# 1

# *Introduction*

Repeatedly measured data are paramount in medicine, epidemiology, public health, psychology, sociology, and many other fields. The simplest case of such data is when a single measurement is collected repeatedly on the same individual or experimental unit. Each repeated measurement is called an *observation* and observations can be obtained over time, over a spatial map, or can be unordered temporally or spatially but nested (clustered) within larger experimental units.

*Clustered data* occur when repeated measures are not ordered and can be considered symmetrical within the larger experimental unit (cluster). For example, members of the same household can be interviewed, and in this case, their responses are repeated measures within the family. The family, rather than the individual, is the experimental unit and serves as the cluster. Observations on different individuals within the cluster are likely to be related to one another because individuals share the same environment and/or genetic predisposition. Similarly, patients may be clustered (or nested) within the same therapy group or clinic. Their treatment responses are also expected to be related because of the common influence of group or clinic, and can be considered repeated measures within the group or clinic. Several layers of clustering can be present in a data set. For example, the individual can be nested within family and the family can be nested within the neighborhood.

*Longitudinal data* occur when repeated measures are collected over time. In clinical trials in psychiatry and related fields, often the same rating instrument is administered to each individual at baseline, at intermediate time points, at the end of the randomized phase, and at follow-up. For example, depression severity can be measured weekly, biweekly, or monthly, in order to assess treatment effects over time. Similarly, in observational studies, the natural progression of a disease or other measures is ascertained repeatedly over time. In animal or human laboratory experiments, often responses from the same individual to different randomly ordered experimental conditions are recorded and compared.

*Spatial data* occur when repeated measures are spatially related. In imaging data sets, voxels are arranged in three-dimensional space where an observed value in a particular voxel is likely related to the observed values in neighboring voxels. In functional imaging studies, brain activation maps are created and often averaged region of interest signals are analyzed in order to measure and compare responses to different stimuli. In epidemiological studies, disease maps over geographical areas are created and analyzed. Methods for voxel-based data analysis of imaging studies and geographic and information systems are beyond the scope of this book, but we consider region of interest analyses of imaging data.

In all these situations, repeated observations within the same individual or cluster are related. Failure to take this interrelationship into account in statistical analyses, can lead to flawed conclusions. In this chapter, we review some terminology relevant to repeated measures data, such as mean response and measures of variability and correlation, present types of studies with longitudinal and clustered data, discuss advantages and challenges of collecting and analyzing repeatedly measured data, describe data sets that are used for

illustration throughout the book, and provide a brief historical overview of approaches for the analysis of correlated data. We focus on continuous (quantitative, dimensional) measures. Later chapters deal specifically with categorical measures which can be dichotomous, ordinal with few ordered levels, and nominal (unordered). Some statistical terminology and basic notation is presented in Section 1.7 and can be skipped by readers who are confident of their statistical knowledge of basic concepts. Statistical Analysis System (SAS) code for the graphs in this chapter and for all models considered in further chapters, together with actual output and available data sets, is available on the book website.

## 1.1 Aspects of Repeated Measures Data

### 1.1.1 Average (Mean) Response

The goals of many studies with repeatedly measured data are to estimate the average response in a population of interest and see whether it changes significantly as a result of treatment, exposure, covariates, and/or time. Herein, *response* is used in the sense of an outcome (outcome variable, dependent variable) that measures the main characteristic in the population of interest. *Population* is the target group of individuals for whom statistical inference should be generalizable and from where the study *sample* is obtained. For example, in depression studies, the response can be depression severity measured by a standard depression rating scale, such as the Hamilton Depression Rating Scale (HDRS), or a dichotomous measure of improvement defined as at least 50% decrease from baseline on the HDRS, and the population can be all individuals with major depression. In substance abuse studies, the response can be the percentage of days without substance use in a particular time period, and the population can be all individuals with alcohol dependence. In functional imaging studies, the response can be activation change in a brain region and the population can be all individuals, healthy or otherwise. The sample should be randomly obtained from the population if it is to be representative of the population of interest.

Average response refers to the mean of the individual responses in the sample or the population. In the simplest case of a single random sample from a population without repeated measures, the sample average response is just the arithmetic mean of all response values for the individuals in the sample (see Section 1.7 for exact formula). The population-average response is the mean response of all individuals in the population and, since it is usually not possible to measure, we use the sample mean to make inferences about the population mean. In longitudinal or clustered data, response is measured repeatedly within the individual over time or within the cluster, and the average response is usually a sequence or collection of numbers that correspond to each repeated measurement occasion. For example, the average response in a depression clinical trial that takes 8 weeks may be a sequence of eight averages of the individual responses (one for each week of the study). The average response in an imaging study may be a collection of several average responses, each corresponding to a different brain region.

Average response usually depends on a number of *predictors*. In clinical trials, we always have treatment as the main predictor of interest while participant characteristics such as age, gender, and disease severity are additional predictors that can also affect the response. Such additional predictors are usually called *covariates*. In observational studies, we might

be interested in the effect of exposures, such as smoking or drinking, on the response. In imaging studies, we may want to measure brain activation while individuals perform different tasks. In all these situations, estimation of the average response and how it depends on different predictors is of primary interest.

### 1.1.2 Variance and Correlation

The variability and interdependence of repeated measures within the individual or cluster are usually of secondary interest, although there are situations where they may be of equal or even higher interest than the estimation of the average response. For example, in clinical trials, the main goal may be to test whether an experimental treatment is on average better than a standard treatment or a placebo in terms of improvement in response over time. The variability in the responses of individuals needs to be taken into account but it is usually not of primary interest. However, it is possible that the experimental treatment may have a very similar average response to the standard treatment, but inter-individual variability in response may be lower (i.e., individuals may respond to treatment more consistently and similarly to one another). In this situation, the new treatment may be preferable and estimation of the variability of response is of interest too.

Variability of observations around the mean from a simple random sample is described by the *variance* or *standard deviation* of the observations (see Section 1.7). The sample standard deviation is often preferable as it provides a measure of variability that is evaluated in the same units as the mean. In repeated measures situations with longitudinal data, often the variability of the response at one particular time point differs from the variability at another, in which case it makes sense to estimate separate variances in order to assess data spread at individual time points. However, in some situations it may be reasonable to assume that the variances on all repeated occasions are the same. In this case, a better statistical estimate of the common variance can be obtained by pooling information from all occasions. Examples of both scenarios are considered in Chapter 2.

Repeated measures within individuals or clusters are often correlated. *Correlation* reflects the degree of linear dependence between two variables and varies between –1 and 1. It is important to emphasize that the definition includes the word "linear." Two variables may be perfectly related in a curvilinear fashion and have a correlation of zero. Correlation values of 1 or –1 correspond to perfect linear dependence between two variables. In these cases, knowing the values of one of the variables exactly predicts the values of the other variable, but does not imply that the two variables take the same value. Correlations are positive when larger values on one of the variables correspond to larger values on the other variable. Correlations are negative when larger values on one of the variables correspond to smaller values on the other variable. Please note that the proper statistical term for the latter case is "negative correlation," not "inverse correlation," as is often erroneously used. Section 1.7 shows how correlations are calculated and illustrates different scenarios.

Repeated measures within individuals, especially in longitudinal studies, are usually positively correlated although, in some situations, it is possible for observations to be negatively correlated. Individual responses tend to be systematically higher or lower than the mean response, which is reflected in a positive correlation. For example, if individuals start a study with higher illness severity than most other individuals, their repeated severity measures are likely to stay above average, at least for a while. Thus, repeated observations on the same individual are positively correlated with stronger correlation the closer the observations are to each other in time. This is very typical of longitudinal studies.

Some situations where negative correlation is more likely to be present are as follows: clustered data where individuals within a cluster may be competing for resources (e.g., individual weights of fetuses within a litter may be negatively correlated), longitudinal data where a positive response on one occasion makes a positive response less likely in another situation (e.g., immunity built after a viral illness may prevent a person from getting sick with the same or similar virus in the future), and clustered imaging data where a positive response in one brain region may occur simultaneously with a negative response in another region. Different types of statistical models provide varied levels of flexibility in specifying the structure of the correlations and variances within a data set. Subsequent chapters deal with this issue in detail.

## 1.2 Types of Studies with Repeatedly Measured Outcomes

Repeated measures can be collected on the same individual over time, on different parts of the body of the same individual, on members of the same family, or on individuals in clusters where measurements are expected to be related to one another, for example, students in schools or patients in clinics. The variability of the individual measures and interdependence between repeated measures can follow different patterns and needs to be properly taken into consideration in the statistical analysis of the data. Herein, we consider different types of studies with repeated measures data and discuss the implications of the patterns of variability in each of these situations for statistical modeling.

In *longitudinal studies* repeated measures are collected on the same individuals over time. Longitudinal studies are often prospective (i.e., individuals are recruited at a particular moment in time and followed up) and most often their focus is on assessing the effect of an intervention or an exposure over an extended period of time. They can also be used to ascertain trajectories of change and to compare temporal patterns of response of different groups of individuals. In some cases, subjects are assessed over time under experimental conditions (i.e., individuals are randomized to receive a particular treatment), whereas other times subjects are simply observed (i.e., when it might be unethical or too expensive to randomize individuals, for example, in studies investigating the effects of smoking or when analyzing the progression of a rare disease). These two types of studies are known as *experimental* and *observational*, respectively.

*Clinical trials* are the most common type of experimental longitudinal studies. Even though the primary endpoint of a clinical trial may be a single measure (e.g., time to remission, relapse, or outcome measured at the end of the trial), data are collected repeatedly on the same individuals over time. Double-blind randomized clinical trials, in which both the patient and the clinician are blind to treatment assignment, are considered the gold standard of evaluating intervention effects and are the only studies where direct causal interpretation is possible because randomization balances the study groups at baseline on potentially confounding variables for the relationship of the main predictor of interest and the outcome.

The simplest and most frequently used clinical trial design is the *parallel group design* where each individual is randomly assigned to an intervention and individuals receiving different interventions are followed in parallel. Participants do not switch treatments in this design unless necessitated for safety or other reasons. Figure 1.1 presents an example of such a design with three groups and four equally spaced assessment points over time.

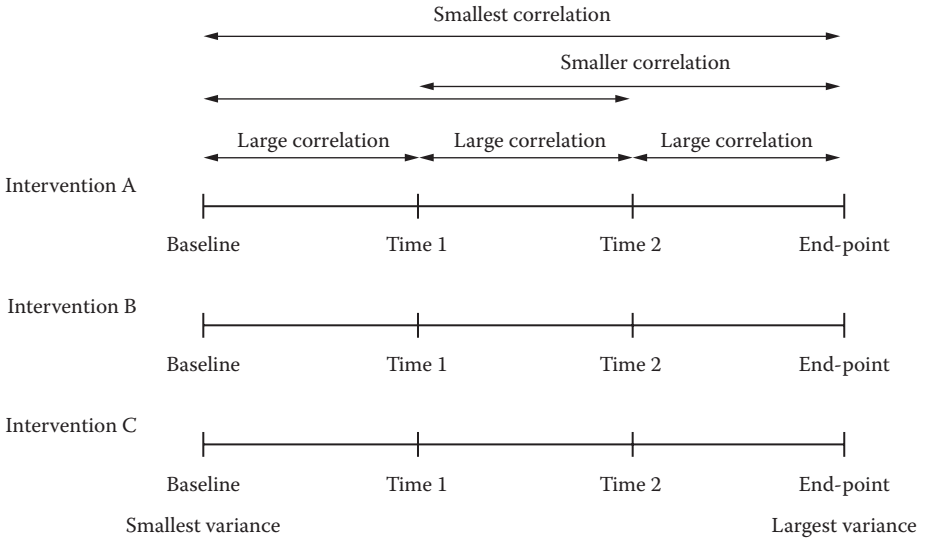


**FIGURE 1.1**
Parallel group clinical trial with three interventions and four repeated measures.

In the data example subsection of this chapter (Section 1.5), we present parallel group clinical trials in depression and alcohol dependence. In these studies, individuals are randomly assigned to different interventions and the outcomes (depression severity in the depression trials and drinking in the alcohol dependence trial) are repeatedly measured on individuals during treatment.

In such parallel group trials, the variability of repeated measures on the same outcome often increases over time because individuals are most comparable at baseline as they need to satisfy a strict set of inclusion/exclusion criteria. With time, differences emerge as some individuals respond more favorably to treatment than others. This leads to an increase in variability of the measurements toward the end of the trial. This increase is sometimes small and can be ignored, but occasionally the increase may be quite dramatic and needs to be taken into account in the data analysis. Furthermore, measurements within the same individual that are closer in time tend to be more highly correlated than measurements that are further apart in time. This is almost always the case and needs to be properly modeled so that statistical inferences are valid. Different ways to take into account the pattern of correlations in the statistical model are discussed in Chapter 3.

