 32 faces. The attractiveness was rated by subjects on a five-point Likert scale. Figure 1.4 shows the mean ratings of the five levels of female composite faces using data from Table 1 in Langlois and Roggeman (1990). The vertical lines in Figure 1.4 indicate the standard deviations at the five composite levels. It is assumed that faces per composite level were rated by different samples of raters.

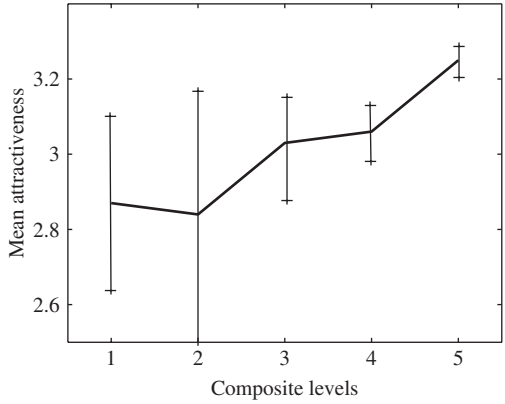


Figure 1.4 Average attractiveness ratings for female composite faces (data from Langlois and Roggeman, 1990).

Figure 1.4 shows that there is a relation between the composite level of the faces and their attractiveness. Overall, the trend shows that the faces were judged more attractive as more faces were entered. There also seems to be more consensus among raters about the attractiveness of the faces as the composite level increases. Figure 1.4 shows that the standard deviation tends to decrease as the composite level increases.

Although a clear trend is shown in the figure, it is not clear how the differences are manifested. Is this relation between attractiveness and composite level mainly a linear relation or is it also curvilinear? In the original design, the same number of raters was used for each composite level. In this study, all five composite levels are needed and so there is no question about the choice of number of levels to include. One can, however, take issue with the number of raters needed at each composite level. What is the optimal number of raters at each level to obtain the most efficient estimate for the linear or curvilinear effect of the composite level on the attractiveness? More details about this problem are discussed in Chapter 4.

1.8.5 Designs for calibration of item parameters in item response theory models

In educational research, a lot of resources are spent on the design of achievement tests, and various researchers have studied the problem of optimally designing achievement tests to decrease the costs of measurement. An example is the computerized GRE® General Test (Graduate Record Examinations), which measures verbal reasoning, quantitative reasoning, critical thinking, and analytical writing skills that have been acquired over a long period and that are not related to any specific field of study. The GRE test scores are frequently used as a criterion for admission to graduate studies in the United States.

Item response theory (IRT) models are usually used to estimate the characteristics of items in such a test. Van der Linden and Hambleton (1997) is

a useful reference for IRT models. An often used model is the two-parameter logistic (2PL) model, which models the probability of a student with ability θ_j to correctly answer an item as a logistic function:

$$p_i(\theta_j) = \frac{\exp\left[a_i(\theta_j - b_i)\right]}{1 + \exp\left[a_i(\theta_j - b_i)\right]}. \tag{1.6}$$

The response probability $p_i(\theta_j)$ for item i is modelled as a function of the ability level θ_j in the interval $-\infty < \theta_j < \infty$. The item parameter b_i represents the location parameter with range $-\infty < b_i < \infty$ and the item discrimination (slope) parameter is represented by a_i with range $0 < a_i < \infty$. The exponential function is $\exp(x) = e^x \approx 2.7184^x$. It should be noted that this is actually a logistic model, but with a quantitative latent variable θ_j. The difference between the logistic model described in Chapter 5 and the IRT model is that the IRT model assumes that the ability levels of students are unknown and have to be estimated as well, whereas the logistic model assumes a manifest independent variable. Usually, marginal maximum likelihood estimation of the parameters is applied with a normal density function for the distribution of the θ_js (Van der Linden and Hambleton, 1997).

Figure 1.5a shows a typical set of nine response functions from the 2PL model. These response functions vary in location and slope. Figure 1.5b shows one of these nine functions for which the slope $a_i = 1$. The dotted lines show that the probability of correctly answering the item is $p_i(\theta_j) = 0.5$ for a person with ability $\theta_j = b_i$. It is seen that as the ability of the student increases, the probability of answering the item correctly also increases, and that the steepness of the probability function depends on the size of the slope parameter a_i.

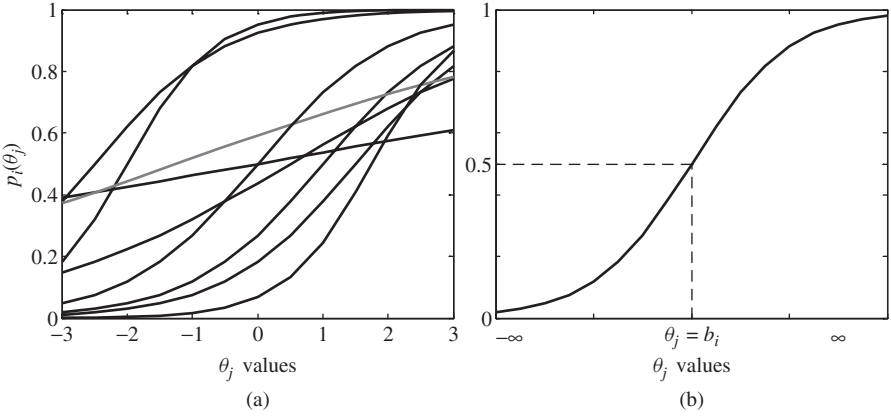


Figure 1.5 A set of nine typical response functions.

A computerized adaptive testing (CAT) setting is often used to estimate the ability levels of students and the item parameters as efficiently as possible. For

example, the GRE General Test is implemented in CAT form where a computer program sequentially selects items for a student to answer. The CAT procedure selects items that are not too difficult and not too easy for the student. It uses the information received from answers of a student to previously administered items. In this way, the ability of the student can be efficiently estimated with fewer items than with a traditional paper and pencil test. However, to adopt such CAT procedures, a sufficient large item bank with calibrated items is needed. To prevent the item bank from being exhausted by the CAT procedure, that is, when the items become out of date or over exposed, these item banks usually contain a huge number of items and builders of item banks have to conduct costly sessions to calibrate all these items.

Different procedures to build such huge item banks have been proposed. In many cases, calibration of items takes place by administering the items to a large fixed sample of students. This may, however, result in large parts of the data with little or no information on the item parameters. More efficient sampling designs have been suggested by Berger (1992, 1994), among others. The characteristics of the items that provide more efficient estimators for the item parameters are explained in Chapter 5.

1.9 Summary

This introductory chapter reviews the basic elements of a research design and lists some key requirements of a 'good' design. These requirements can be methodological, statistical or ethical in nature. The main focus of this book is to construct research designs that are as efficient as possible at minimal cost. Efficiency here can refer to the accuracy of the estimates for the model parameters or the power level of test or tests of interest. We distinguish two types of designs in the literature: exact and approximate (or continuous) designs. We also provide examples from the social and biomedical fields to illustrate different types of design problems we may encounter in practice and what design issues actually entail for each of the problems.