Another frequently used clinical trial design, is the *cross-over design*, which is popular in both human and animal laboratory studies. In this design, individuals are assigned several treatments in randomized order, that is, they are randomized to a particular sequence of treatment assignments. Figure 1.2 shows an example of a cross-over design with three treatments. In Section 1.5, we present a human laboratory study in which smokers received different doses of nicotine and menthol in randomized order. The outcome was nicotine reinforcement and was measured for each menthol and nicotine dose combination. The study focused on assessment of the independent and interactive effects of nicotine and menthol.

The cross-over design can be particularly useful when individuals vary considerably in their response from one another, repeated measures on the same individual are substantially correlated, when interventions are relatively short in duration, and when there is no or low possibility of carry-over effects from one treatment to another. Carry-over

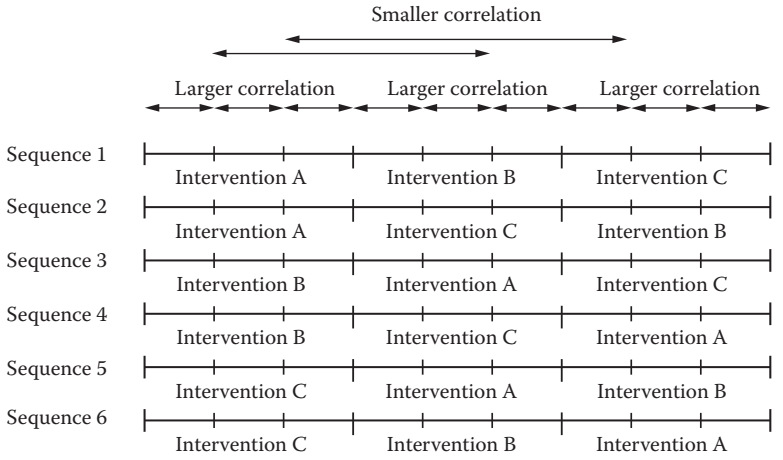


**FIGURE 1.2**
Cross-over trial with three interventions and three repeated measures within intervention period.

effects are minimized when a sufficient washout period is allowed in between treatments. This design is more efficient (i.e., can detect differences between treatments with greater power or with fewer individuals) than the parallel group design when there is large between-subject variability because it allows a direct comparison of the treatment effects within each individual, i.e., each subject serves as their own control. However, this design may also be associated with higher probability of dropout and order effects must be controlled. It is also difficult to implement in scenarios when there are carry-over effects and treatment effects take long to manifest. In cross-over designs, correlations between repeated measures on the same individual within the same treatment period are usually higher than correlations between repeated measures on the same individual from different treatment periods. Within each period, correlations can be modeled using the same approach as correlations from a parallel group clinical trial.

*Observational longitudinal studies* usually follow groups of individuals over time. For example, in the Health and Retirement Survey, presented in Section 1.5, individuals aged 55 and older were followed more than 10 years with interviews every 2 years. Participants in the study of association between unemployment and depression were interviewed up to three times in a period of up to 16 months after job loss.

Observational longitudinal studies are used when it is not possible or ethical to randomize individuals, or when it may be too expensive to perform an experimental study. For example, assessing the effect of smoking, or of genetic factors, on the emergence or progression of some disease over time can only be performed using observational studies. In these studies, the same issues about modeling variability and correlations between repeated measures on the same individual as in randomized clinical trials apply.

In studies where the same outcome is measured on related individuals (e.g., twins, sibling pairs, or members of the same family) or in clustered settings (e.g., individuals within the same clinic or treated by the same doctor), correlations are also expected to be present. For example, in the Health and Retirement Survey, data were collected on married individuals. In the mother–infant study presented in Section 1.5, there are positive correlations between mothers and their infants. Correlations may be naturally occurring (e.g., individuals living together or genetically related are expected to exhibit some level of correlation on some responses) or may be introduced by the researcher via the study design. For

example, in cluster-*randomized clinical trials*, the units receiving a particular intervention are clinics, not individual patients. However, the responses to the intervention are usually measured on the individual patients within clinics. Failure to take into account the correlation between the observations on different patients within the same clinic can result in erroneous conclusions. Regardless of the reason for correlations between the observations in clustered settings, statistical methods need to appropriately model this correlation in order to provide valid results.

Unlike in situations with longitudinal data, where correlations are stronger or weaker depending on the time lag between observations on the same individual, in clustered settings with only one level of clustering it is likely that observations within clusters are equally correlated while observations from different clusters are uncorrelated. For example, observations on individuals within the same clinic may be correlated but observations on individuals from different clinics should be uncorrelated. Additionally, the variances of the observations on individuals in the same cluster are expected to be the same. This structure of variances and correlations is the simplest possible structure in repeated measures scenarios and is called *compound symmetry structure*. This structure will be presented in more detail in Chapters 2 and 3.

Different levels of clustering are also possible. For example, individuals can be nested within families and families can be nested within neighborhoods, which leads to different levels of correlations within the family and within the neighborhood. In this case, a multilevel version of the compound symmetry structure may arise, such that observations on members of the same family are strongly correlated, observations on members of different families but living in the same neighborhood may be weakly correlated, and observations on members of different families in different neighborhoods are uncorrelated. The variance of each observation in this case can be represented as the sum of the variance due to neighborhood, the variance due to family, and the variance due to the individual. Figure 1.3 illustrates the situation with two different levels of clustering: individuals are

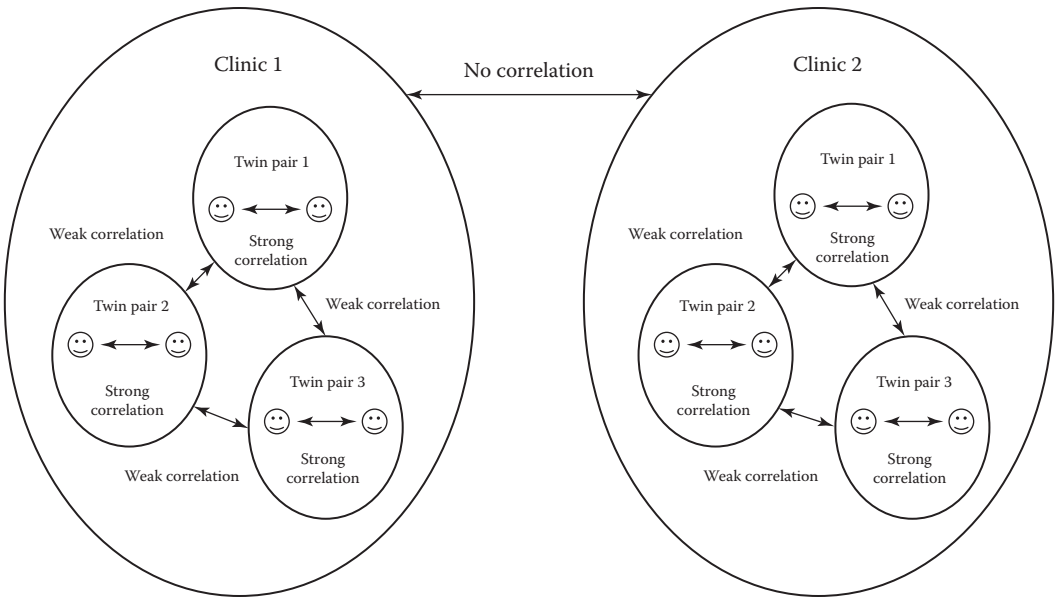


**FIGURE 1.3**
Clustered data with two levels of clustering.

nested within twin pairs and twin pairs are nested within clinics where they might be undergoing a particular treatment together with other twin pairs. More than two levels of clustering may also be present and need to be taken into account for proper statistical inference.

It is also possible to have correlations in a data set due to both clustering and to repeated observations on the same individual over time. For example, some interventions in psychiatry and related fields (e.g., some behavioral interventions) are administered in a group setting (e.g., therapy group) and then participants' responses are repeatedly assessed over time. Since the same therapist is providing treatment to a group of individuals at the same time, correlations may arise between measurements on individuals in the same group. These correlations are in addition to the correlations that exist between repeated measures on the same individual over time. This leads to layers of correlation in the data. In particular, measurements on the same individual at two adjacent time points are usually strongly correlated and measurements on the same individual further apart in time are weakly correlated. Furthermore, measurements on different individuals within the same therapy group are correlated, with the strongest correlation observed for pairs of observations at the same time point and correlations decreasing with time lag between observations. The models considered in Chapters 3 and 4 demonstrate how such fairly complicated situations can be seamlessly and appropriately handled and illustrate the methods on data introduced in Section 1.5.

*Imaging data* present their own set of challenges in accounting for correlations between observations because of the spatial relationship between units of analysis. Different imaging techniques (e.g., magnetic resonance imaging [MRI], functional magnetic resonance imaging [fMRI], and diffusion tensor imaging [DTI]) and different resolution levels may require different techniques for the handling of correlated measures. At the highest resolution level of MRI or fMRI analyses, signal intensities at adjacent voxels (points in three-dimensional space) are expected to be more highly correlated than observations at voxels that are more distant. Also, variability of signals measured at individual voxels do not in general vary considerably. However, at the level of region of interest (ROI), where sets of voxels have been combined into anatomically or functionally defined areas and overall measures have been calculated over the entire ROI, distance between regions is no longer of utmost importance and measures in regions closer in space are not necessarily more highly correlated than measures from regions that are further apart. Moreover, variances of measures in different regions may be vastly different from one another due to the size of the region or other factors. In-depth consideration of issues and techniques of analysis of different types of imaging data are beyond the scope of this book. However, in Chapter 3, we discuss techniques for analysis of ROI data and illustrate with the data from the fMRI study of working memory in schizophrenia, introduced in Section 1.5. For details on issues in brain imaging analysis the interested reader is referred to Chung (2014). Friston (2007) provides a detailed overview of statistical parametric mapping for functional brain images.

## 1.3 Advantages of Collecting and Analyzing Repeatedly Measured Data

The main advantage of collecting repeatedly measured data is that each individual or experimental unit serves as their own control. When an intervention or exposure can

be varied within an individual, repeated measures allow one to assess or compare the effects within the subject. This means that the effects of potentially confounding variables that vary between subjects can be controlled and, as a result, the variability of estimating effects of an intervention or exposure is reduced, compared to studies at a single time point (*cross-sectional studies*). This is reflected in increased power for within-individual or within-cluster comparisons, that is, there is higher probability of finding differences in response within clusters or individuals when true differences exist.

Furthermore, patterns of change over time can be assessed when longitudinal data are collected. Prospective studies collect information over a period of time, starting at study entry. Clinical trials and cohort studies are examples of prospective studies. In such investigations, repeated measures on a number of variables can be collected on the same individuals. This allows one to estimate trajectories over time, to test between-group differences on trajectories over time, and to assess variability of measurements both within and between subjects. Such studies have greater *statistical power* for testing time effects and differences between groups over time than corresponding cross-sectional methods. Additionally, such analyses that take into account the variability and correlation of repeated measurements can better control the probability of finding differences where true differences do not exist (i.e., better control of *type I error* in statistical testing).

## 1.4 Challenges in the Analysis of Correlated Data

To analyze a data set with correlated data, proper statistical models need to be constructed. In this section, we consider parametric models in which all aspects of the models need to be specified. *Non-parametric models* that do not make specific assumptions about the distribution of the response variable are considered in Chapter 5. *Semi-parametric models* that make assumptions about some aspects of the response distribution and leave others unspecified, such as *generalized estimating equations* (GEE), are considered in Chapter 4.

The first aspect of the statistical model, is the specification of the patterns of the means of the response variable within a cluster or over time. The means are usually assumed to depend on individual or cluster characteristics, on predictors that may vary within a cluster or over time, and very often on time and treatment. The model should provide a good smoothing of the unknown true relationship between predictors and the response that is useful for a relatively simple description of reality. In the model definition, the relationship is described by a mathematical equation, which is usually a linear function of the predictors and hence is called a *linear predictor* (see Section 1.7 for model definition). Some statistical models are more general and assume non-linear relationships. But in all cases, the equation that describes how the mean response varies as a function of the predictors needs to be matching reality reasonably well. This requires that measured predictors that affect the mean be included in the equation and that the form of the equation corresponds to the relationship between the predictors and the mean outcome. An example of poor correspondence between model and reality is when the model assumes that the response varies linearly with the predictor but in fact the relationship is curvilinear. If the mean response is not correctly specified, other aspects of the model definition can be affected and statistical inferences may be misleading.

The second aspect of the statistical model is the *distribution of the response variable*. When the response variable is continuous or approximately continuous (e.g., scores on

instruments assessing symptoms of depression or schizophrenia), the natural and mathematically most convenient distribution that is considered first is the normal distribution. However, if a histogram of the response distribution does not appear approximately bell-shaped (with most observations in the middle and a few large and small observations in the tail of the distribution), then directly assuming normal distribution can lead to problems with the conclusions from the statistical analysis, especially in small samples. One possible solution to this issue is to apply a transformation to the response variable prior to analysis, in order to have the distribution of the transformed variable more closely resemble the normal distribution, and then use models for normally distributed data. This is mathematically most convenient but is not straightforward to interpret as all statistical estimates are on the transformed scale and in general can't be directly transformed back. Chapter 3 is devoted to models for repeated measures with a normal response. Alternative distributions can be considered for continuous response variables and some of these models are considered in Chapter 4. Chapter 4 also covers models for dichotomous (binary) and count data.

The third aspect of the model formulation is describing the variances of repeated observations and the correlations between repeated measures within clusters and/or over time. In Chapters 3 and 4, we consider different approaches for accounting for the variance and covariance, based on mixed-effects models and estimating equations.

In summary, due to the complexity of the variances and correlations between repeatedly measured observations, formulating an appropriate statistical model is challenging and should be done in steps with proper checks of each aspect of the model formulation. Descriptive statistics and data visualization techniques are used to inform decisions about model formulation. Such techniques are illustrated using the data sets in the next section.

## 1.5 Data Sets

Several data sets are considered for illustration of the methods described throughout the book. We consider data sets from both clinical trials and observational studies, with longitudinal and clustered data, with balanced and unbalanced designs. Most data sets have missing data which can present problems for analysis. Specific features of the different data sets are emphasized and graphical and tabular methods for data exploration are presented.

### 1.5.1 Augmentation Treatment for Depression

The first data set is from a parallel group clinical trial of an *augmentation treatment for depression* (Sanacora et al., 2004). In this study, 50 patients were randomly assigned to either fluoxetine + yohimbine (augmentation treatment) or fluoxetine+placebo (control treatment) for six weeks. The main study hypothesis was that the augmentation treatment group would show faster improvement than the control group on the HDRS total score, which measures severity of depression symptoms. The primary analysis of these data reported in Sanacora et al. (2004) showed that patients in the augmentation group achieved responder status (HDRS score of 10 or less) faster than the control group. In subsequent chapters (e.g., Chapters 3, 6, and 7) we use this data set to illustrate model fitting that allows for comparison of the treatment response profiles of the two groups over time.

But before we proceed with consideration of the statistical models, we explore the data with some graphical representations. Figure 1.4 shows two *profile plots* of the HDRS scores of all patients in the study by treatment group. This type of plot (also known as a *spaghetti plot*) is useful for visualizing longitudinal data in small- to medium-sized data sets, as it shows the individual trajectories of observed responses over time. It provides a visual impression of the mean trend over time, the variability of observations, and the strength of correlation between adjacent observations of individuals over time. From Figure 1.4 we see that the individuals in the two treatment groups appear to have similar baseline scores, although the scores in the control group are slightly higher. We also see that most patients have substantial decrease in depression severity over time and that variances appear to slightly increase from baseline, especially for the control group in the middle of

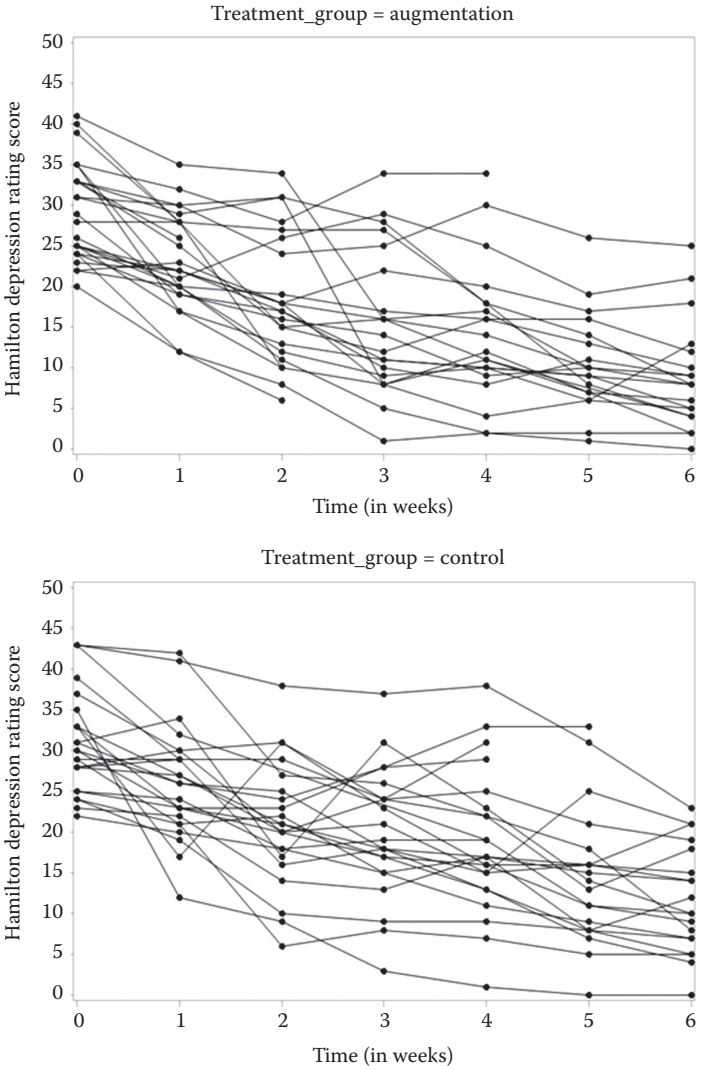


**FIGURE 1.4**
Profile plots of Hamilton Depression Rating Scale scores of all subjects in the augmentation depression data set by treatment group.

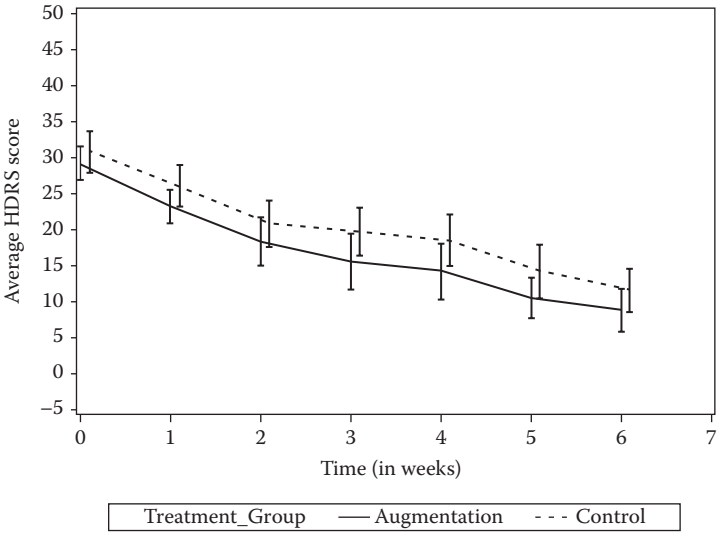


**FIGURE 1.5**
Means and standards error of the Hamilton Depression Rating Scale scores over time by treatments group in the augmentation depression data set.

the observation period. Furthermore, most patients' responses tend to stay either above or below the corresponding average scores of their respective treatment group which suggests that correlations between repeated observations within the individual are positive.

Figure 1.5 presents a plot of the means and standard errors of the means over time for each treatment group, based on all available observations at each time point. This plot confirms our observations from Figure 1.4 concerning the average trend and the variability over time. However, it does not provide any information about the correlation of observations within individuals. Also, in the presence of missing data, it may present a distorted picture of the average trends over time. For example, if subjects in the control group selectively drop out due to inefficacy of the treatment and subjects in the active group drop out due to side effects, the between-group differences shown in the figure at later time points may be smaller than the real differences. Analysis in the presence of missing data is considered in detail in Chapter 7 and these data are used for illustration.

Despite the limitation of the mean plot, it is quite useful for spotting changes in average treatment response between groups over time and can be used with a data set of any size. Since standard errors of the means decrease with increasing sample size and the corresponding error bars get tighter around the means, sometimes the same type of plot is created with bars corresponding to standard deviations, rather than standard errors of the mean. Standard deviation estimates do not in general decrease with increasing sample size and provide an estimate of the variation of individual observations in the sample, rather than of the means.

### 1.5.2 Sequenced Treatment Alternatives to Relieve Depression (STAR*D)

The second data set is from the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) clinical trial (Gaynes et al., 2008; Trivedi et al., 2006). STAR*D is the largest randomized prospective study of outpatients with major depression to date. The first stage of this study was a 12-week course of citalopram, a selective serotonin reuptake inhibitor

(SSRI) antidepressant and the outcome of interest was improvement in the total score from the Quick Inventory of Depression Symptomatology (QIDS) (Rush et al., 2006) questionnaire. Four thousand and nineteen subjects provided data on QIDS in the first phase. The total QIDS score is similar to the total HDRS score considered in the previous example and reflects total depression severity.

Prediction of initial response to antidepressant treatment in STAR*D was recently considered by Chekroud et al. (2016) with responder status over the entire 12 weeks defined, based on the improvement in total QIDS score. Subsequent analyses identified three different clusters of depression symptoms (core depression symptoms, sleep symptoms, and atypical symptoms) in this study and in two other large clinical trials in depression, and showed that treatments are not equally effective for the three clusters (Chekroud et al., 2017) across trials. In this book, we focus on the effects of citalopram treatment in the first phase of the STAR*D trial on the three clusters and illustrate how the models introduced in Chapter 3 can be used to model the three aspects of depression severity simultaneously.

The design of the study was intended to be balanced with subjects scheduled for visits at weeks 0, 2, 4, 6, 9, and 12. However, participants were sometimes seen in intermediate weeks and occasionally had repeat visits during the same week. Thus, the measurement times of subjects were somewhat different and the design was actually unbalanced. This limits the set of possible approaches that could be used for such data and requires that one makes the assumption that the time points are independent of the outcome and of the other effects in the model. This assumption is considered in more detail in Chapter 3.

Figure 1.6 shows a *panel plot* of the observed symptom cluster scores over time of three participants in the study. The rows correspond to different individuals and the columns correspond to different clusters. Each dot in the graph is the average of the scores on the individual items for the participant in the cluster of symptoms. The observation times are different for the three individuals but are all within the 12-week period. The superimposed regression lines are based only on the observations in the graph and are used to illustrate visually the linear trend in change over time. Note that the linear trend does not necessarily fit the data well. In particular, the trends of change in the sleep cluster appear curvilinear for these three individuals, but since these are only a few of the participants, we can't make any conclusions about the pattern of change over time for the entire sample.

The panel plot is an alternative way of presenting individual change over time to the spaghetti plot, but it is limited to showing only a few individuals at a time. We consider these data in more detail in Chapters 3 and 10.

### 1.5.3 Combined Pharmacotherapies and Behavioral Interventions for Alcohol Dependence (COMBINE) Study

The Combined Pharmacotherapies and Behavioral Interventions for Alcohol Dependence (COMBINE) Study (Anton et al., 2006) represents the largest study of pharmacotherapy for alcoholism in the United States to date. This parallel group clinical trial was designed to answer questions about the benefits of combining behavioral and pharmacological interventions on drinking outcomes in individuals with alcohol dependence. Eight groups (1226 participants in total) received medication management and a combination of active or placebo naltrexone, active or placebo acamprosate, and combined behavioral intervention (CBI), or no CBI in a $2 \times 2 \times 2$ factorial design (Figure 1.7). A ninth group received only CBI without medication management and is not considered herein. Double-blind medication treatment (naltrexone and acamprosate) and CBI were provided for approximately four months and participants were followed for up to one year after randomization.

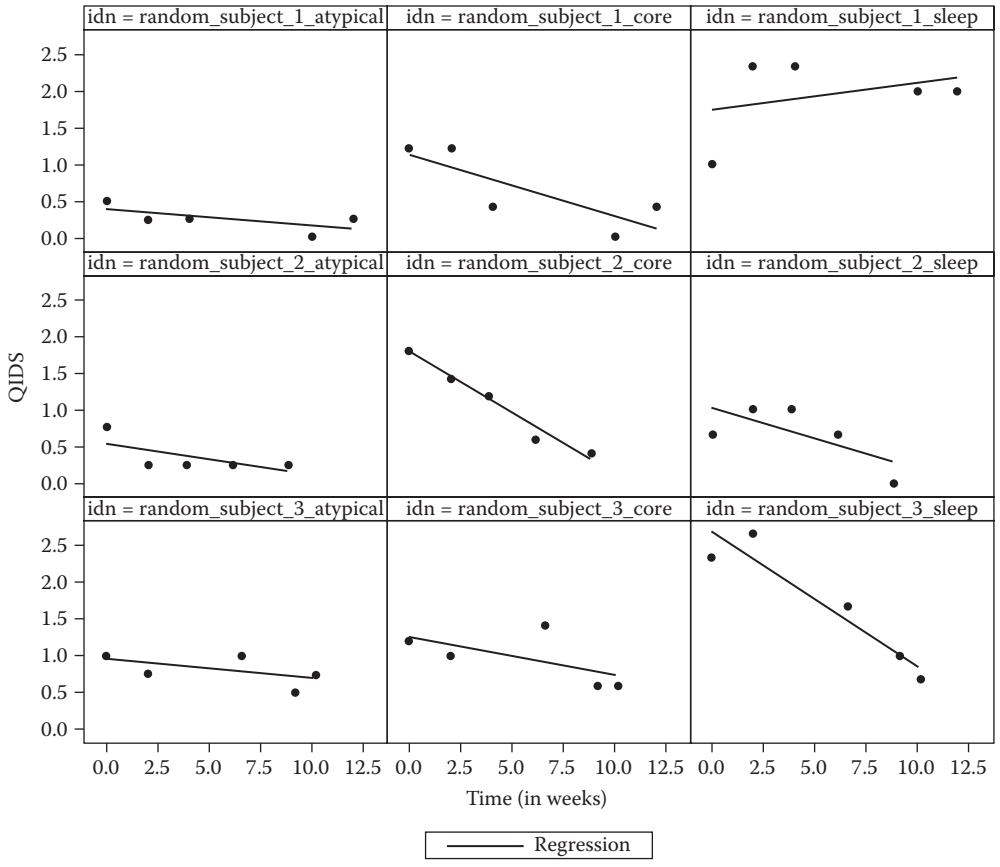


**FIGURE 1.6**
Cluster symptom scores of three individuals from the STAR*D trial in depression.

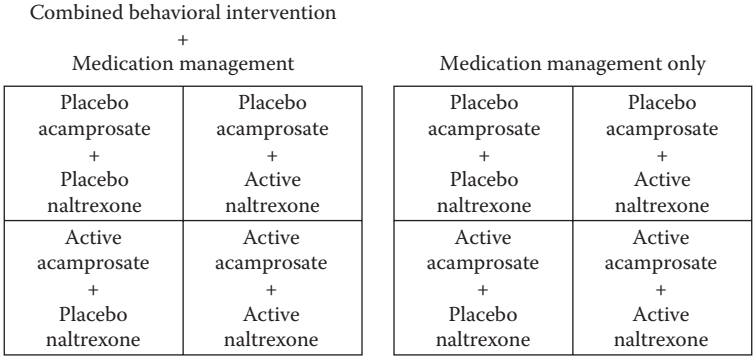


**FIGURE 1.7**
Design of the COMBINE study.

Participants were required to abstain from drinking for at least four days prior to randomization and the main outcomes in the primary analyses were: time to first heavy drinking days and percent days abstinent over the entire treatment period. The primary analyses found significant benefit of naltrexone and CBI on drinking outcomes, but the combination of naltrexone and CBI was not better than the monotherapies. The effects of acamprosate were not significant.

Drinking data in this study were collected daily using the timeline follow-back method (TLFB). This method was also used to collect daily drinking data for the 90 days prior to the baseline assessment and during follow-up. Since daily drinking data are available, it is possible to look at changes in drinking patterns over time pre-treatment, during treatment, and during follow-up. In this book, we use monthly summaries of drinking data for illustration, which allow estimation of trajectories of treatment response over time and allow us to ignore variability in the daily measures due to the day of the week. Depending on the model that we are illustrating, we focus on several different measures: average number of drinks per day, average number of drinks per drinking day (day on which drinking occurred), and number of drinking days.

Figure 1.8 shows the average number of drinks per drinking day in four treatment arms during the treatment period. Note that this measure is calculated only for the drinking days and thus reflects only one aspect of drinking behavior (i.e., intensity of drinking). Additional outcomes, such as percent of drinking days, need to be considered in order to describe other aspects of drinking (i.e., frequency of drinking). We consider these aspects separately in subsequent chapters. Joint analysis of the different aspects is also possible but requires more sophisticated statistical models and is beyond the scope of this book. Interested readers are referred to Liu et al. (2008) and Liu et al. (2012) for more details.

In Figure 1.8, we omit the standard errors of the means from the graph and also ignore acamprosate assignment in order to have a less cluttered figure. This type of figure shows

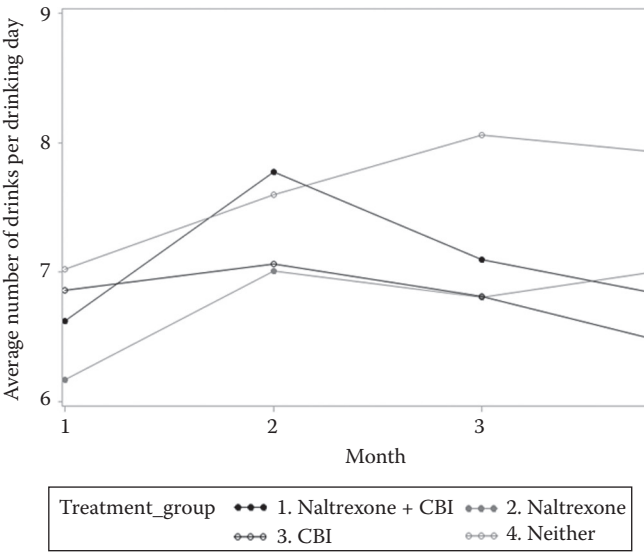


**FIGURE 1.8**
Average number of drinks per drinking day by treatment group in the COMBINE data set.

the average trends over time but does not provide information about the data distribution at each time point. Profile plots will not be very useful for data visualization in this study because there are several hundred participants within each group and it will be difficult to distinguish the individual trajectories. Standard error or standard deviation bars will also be hard to distinguish if added to Figure 1.8 because there are several groups and many time points. Instead, histograms or box plots can be used to visualize the data distribution at each time point. Figure 1.9 shows *box plots* of the outcome (average number of drinks per

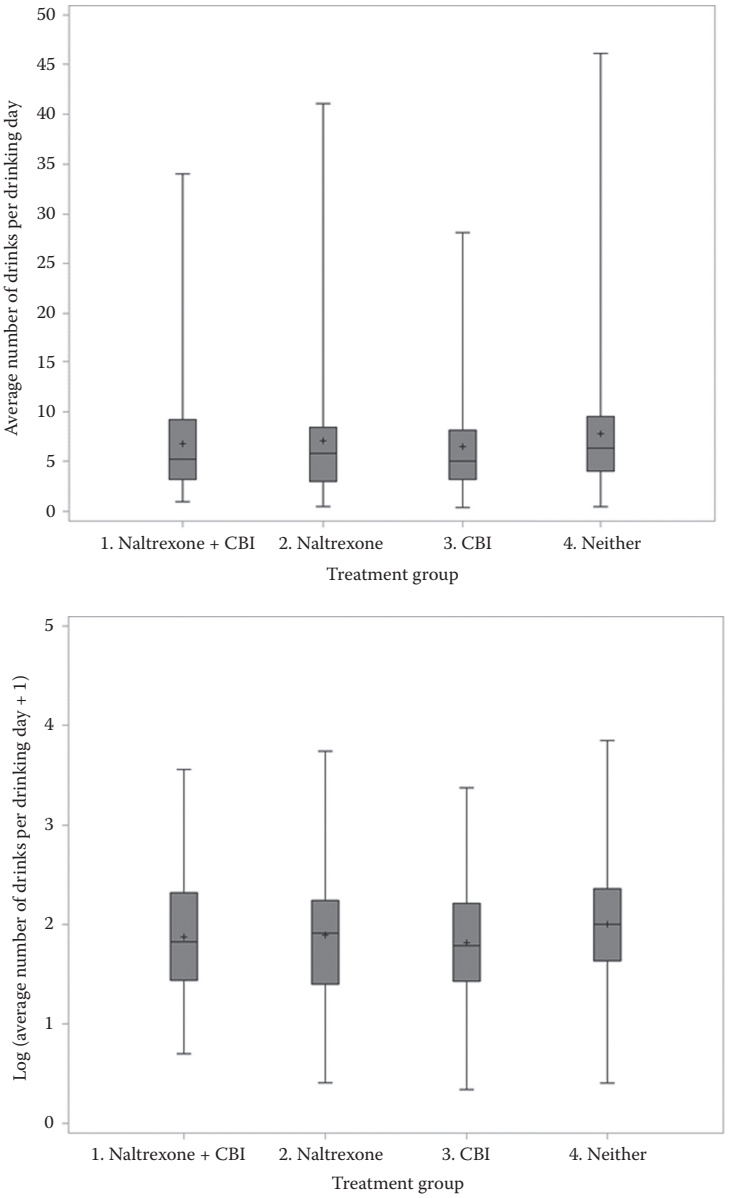


**FIGURE 1.9**
Box plots of average number of drinks per drinking day by treatment group at month four in the COMBINE study.

drinking day) at month four by treatment group. In the top panel of the figure are four box plots of the original response variable. In each box plot, the middle line shows the median of the data (i.e., the value below which 50% of the observations lie), the lower and upper ends of the box show the 25th and the 75th percentile, respectively (i.e., the values below which we have 25% and 75% of the observations, respectively), the plus (+) sign shows the mean of the data and the whiskers of the box plot show the minimum and maximum value in the corresponding treatment group. From this set of box plots, we see that drinking data are right-skewed (i.e., there are a few large observations whereas the majority of the observations are clustered together at the lower end of the scale). When data are right-skewed, usually a transformation is applied prior to statistical analysis and the log transformation is most commonly used. The bottom panel of Figure 1.8 shows the box plots of the data after the data have been transformed using the log transformation. We add 1 to each observation prior to transformation in order to avoid problems with taking log of values that are equal or close to 0. The box plots of the transformed data show that the transformation makes the data more symmetric and the medians and the means are much closer to one another than before the transformation.

In subsequent chapters, we show how to fit different statistical models to the COMBINE data to assess changes over time and the effect of baseline covariates on trajectories over time. We use these data to illustrate models for longitudinal data with continuous and categorical responses, mixture models for empirical derivation of heterogeneous trajectories over time, and assessment of moderating and mediating effects. We also demonstrate how to interpret significant interactions and main effects via appropriate post-hoc comparisons.

### 1.5.4 The Health and Retirement Study

The Health and Retirement Study (HRS, http://hrsonline.isr.umich.edu/) is a longitudinal survey among American citizens born between 1931 and 1941 and their spouses that assesses changes in labor force participation and health status over the transition period from working to retirement, and the years after. The initial HRS panel (N=12,652) was first interviewed in 1992 with subsequent interviews taken every two years. This survey is an observational longitudinal study that provides a wealth of information to address important questions about aging and the transition from working to retirement. In this book, we focus on changes in self-rated health (SHLT) and body-mass index (BMI), and the effects of covariates on these changes. Body-mass index is calculated as weight, in kilograms, divided by the square of height, in meters, and is considered a continuous measure. Self-reported health is an ordinal measure that takes the following possible values: excellent (coded as 1), very good (2), good (3), fair (4), and poor (5). BMI increases on average over time while SHLT deteriorates on average over the first seven waves of data. Mean plots with or without variance estimates could be created to illustrate this, as shown for the previous data sets.

Within this data set, correlations are present between repeated measures on the same variable within individuals and on different variables within individuals. Figure 1.10 is a scatterplot matrix that shows the distributions of BMI and SHLT at the first wave and at the seventh wave, and gives visual clues as to whether the measurements are correlated and in what direction the correlation is. For this plot, we chose a subset of 250 individuals since plotting the entire data set of several thousand individuals would have made the graphs hard to read. Each of the individual plots in the scatterplot matrix illustrates the relationship between two variables and to a certain extent the distribution of each variable.

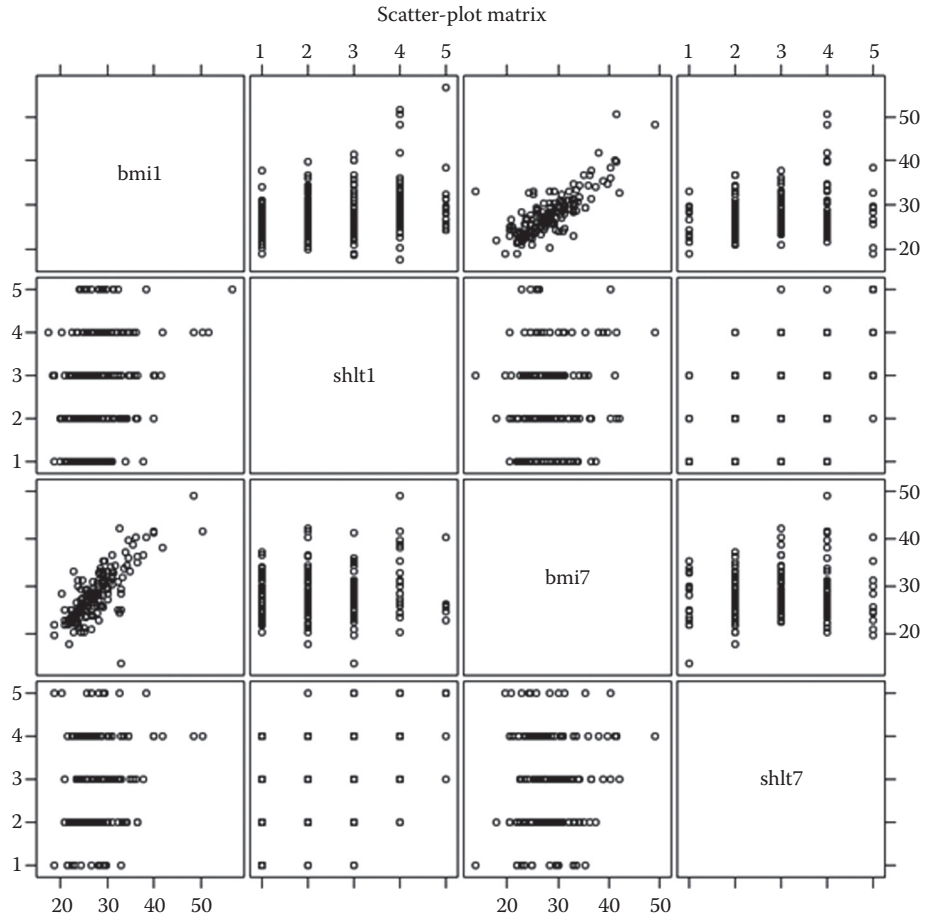


**FIGURE 1.10**
Scatterplot matrix of body-mass index (BMI) and self-rated health (shlt) measures in wave 1 and wave 7 of the Health and Retirement Study.

For example, the plot in the first row and second column is a scatterplot for the relationship between BMI at wave 1 (bmi1) and self-rated health at wave 1 (shlt1). Since SHLT takes only five possible values, the observations are concentrated in five columns of individual circles, each circle corresponding to an individual in the sample. The increasing height of the columns from left to right indicates that participants with higher BMI tend to have worse self-reported health (i.e., higher scores on the self-rated health variable). Also, some outliers in terms of BMI are noted in the upper right corner of this plot.

Similarly, the plot in the first row and third column visualizes the relationship between BMI at wave 1 (bmi1) and BMI at wave 7 (bmi7). As expected BMI measures are strongly positively correlated even though the observations are 14 years apart. The plot in the second row and the fourth column visualizes the relationship between self-rated health during wave 1 (shlt1) and wave 7 (shlt7). Self-rated health measures are also positively correlated but this is harder to see since this variable is ordinal with five levels and the dots representing individuals are on top of each other. We notice that there is nobody in this sample with poor (5) or fair (4) self-rated health at wave 1 who is with excellent health (1)

at wave 7. Participants with poor (5) self-rated health at wave 1 also do not have very good (2) health at wave 7 and participants with excellent health (1) at wave 1 do not have poor health (5) at wave 7.

Note that the plots in the lower left part of this scatterplot matrix are flipped images of the plots in the upper right part of the matrix. For example, the plot in the second row and first column represents the same information concerning the relationship between BMI and SHLT at wave 1 as the plot in the first row and second column but with the vertical and horizontal axes switched. Choosing which of the two plots to focus on is a matter of convenience and depends on the application.

Scatterplot matrices are very useful for visualization of relationships between variables since they allow simultaneous consideration of multiple variables. However, scatterplot matrices that are too large should be avoided since detail may be hard to see. The HRS data set will be used in subsequent chapters for illustration of models for correlated data and also for the effects of different types of missing data on inferences.

### 1.5.5 Serotonin Transport Study in Mother–Infant Pairs

This study evaluates the effects of maternal treatment with an antidepressant (sertraline) for post-partum depression on serotonin transport in breastfeeding mother–infant pairs (Epperson et al., 2001). Treatment with selective serotonin reuptake inhibitors (SSRIs) is associated with significant blockade of serotonin reuptake in patients. Infants of breastfeeding mothers are exposed to sertraline through maternal breast milk. The critical question is whether SSRI exposure is safe for infants. One aspect of this assessment is to test whether sertraline exposure is associated with significant blockade of serotonin reuptake in infants and to compare magnitude of blockade between mothers and infants. The data set consists of 14 mother–infant pairs with serotonin measurements in both mothers and infants before and after exposure to sertraline. Figure 1.11 shows serotonin levels in mothers and infants before and after antidepressant treatment. The plot clearly shows a decrease in mothers after treatment and no change in their infants.

There are two types of correlations in these data: between measurements on mothers and infants within the same pair, and between pre- and post-measurements for each infant or mother. These correlations are shown in a table form in Table 1.1. In this table *mpre* is the variable "serotonin level in mother before the intervention," *mpost* is "serotonin level in mother after the intervention," *cpre* is "serotonin level in child before the intervention," and *cpost* is "serotonin level in child after the intervention." All correlations are positive and most are fairly strong. The only two that are not statistically significantly different from zero are between mothers' and children's measures after treatment, and between mothers' measures after the treatment and children's measures before the treatment. The correlations, as well as the apparent differences in variability of the measurements of mothers before and after the intervention, need to be taken into account for proper statistical analysis. We use these data to illustrate how mixed models can be fitted to analyze clustered correlated data.

### 1.5.6 Meta-Analysis of Clinical Trials in Schizophrenia

This study (Woods et al., 2005) assessed whether the degree of improvement with antipsychotic medication in clinical trials differed depending on control group choice. The meta-analysis evaluated 66 treatment arms from 32 studies of four medications (risperidone, olanzapine, quetiapine, and ziprasidone) for the treatment of schizophrenia symptoms

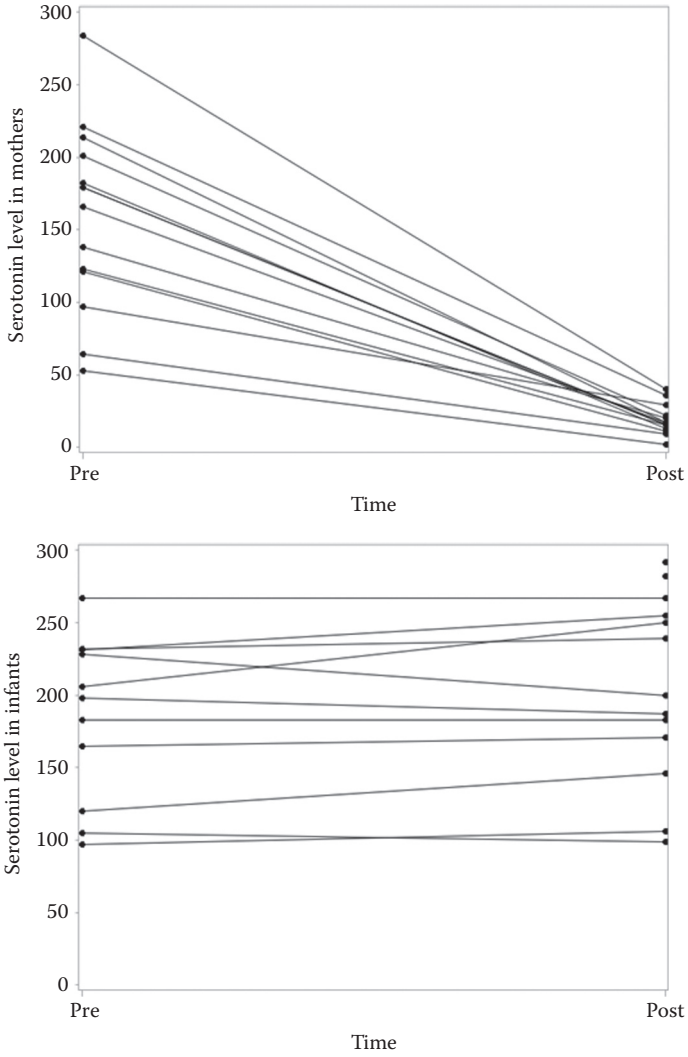


**FIGURE 1.11**
Serotonin levels in mothers (on the left) and their infants (on the right) before and after antidepressant treatment in the serotonin transport study in mother–infant pairs.

**TABLE 1.1**

Correlations between Serotonin Level Measurements on Mothers and Infants before and after Antidepressant Treatment

| | **mpre** | **mpost** | **cpre** | **cpost** |
|---|---|---|---|---|
| **mpre** | 1.00 | 0.70 | 0.71 | 0.56 |
| **mpost** | | 1.00 | 0.48 | 0.34 |
| **cpre** | | | 1.00 | 0.91 |
| **cpost** | | | | 1.00 |

**TABLE 1.2**

Average Change in BPRS Scores from Baseline to Endpoint and Standard Deviations by Dose and Type of Design in Antipsychotic Clinical Trials

| Type of Design Dose | Placebo-Controlled Mean (SD) | Low Dose-Controlled Mean (SD) | Active Control Mean (SD) |
|---|---|---|---|
| Effective dose | 9.0 (2.9) | 11.2 (2.7) | 14.7 (4.0) |
| Intermediate dose | 6.3 (2.7) | 9.8 (3.9) | — |
| Ineffective dose | 2.8 (0.8) | 6.7 (2.2) | — |

with a total of 7264 patients. Average improvements and corresponding standard deviations by dose and type of design are shown in Table 1.2. Based on this table, the largest improvement from baseline occurs in studies with active control at effective medication doses. The average improvement in this type of design is by about 50% more than the average improvement at the same dose level in placebo-controlled trials (14.8 compared to 9.0). This corresponds to more than one standard deviation difference, which is considered a substantial effect. The original study (which applied mixed models to these data) concluded that the degree of improvement with antipsychotic medication in clinical trials differed significantly depending on control group choice. The modeling took into account the correlations between measurements on different treatment arms within the same study that are due to sampling individuals from the same population and the common effect of the environment in which the treatments within the same study were offered. In Chapter 3, we describe how an appropriate model is constructed and show the results from the analysis.

### 1.5.7 Human Laboratory Study of Menthol's Effects on Nicotine Reinforcement in Smokers

Menthol is a common ingredient in e-cigarettes and in other modified tobacco products that may facilitate the development and maintenance of addiction, especially in young adults who increasingly use e-cigarettes. This study (Valentine et al., under review) used a two-level cross-over experimental design to examine whether menthol at different doses, compared to placebo, altered nicotine reinforcement in young adult smokers. Smokers of mentholated and smokers of non-mentholated cigarettes were randomized to receive the three doses of menthol (high dose, low dose, and no menthol) by an e-cigarette in random order and in a double-blind fashion. Each menthol dose was given on a separate test day. On each of these days, all three nicotine doses (saline, 5, 10 μg/kg) were infused in random order. The design is illustrated with the schematic in Figure 1.12. This type of design leads to increased power for testing the main and interactive effects of the two factors (nicotine and menthol) compared to a parallel group design. The main outcome of interest in this study is the rewarding effect of nicotine measured by the Drug Effects Questionnaire. The hypothesis is that concurrent menthol and nicotine administration, as compared to nicotine and control flavor, or saline and control flavor, enhances the rewarding effects of nicotine. We use these data to illustrate how to model correlations within subjects between repeated observations on the same test day and on different test days.

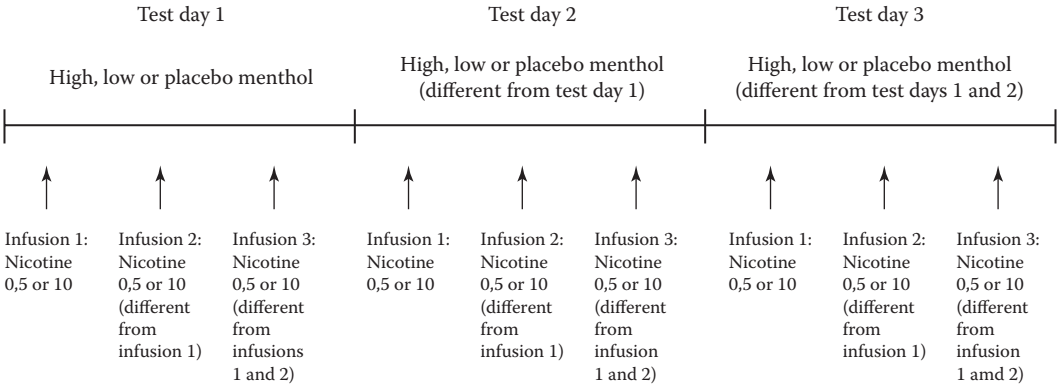


**FIGURE 1.12**
Double cross-over design of the human laboratory study of menthol's effects on nicotine reinforcement in smokers.

### 1.5.8 Functional Magnetic Resonance Imaging (fMRI) Study of Working Memory in Schizophrenia

In this study, 14 patients with schizophrenia and 12 healthy comparison participants were tested on a spatial working memory task with two difficulty levels (Driesen et al., 2008). Brain activation during three distinct phases (encoding, maintenance, and response) was recorded using fMRI. The study assessed phase-specific deficits in cortical function that contribute to cognitive impairments in schizophrenia. The relationship between task performance and brain activation was also assessed. Herein, we focus on averaged activation measures in pre-specified regions of interest (SMFG=superior medial frontal gyrus, MFG=middle frontal gyrus, IFG=inferior frontal gyrus, and VIFG=ventral inferior frontal gyrus) in the pre-frontal cortex as dependent measures.

This is an example of clustered data with the individual being the cluster. There are several sets of correlations within the cluster: correlations between the three different phases, the four different regions, and the two different task difficulty levels. Furthermore, variances in the different regions, phases, and difficulty levels are different. A proper statistical model for the analysis of these data needs to take all these features into consideration. We use this data set to illustrate how mixed models can be used to account for the complicated variance–covariance pattern of the data so that testing of the main hypotheses involving group differences can be accomplished.

Table 1.3 shows the means, variances, and covariances of a subset of the data. We provide descriptive statistics only for the encoding phase (eight repeated measures: one for each region by difficulty level combination), since it is difficult to visually inspect all 24 repeated measures simultaneously. In general, to obtain a preliminary impression of data with many repeated measures, one needs to separate the data into meaningful parts, examine the parts one at a time, and then assess the interrelationships between the different parts. More detailed exploration of these data is undertaken in subsequent chapters. From Table 1.3, we see that means in some regions (e.g., IFG) appear higher than means in other regions (e.g., VIFG) across task difficulty levels. Standard deviations are in general similar, although in regions with lower average they tend to be slightly lower. Also, all correlations within individuals are positive and sizeable, especially correlations within the same region and within the hard-working memory task.

**TABLE 1.3**

Means (M), Standard Deviations (SD) and Correlations (*r*) between Repeated Measures during Hard and Easy Working Memory Tasks in Four Different Brain Regions in the Schizophrenia Data Set

| | **Hard MFG** | **Hard IFG** | **Hard VIFG** | **Hard SMFG** | **Easy MFG** | **Easy IFG** | **Easy VIFG** | **Easy SMFG** |
|---|---|---|---|---|---|---|---|---|
| **Hard MFG** | *M*=0.49 SD=0.23 | *r*=0.57 | *r*=0.59 | r=0.17 | r=0.63 | r=0.44 | r=0.35 | r=0.24 |
| **Hard IFG** | | M=0.42 SD=0.25 | r=0.71 | r=0.62 | r=0.56 | r=0.80 | r=0.52 | r=c0.56 |
| **Hard VIFG** | | | M=0.27 SD=0.16 | r=0.38 | r=0.55 | r=0.63 | r=0.79 | r=0.52 |
| **Hard SMFG** | | | | M=0.44 SD=0.32 | r=0.42 | r=0.62 | r=0.44 | r=0.75 |
| **Easy MFG** | | | | | M=0.35 SD=0.17 | r=0.59 | r=0.57 | r=0.65 |
| **Easy IFG** | | | | | | M=0.32 SD=0.16 | r=0.70 | r=0.70 |
| **Easy VIFG** | | | | | | | M=0.24 SD=0.16 | r=0.70 |
| **Easy SMFG** | | | | | | | | M=0.35 SD=0.25 |

### 1.5.9 Association between Unemployment and Depression

These data are from a study of 254 recently unemployed individuals who were followed for up to 16 months after a job loss (Ginexi et al., 2000). At each of three interviews after the initial job loss, conducted at different times for different individuals, depression symptoms were measured using the Center for Epidemiologic Studies Depression (CES-D) questionnaire, which asks participants to rate the frequency with which they experience each of twenty symptoms of depression. The total CES-D score was calculated as the sum of the answers to the 20 individual questions with a possible range of 0–80. Unemployment status at each interview was also recorded.

Figure 1.13 shows the change in individual depression scores over time for subjects who were re-employed by the end of the study and for subjects who were unemployed at the last available interview. Visually, the range of CES-D scores for those who remained unemployed (or were employed and then laid off again) is wider than for those who were re-employed. Thus, the study hypothesis that depression is associated with higher levels of depressive symptoms appears plausible. Appropriate statistical models for these unbalanced data are presented in Chapter 3, whereas Chapter 8 uses the same data to illustrate the use of time-dependent covariates.

## 1.6 Historical Overview of Approaches for the Analysis of Repeated Measures

The first method for analysis of repeated measures data was the analysis of variance (ANOVA) model with a single random subject effect that dates back to the work of

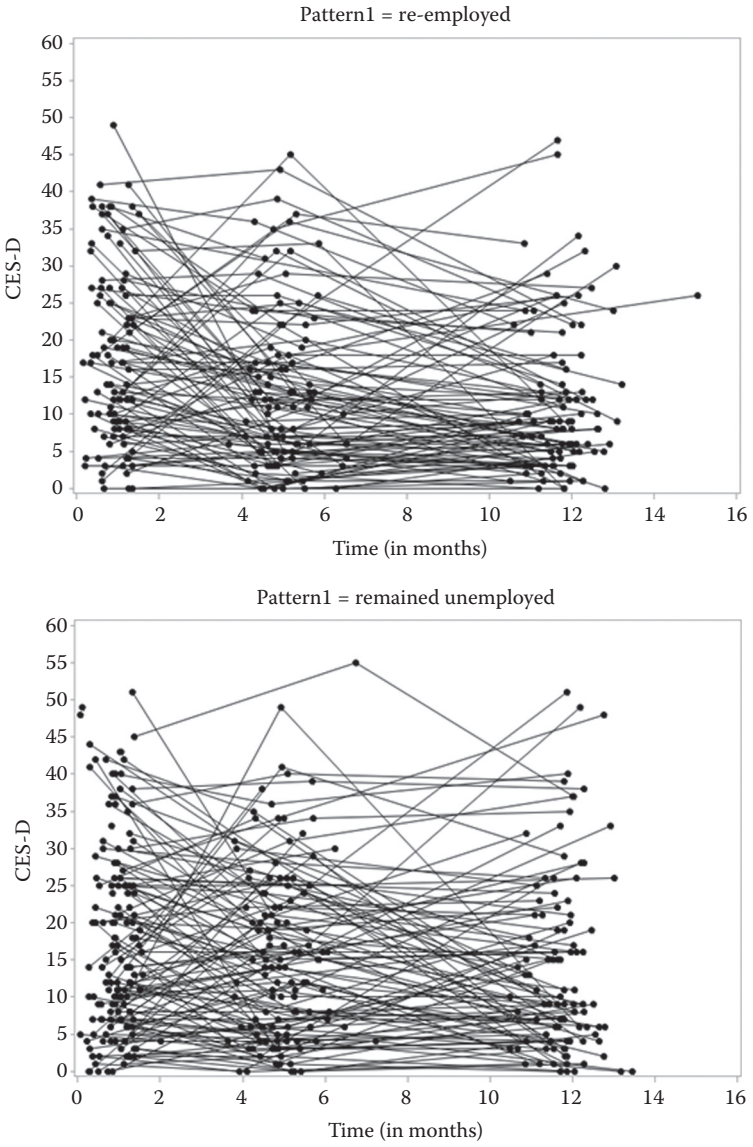


**FIGURE 1.13**
Profile plots of CES-D scores of all subjects by employment pattern in the study of the association between unemployment and depression.

R. A. Fisher (Fisher, 1925). This approach is also known as the *univariate repeated measures ANOVA* (rANOVA) and assumes equal variances of the repeated observations and equal correlations among all repeated observations within individuals or clusters. This assumption is likely to be satisfied in randomized block designs where the observational units within a block are deemed exchangeable, but is unlikely to be satisfied in more complicated clustered or longitudinal designs where variances and correlations can differ within individuals or clusters. Recognizing this problem, Greenhouse and Geisser (1959) and Huynh and Feldt (1976) developed corrections to the statistical tests in univariate repeated

measures ANOVA so that this approach could be used for hypothesis testing when variances and correlations vary within individuals. Despite the correction, this approach is not very flexible for the analysis of longitudinal data and, as incorporated in most statistical packages, does not allow for missing data on an individual. It is also appropriate when the number of occasions per individual or cluster is the same while in many studies with longitudinal and clustered data, the number of observations may differ.

A modification of the ANOVA approach that requires more extensive computations, is *the multivariate repeated measures analysis of variance* (rMANOVA) model. MANOVA was developed for testing between-group differences on multiple distinct response measures simultaneously. The repeated measures situation is different from the situation with distinct response measures in that the same response variable is measured repeatedly over time or within clusters. Nevertheless, in both situations the response observations are correlated and both situations fall within the same framework. An advantage of the rMANOVA approach over the rANOVA approach, is that it allows the variance–covariance structure to be completely general. However, when there are missing data on an individual it excludes all data on this individual from analysis. It is also less powerful than the rANOVA approach when exchangeability is in fact satisfied. Like rANOVA it also requires that the number of repeated occasions per individual is constant.

A special case of rMANOVA analysis is *profile analysis* (Box, 1950) which constitutes a MANOVA analysis of multiple derived variables that are linear combinations of the repeated observations on an individual. This approach is most commonly used with longitudinal data and allows simultaneous testing of mean differences across occasions and trends over time between groups.

For the sake of simplicity of interpretation, often studies with repeated measures data are analyzed based on single summary measures of the observations within individuals. In clustered data studies, one can calculate the means for each cluster and then compare these means between groups using usual ANOVA models. In longitudinal data studies, one can calculate the change from baseline to endpoint and then perform ANOVA comparison on these derived measures. Alternatively, scores at the end of treatment can be compared using ANOVA or ANCOVA (analysis of covariance) with control for baseline scores. This approach has severe limitations, as it ignores a large portion of the data and often requires imputation of missing data. Missing data are very common in longitudinal studies and one of the earliest approaches for dealing with missing data was to impute missing values with baseline values carried forward, last observation carried forward, or mean values calculated based on all individuals. The observation carried forward approaches virtually always lead to biased estimates of effects and, although they were originally proposed as being conservative (i.e., having lower probability of false positive results) in clinical trials and observational studies, they can also be anti-conservative or too liberal (i.e., having higher probability of false positive results). Recently, more sophisticated methods, such as multiple imputation approaches, have been used, which provide valid conclusions under general assumptions about missing data.

The state-of-the-art approaches for analysis of repeatedly measured data nowadays are *mixed-effects models*. They are also known as *random effects models* (Laird and Ware, 1982; Ware, 1985), *random regression models* (Goldstein, 1987), *hierarchical linear models* (Bryk et al., 1987) and *empirical Bayes models* (Casella, 1985). Random effects models assume that individuals deviate randomly from the overall average response. The correlation between repeated observations on the same individual can arise from common random effects or the pattern of variance and correlations can be directly specified. A combination of both approaches is also possible. The specified structures can vary in complexity, from equal

variances at all time points and equal correlations between any two measurements on the same individual (i.e., the structure assumed in rANOVA models), to no restrictions at all (i.e., the structure assumed in rMANOVA models). As an intermediate complexity, one can assume that correlation between observations decreases with increasing time lag. Mixed models are very flexible because they can consider many different variance–covariance structures and select the best-fitting one in the process of selecting the best model. They also use all available data on an individual and give unbiased estimates when data are missing at random. Chapters 3 and 4 present mixed-effects models in detail. Different missing data assumptions are discussed in Chapter 7. A more detailed historical overview at a non-technical level of methods for the analysis of repeated measures data can be found in Gueorguieva and Krystal (2004). Other fairly non-technical books on longitudinal data are Singer and Willett (2003) and Twisk (2013a).

## 1.7 Basic Statistical Terminology and Notation

For simplicity of explanation, we consider that the repeated observations occur within the individual. That is, the individual is the clustering factor. The relevance of the notation to other clustering situations (e.g., individuals nested within families or other larger units) is clear when in the rest of this section "individual" is replaced by "cluster." We also consider the *augmentation treatment for depression* study (Example 1.5.1) in order to illustrate the concepts.

The technical notation is kept to a minimum in this presentation. If, even at that level, it presents a challenge, the book of Altman (1991) can be consulted for a review of basic concepts. For more comprehensive notation and technical details, interested readers are referred to other books: Lindsey (1999), Weiss (2005), and Hedeker and Gibbons (2006). Fitzmaurice et al. (2009) provides a very comprehensive reference for longitudinal data analysis.

### 1.7.1 Response

The response (outcome, dependent) variable is denoted by $Y$. When there is a single observation per individual, the individual responses are denoted as $Y_i$ where $i$ corresponds to the $i$th individual and there are $n$ individuals in the sample. The average of all observations in the sample is the sample mean and is calculated as

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

That is, all observations are summed and the sum is divided by the number of observations.

The sample variance reflects the entire variability in the sample and is calculated as follows:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$

That is, it represents an average of the squares of the deviations of the individual observations from the sample mean. In the denominator, we use $n - 1$ rather than $n$ in order to take into account that we estimate the mean rather than use the true unknown value. The more spread out the observations are, the more variability there is and the larger the calculated variance will be. Since the variance is measured in squared units compared to the response, a more interpretable measure of variability is the standard deviation of the observations, which is measured in the same units as the mean and is obtained as the square root of the variance:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}.$$

The formulae for variance and standard deviation are presented for completeness but are not crucial for understanding the material presented in this book.

In repeated measures data, we have multiple observations per individual and an additional subscript is needed to annotate the responses. The response is $Y_{ij}$, where $i$ corresponds to the $i$th individual and $j$ corresponds to the $j$th observation within the individual. The number of individuals in the sample is usually denoted by $N$ and the number of observations within the $i$th individual is $n_i$. The subscript here is necessary to indicate that the number of observations within the individual does not need to be the same.

The simplest case that we consider in Chapters 2 and 3, is with a quantitative (continuous) response, that is, the response takes values over an interval of possible values. For example, weight and height measurements, rating scales over large intervals, are considered quantitative responses. Chapter 4 presents models for responses that are dichotomous (binary), ordinal, or represent counts. All responses are assumed to be random variables, that is, there is uncertainty in the values that are observed.

In the considered example of the augmentation study in depression, the response is a measure of depression severity (HDRS). We usually start the statistical analysis by calculating means and standard deviations for each group at each time point and visualize the data. In statistical notation, since we have as many means as there are repeated occasions, we denote the sample means and standard deviations as follows:

$$\bar{Y}_j = \frac{1}{N}\sum_{i=1}^{N} Y_{ij}$$

$$s_j = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(Y_{ij} - \bar{Y}_j)^2}$$

Herein, we used $N$ to denote the number of individuals on each of the repeated occasions but the number of individuals does not need to be the same. In longitudinal data especially, participants drop out so the number of individuals decreases over time. Figure 1.5 shows the means and standard deviations by treatment group in the depression example.

### 1.7.2 Predictors

In statistical notation, the predictors are usually denoted by $X$ and can vary between or within individuals. This dependence is often reflected in the subscripts of the predictor. $X_i$ usually denotes the value of a predictor for the $i$th individual and the subscript $i$ means that the covariate has the same value on all repeated occasions within the individual, but in general, has different values for different individuals. In this case, the predictor is said to vary between individuals but not within individuals. In the considered example, the treatment group is a predictor that varies between individuals but not within individuals.

Similarly, $X_{ij}$ denotes the value of a predictor for the $j$th observation on the $i$th individual. The additional subscript $j$ is used in order to distinguish different values of this predictor on different observation occasions within individuals. In this case, the predictor is said to vary within individuals and it can also vary between individuals. In the depression example, time is a within-subject predictor as it varies within each individual (i.e., each individual has observations at several different time points). The time points may or may not be the same for different individuals. Other subject characteristics can vary over time, such as blood pressure or weight, or subjects can switch treatments over time and hence treatment can also vary within individuals. This is the case in cross-over studies.

When there are multiple predictors, a third subscript $k$ can be used to denote the $k$th predictor. Time and group are almost always present as predictors in repeated measures studies with longitudinal data. Additional variables that may be affecting the response are usually referred to as covariates and they are also considered predictors. For example, history of depression and concurrent medication use may be additional covariates in the depression study.

Predictors are assumed to be exactly observed, that is, there is no uncertainty in the values of the predictors and they are considered fixed, not random. Time and group are usually exactly observed but other predictors may be measured with error and may need to be considered as random variables themselves. For example, self-reported medication compliance may be imprecisely ascertained. When predictors are also assumed to be random variables, estimation and inference are more complicated. Interested readers are referred to Fuller (1987) since this situation is not considered in this book.

### 1.7.3 Linear Model

The statistical relationship between the predictors and the response in the population and its change by occasion or time can be described using a statistical model. This represents our theoretical understanding and assumptions about the relationship between the predictors and the response and may or may not correspond to reality. The linear model assumes that the association is linear in the coefficients (betas in the formula below) and that the effects of the predictors are additive (i.e., they add onto one another rather than multiply or act together in another fashion). This can be expressed using the following formula:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \ldots + \beta_p X_{ijp} + \varepsilon_{ij},$$

Where some of the predictors vary between individuals, and some vary within individuals, the beta coefficients are unknown parameters that can be estimated from the data and reflect the direction and magnitude of the association between the predictors and the

response, and the epsilons denote the errors that describe the uncertainty or residual variability of the measurements, apart from what the predictors explain.

In the depression example, the equation describes how each observation varies depending on the predictors and can be written as

$$HDRS_{ij} = \beta_0 + \beta_1 Group_i + \beta_2 Time_j + \beta_3 Group_i \times Time_j + \varepsilon_{ij},$$

where:

$Group_i$ takes the value of 1 if the $i$th individual belongs to the augmentation group and 0 if this individual belongs to the control group
$Time_j$ is the week (coded 0 through 6) when the $j$th observation is taken
$Group_i$*$Time_j$ denotes the interaction between $Group_i$ and $Time_j$ (i.e., is the product of $Group_i$ and $Time_j$)

The interaction reflects how the responses for the different groups differ from one another over time. Each individual's response can be described by substituting the appropriate values for $Group_i$ and $Time_j$ in the equation. For example, the baseline (i.e., Time=0) response for an arbitrary chosen patient $i$ from the augmentation group is described as $HDRS_{i0}=\beta_0+\beta_1+\varepsilon_{i0}$ while the response at week 6 for another arbitrary chosen patient $l$ from the control group is described as $HDRS_{l6}=\beta_0+\beta_2\,6+\varepsilon_{l6}$. The errors reflect the expected deviations for particular individuals and particular occasions from the average response of all patients in the corresponding hypothetical population measured on the corresponding occasion.

### 1.7.4 Average (Mean) Response

The average response for a particular combination of values of the predictors is described according to the linear function above and is

$$EY_{ij} = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \ldots + \beta_p X_{ijp}$$

Here, $E$ stands for expectation and this formula describes how the expected (average) response in the population varies depending on the predictors. The predictors may appear by themselves in the formula or two (or even more) predictors can multiply each other (for example the $kth$ predictor can be a product of some of the other predictors, e.g., $X_{ijk}=X_{ij1}X_{ij2}$). When predictors appear by themselves, they represent main effects. The beta coefficients are then interpreted as the differences in mean response that correspond to a unit change in the predictor (for continuous predictors) and the difference in mean response that corresponds to comparing a particular level of a categorical predictor to a reference level of this predictor (e.g., experimental to control group, or later time point to an earlier time point). When there are interactions between the predictors (i.e., when the predictors multiply each other, e.g., $X_{ijk}=X_{ij1}X_{ij2}$ in the formula above), the interactions need to be interpreted first. Interactions may involve two or more predictors and become increasingly complex to explain, especially in designs with multiple factors. We consider different interactions and their interpretations in subsequent chapters. Herein, a simple situation is considered in the context of the depression example.

In the depression example, the average response can be described as

$$E\{HDRS_{ij}\} = \beta_0 + \beta_1 Group_i + \beta_2 Time_j + \beta_3 Group_i \times Time_j.$$

Let us assume for a moment that $\beta_3$ is equal to 0, which means that there is no interaction between group and time, that is, the change over time in the two groups has the same form (i.e., average responses in the two groups can be described by parallel lines). The coefficient $\beta_1$ is interpreted as the difference in mean HDRS scores between the two groups averaged over the entire time period. Using the specified coding above for a group (1 for the augmentation group and 0 for the control group), positive values mean that the augmentation group has higher scores on average, while negative values mean that the control group has higher scores on average. The coefficient $\beta_2$ is interpreted as the change in average HDRS score per week (i.e., one unit change in time). To estimate how much the HDRS scores change on average over the entire study period, we need to multiply $\beta_2$ by the study duration in weeks (i.e., $\beta_2$ times 6). Note that this model assumes that the rate of change stays constant over the study period, that is, the change over time is described by a straight line. This is often an untenable assumption although, in some situations, it may be a convenient approximation.

If $\beta_3$ is not equal to 0, then there is an interaction between group and time. In this case, slopes of average change in the two groups over time are different and the average between-group differences change depending on which time points we consider. At time 0, the difference in average response between the two groups is described by $\beta_1$ but at time 6, for example, the difference in average response between the two groups is described by $\beta_1+\beta_3.6$. Depending on the signs of $\beta_1$ and $\beta_3$ the difference may be smaller or larger. The $\beta_3$ coefficient is interpreted as the difference in slopes (linear rates of change) between the two groups over time. When change over time is described by a more complicated function, rather than linear change, interpreting between-group differences becomes more challenging.

In the general linear model, a unique linear combination of the predictors corresponds to each observation time point and group. For the control group, the average response at baseline (time 0) is $HAMD_{i0}=\beta_0$, while for the experimental group, the average response at week 6 is

$$E\{HDRS_{i6}\} = \beta_0 + \beta_1 + \beta_2.6 + \beta_3.6.$$

In repeated measures data, estimation of the relationship between predictors and the mean response is usually of primary interest. In order to assess this relationship, all beta parameters need to be estimated. The deviations from the mean response need to be taken into account but are often of secondary interest.

### 1.7.5 Residual Variability

To perform statistical inference (i.e., construct confidence intervals or test hypotheses about the beta parameters), certain assumptions need to be imposed on the errors in the statistical model formulation above. Usually, when the response is continuous, the errors are assumed to be normally distributed with zero mean. Note that the distribution of the errors determines the distribution of the response in the sample when there are no other random effects. Thus, if the errors are normally distributed then the response is also

normally distributed. A histogram of the responses on each occasion and within a group indicates whether the data are indeed approximately bell-shaped distributed and hence whether the normal distribution is appropriate.

The zero mean assumption is reasonable if we have included all important predictors of the response in the linear predictor and have specified the nature of the relationship correctly. That is, we have not omitted predictors that substantially affect the response and have used the appropriate form of each predictor. A classic example of miss-specified form is if the relationship between time and the response when plotted seems to be curvilinear but we include only the linear effect of time in the model. In this case, on some occasions the errors will have means that are larger than 0 and on some other occasions they will have means that are smaller than 0. Such a discrepancy will need to be corrected in order to reach justifiable conclusions for the relationship between predictors and response.

In classical regression and analysis of variance models, where each individual contributes a single observation, the errors are assumed to be independent of one another and to have equal variances. However, in repeated measures situations, error variances often vary by occasion and errors are correlated within individuals. In the depression example, and in similar longitudinal studies, it is likely that variances increase over time because individuals are usually selected to satisfy certain conditions for study entry, and then, some individuals show significant improvement in their response, some show no change, and some deteriorate. Thus, models that assume equal variances may not be appropriate.

Furthermore, the errors $\varepsilon_{ij}$ corresponding to different individuals are assumed to be independent while different $\varepsilon_{ij}$'s corresponding to the same individual are assumed to be related. This is reasonable, as we expect repeated observations on the same individual on different occasions to deviate in a systematic way from the average for similar individuals. Thus, the errors for the same individual are more likely to be in the same direction (i.e., mostly positive or mostly negative) and their magnitudes are likely to be related. To assess whether the data support such assumptions, correlations of repeated observations on different occasions can be calculated and examined, either in table form or in figures. The *sample correlation* between repeated measures on occasion $k$ and $l$ within the individual is calculated as follows:

$$r_{kl} = \frac{\sum_{i=1}^{n} (Y_{ik} - \bar{Y}_k)(Y_{il} - \bar{Y}_l)}{\sqrt{\sum_{i=1}^{n} (Y_{ik} - \bar{Y}_k)^2} \sqrt{\sum_{i=1}^{n} (Y_{il} - \bar{Y}_l)^2}}$$

Correlations measure the degree of linear dependence between variables. Correlations vary between –1 and 1 with 0 corresponding to no linear relationship, –1 corresponding to a perfect negative relationship, and 1 corresponding to a perfect positive relationship. Figure 1.14 shows these three situations, as well as an example of a strong positive correlation, weak negative correlation, and curvilinear relationship, where the correlation is close to 0 but the two variables are related. When there are repeated measures within individuals some degree of linear dependence is expected and the dependence is usually positive.

The statistical notation for the error distributions in models with repeated observations is $\varepsilon_{ij} \sim N(0, \sigma_j^2)$, where $\sigma_j^2$ may be different for different occasions within individuals, $N$ denotes normal distribution and errors are randomly spread out around 0. The errors for measurements on occasion $k$ and $l$ within a randomly chosen individual in the population are assumed to be correlated, that is $Corr(\varepsilon_{ij}, \varepsilon_{il}) = \rho_{jl}$ and this correlation is most often

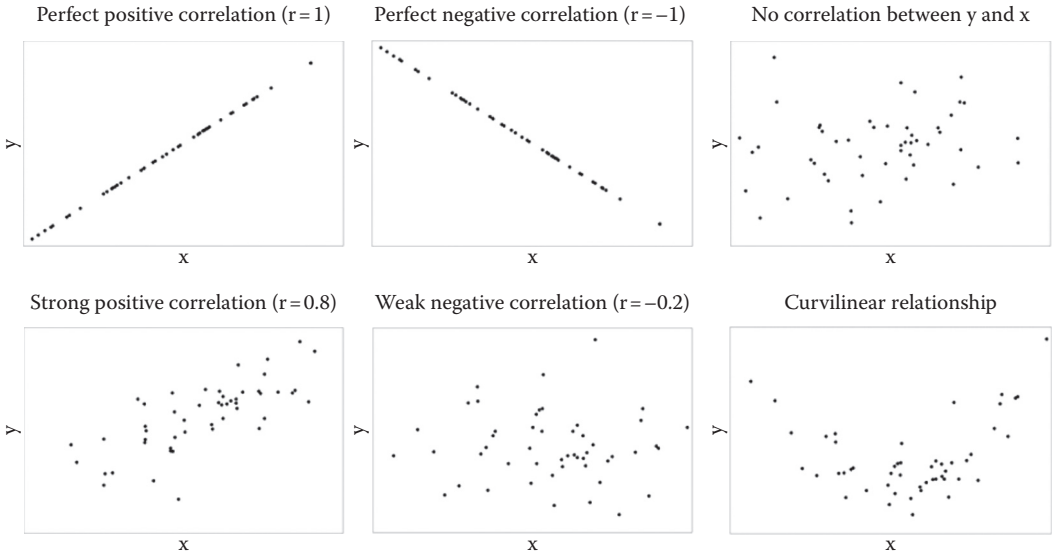


**FIGURE 1.14**
Examples of different corrections between two variables.

positive. With longitudinal data, the correlation is generally stronger for observations that are closer to each other in time and decreases with increasing time lag. Examining sample correlations between any two repeated measurements, either in table form or in a graph, provides information about appropriate correlation structures in the population. Plots to examine correlations in longitudinal data are described in Weiss (2005) and in Dawson et al. (1997).

In the depression example, if an individual starts with a HDRS score that is significantly higher than the average HDRS score in the sample, this individual's HDRS scores are more likely to stay above the average on the next few occasions, compared to the scores of an individual who starts with below average HDRS scores (see Figure 1.4). With increasing time lag, however, the probability of such systematic deviation becomes weaker and weaker.

### 1.7.6 Estimation

Estimates of the beta parameters are obtained using some statistical method so that the residual variability in the data is minimized. The method of *maximum likelihood* is the most commonly used approach and it finds the values of the beta parameters that maximize the likelihood that we observe the data in the sample, given our assumptions about how the data were generated, as reflected in the model formulation. Substituting the estimated beta values in the linear model, instead of the true unknown beta values, gives an estimate of the average response from the data. The obtained estimates are *unbiased* in large samples (i.e., they do not deviate in a systematic way from the true values of the parameters) and they are *efficient* (i.e., the uncertainty of these estimates is as small as possible). Uncertainty is measured by the standard errors of the estimates of the beta parameters and the standard errors are obtained in the estimation process. The parameter estimates are also approximately normally distributed which makes construction of confidence intervals and hypothesis tests straightforward.

### 1.7.7 Statistical Inference

Traditionally, the type of statistical inference of most interest in the subject-matter literature, has been testing whether one or more of the beta coefficients are zero, which corresponds to testing whether there is any effect of the corresponding predictor(s) on the response when keeping the values of the other predictors in the model constant (commonly referred to as controlling for the effects of the other predictors). In the simplest case of *testing a hypothesis* concerning a single beta coefficient, the null hypothesis is that the coefficient is zero. The alternative hypothesis is that this coefficient is different from zero if a two-sided test is performed, or that it is greater (or smaller) than zero in the corresponding one-sided tests. One-sided tests are rarely used because if the direction of the relationship between the predictor and the response is opposite to the hypothesized one, a one-sided test will fail to find a significant effect. For example, if an experimental treatment is compared to a standard treatment and the alternative hypothesis is that the experimental treatment is better than the standard treatment, there is no possibility to conclude based on a one-sided hypothesis test that the experimental treatment is worse than the standard treatment.

*Test statistic*, for testing whether a single beta coefficient is zero, is usually just the estimate of this coefficient over its estimated standard error. Large absolute values of this ratio indicate that it is unlikely that the beta parameter is zero. In such a case, the interpretation is that the predictor is significantly associated with the response.

Most statistical testing is based on the calculated *p-value*, which is the probability that the test statistic is at least as extreme as observed if there is no relationship between the predictor and the response. Note that if there is no true relationship, we would expect the parameter estimate to be close to zero and the test statistic to be small. If the p-value is smaller than a pre-specified cut off called significance level $\alpha$ (0.05% or 5% is most commonly used), then the conclusion is that it is unlikely that there is no relationship between the predictor and the response, and the relationship is declared to be statistically significant.

Two types of error can occur in this inference. When there is no relationship (of the form specified in the model) between the predictor and the response, but the hypothesis test concludes that the relationship is statistically significant, a *type I error* has occurred. This is a false positive result and by selecting a low significance level, we guard against this type of error. At 5% significance level, we would expect 5% of tests, when there is no significant relationship between the predictor and the response, to result in this type of error.

The other error occurs if there is a relationship between the predictor and the response but the hypothesis test results in failure to reject the null hypothesis of no relationship and the conclusion is that the relationship is not statistically significant. This type of error is called *type II error* and is a false negative result. How large the probability of this error is depends on the magnitude of the beta coefficient and the population variability. When the beta coefficient is small and/or the variability is large, there is a higher chance to commit this type of error. This type of error is also directly related to the power of the statistical test.

*Power* is the probability to reject the null hypothesis (i.e., declare that a statistically significant relationship exists) when the alternative is true (i.e., there is a relationship between the predictor and the response). Power changes with changing values of the beta coefficients and changing variability. It increases with increasing beta values and decreases with increasing variability. Issues of significance level and power of tests are considered in more detail in Chapter 11.

Hypothesis tests provide clear conclusions regarding the significance of the relationship between predictors and response. However, they are heavily dependent on sample size and do not provide estimates of the magnitude of the effects and the uncertainty in the estimates. *Confidence intervals* contain more information than hypothesis tests as they give a range for the magnitude of the effect of the predictor on the response with a certain level of confidence. Most commonly, 95% confidence intervals are constructed as the corresponding beta estimate plus/minus 1.96 times the standard error of this estimate. The confidence level 95% means that 95% of the time the true parameter falls within the limits of the confidence interval and is interpreted as the level of confidence that we have that we have captured the true underlying parameter in the confidence interval. Confidence intervals can also be used to evaluate whether the corresponding beta coefficient is statistically significantly different from zero (or any other value) or not, i.e., they can be used to perform the corresponding hypothesis test. If the confidence interval contains zero then the beta coefficient is declared not to be significantly different from zero and if the confidence interval does not contain zero then the beta coefficient is declared to be significantly different from zero.

In recent years, more emphasis is placed on reporting confidence intervals rather than p-values and for a very good reason. Rather than giving a yes/no answer to a sometimes contrived or overly simplified question as hypothesis tests do, they provide an estimate of the magnitude of an effect with an associated level of confidence. Thus, the reader or independent researcher can make their own judgment call on whether a particular result is clinically or practically meaningful or not. As a simple example, consider a hypothetical situation with two treatments (A and B) for depression. A very large clinical trial finds that treatment A improves a measure of depression severity on average by 0.1 standard deviations more than treatment B over a period of 8 weeks with a 95% confidence interval between 0.05 and 0.15. While this corresponds to a statistically significant result because the confidence interval does not contain 0, most doctors would probably not consider such a change as clinically meaningful and they would decide which treatment to use based on other considerations than differences in clinical efficacy.

Hypothesis tests are still useful in situations when some guidance is needed as to which effects to estimate, as in models with multiple possible interactions or in multiple comparison problems. But even in these cases, confidence intervals are still recommended as post-hoc analyses in order to obtain estimates of the magnitudes of effects. The discussion of the choice between hypothesis tests and confidence intervals and the joint use is continued in further chapters of the book.

### 1.7.8 Checking Model Assumptions

The errors in the linear model formulation are not directly observable. However, when estimates of the beta coefficients are obtained, these allow estimation of the errors by taking the difference between the individual responses and the predicted mean: $Y_{ij} - \hat{Y}_{ij}$. These quantities are called *residuals* and they give information about the fit of the model to the data. Since they are estimates of the unknown errors, assessing their distribution and variability can help assess whether the corresponding assumptions about the errors are approximately satisfied. Residual plots can be used to assess whether the assumptions of linearity, normality, and variance pattern are appropriate. If assumptions about the errors are not satisfied then remedial measures must be taken by either considering a more general model, adding covariates, transforming the data, or using statistical methods that make fewer assumptions about the data, such as non-parametric methods. A good

reference for checking model assumptions and remedial measures for linear models is Kutner et al. (2005). Model diagnostics for models for repeated measures data are briefly considered in Chapters 3 and 4 where further references are also provided.

### 1.7.9 Model Fit and Model Selection

Many different models can be fit to any particular data set. Statistical criteria can be used in order to select the best-fitting model among a set of different possible models fitted to the same data set. Perhaps the most commonly used are different versions of information criteria such as the Akaike Information Criterion (AIC) and the Schwartz-Bayesian Information Criterion (BIC). We consider these in more detail in Chapters 3 and 4.

## 1.8 Summary

In this chapter, we described what repeated measures are, introduced different types of studies with repeated measures, discussed advantages of such studies, provided a brief historical overview of statistical methods for clustered and longitudinal data, reviewed some basic statistical terminology and notation, and introduced several data examples that are further considered in subsequent chapters. We focused on issues of describing mean response across repeated measures and accounting for variability and interdependence in the data. In Chapter 2, we consider the traditional methods for repeated measures analysis in more detail, while the rest of the book focuses on the state-of-the-art methods for such analysis and on different aspects of the analysis and design of studies with repeated measures.