Chapter 1

Introduction

1.1 Introduction

Longitudinal studies are defined as studies in which the outcome variable is repeatedly measured; i.e. the outcome variable is measured in the same subject on several occasions. In longitudinal studies, the observations of a subject over time are not independent of each other, and therefore it is necessary to apply special statistical methods, which take into account the fact that the repeated observations within a subject are correlated. The definition of longitudinal studies (used in this book) implicates that statistical methods like survival analyses are beyond the scope of this book. Those methods basically are not longitudinal data analysing methods because (in general) the outcome variable is an irreversible endpoint and therefore strictly speaking only measured at one occasion. After the occurrence of an event no more observations are carried out on that particular subject.

Why are longitudinal studies so popular these days? One of the reasons for this popularity is that there is a general belief that with longitudinal studies the problem of causality can be solved. This is, however, a typical misunderstanding and is only partly true. Table 1.1 shows the most important criteria for causality, which can be found in every epidemiological textbook. Only one of them is specific for a longitudinal study: the rule of temporality. There has to be a time-lag between the outcome variable (effect) and the covariate (cause); in time the cause has to precede the effect. The question of whether or not causality exists can only be (partly) answered in specific longitudinal studies (e.g. randomized controlled trials) and certainly not in all longitudinal studies. In Chapter 6 the problem of causality in observational longitudinal studies will be discussed, while Chapter 10 deals with the analysis of data from randomised controlled trials.

Table 1.1 Criteria for causality

Strength of the relationship
Consistency in different populations and under different circumstances
Specificity (cause leads to a single effect)
Temporality (cause precedes effect in time)
Biological gradient (dose–response relationship)
Biological plausibility
Experimental evidence

What then is the advantage of performing a longitudinal study? A longitudinal study is expensive, time consuming, and the data are difficult to analyse. If there are no advantages over cross-sectional studies why bother? The main advantage of a longitudinal study compared to a cross-sectional study is that the individual development of a certain outcome variable over time can be studied. In addition to this, the individual development of an outcome variable can be related to the individual development of particular covariates.

1.2 Study Design

Medical studies can be roughly divided into observational and intervention studies (see Figure 1.1). Observational studies can be further divided into case-control studies and cohort studies. Case-control studies are never longitudinal, in the way that longitudinal studies were defined in Section 1.1. The outcome variable (a dichotomous outcome variable distinguishing case from control) is measured only once. Furthermore, case-control studies are always retrospective in design. The outcome variable is observed at a certain time-point, and the covariates are measured retrospectively.

In general, observational cohort studies can be divided into prospective, retrospective and cross-sectional cohort studies. A prospective cohort study is the only cohort study that can be characterized as a longitudinal study. Prospective cohort

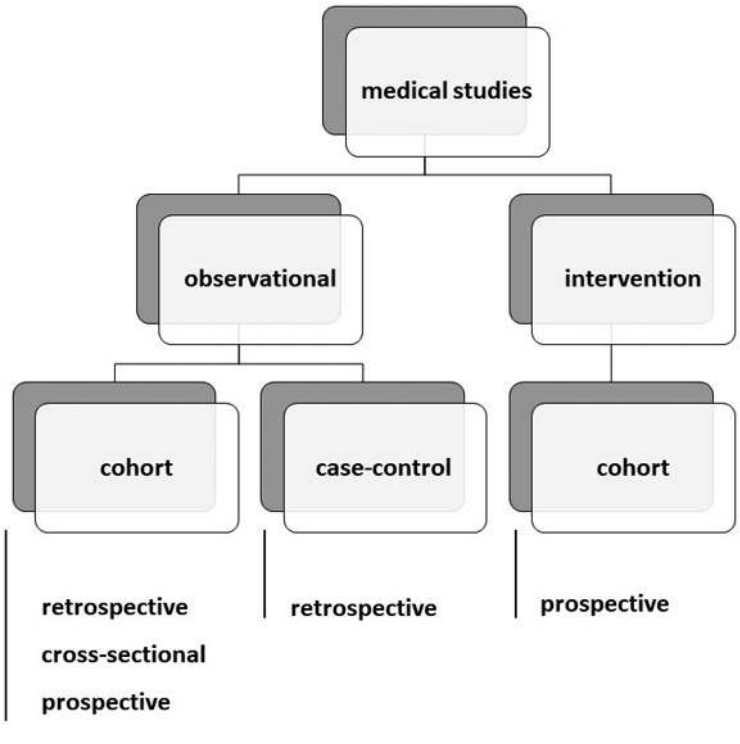

Figure 1.1 Schematic illustration of different medical study designs.

studies are usually designed to analyse the longitudinal development of a certain outcome over time. It is argued that this longitudinal development concerns growth processes. However, in studies investigating the elderly, the process of deterioration is the focus of the study, whereas in other developmental processes, growth and deterioration can alternately follow each other. Moreover, in many studies one is interested not only in the actual growth or deterioration over time, but also in the longitudinal relationship between an outcome and several covariates. Intervention studies, e.g. randomised controlled trials, are by definition prospective, i.e. longitudinal. The outcome variable is measured at least twice (the classical pre-test, post-test design), and other intermediate measures are usually also added to the research design in order to evaluate short-term and long-term effects of the particular intervention.

1.2.1 Observational Longitudinal Studies

In observational longitudinal studies investigating individual development, each measurement taken on a subject at a particular time-point is influenced by three factors: (1) age (time from date of birth to date of measurement), (2) period (time or moment at which the measurement is taken), and (3) birth cohort (group of subjects born in the same year). When studying individual development, one is mainly interested in the age effect. One of the problems of most of the designs used in longitudinal studies of development is that the main age effect cannot be distinguished from the period and cohort effects.

There is an extensive amount of literature describing age, period and cohort effects (e.g. Lebowitz, 1996; Robertson et al., 1999; Holford et al., 2005). However, most of the literature deals with classical age–period–cohort models, which are used to describe and analyse trends in (disease-specific) morbidity and mortality (e.g. Kupper et al., 1985; Mayer and Huinink, 1990; Holford, 1992; McNally et al., 1997; Robertson and Boyle, 1998; Rosenberg and Anderson, 2010). In this book, the main interests are the individual development over time, and the longitudinal relationship between an outcome and several covariates. In this respect, period effects or time of measurement effects are often related to a change in measurement method over time, or to specific environmental conditions at a particular time of measurement. A hypothetical example is given in Figure 1.2. This figure shows the

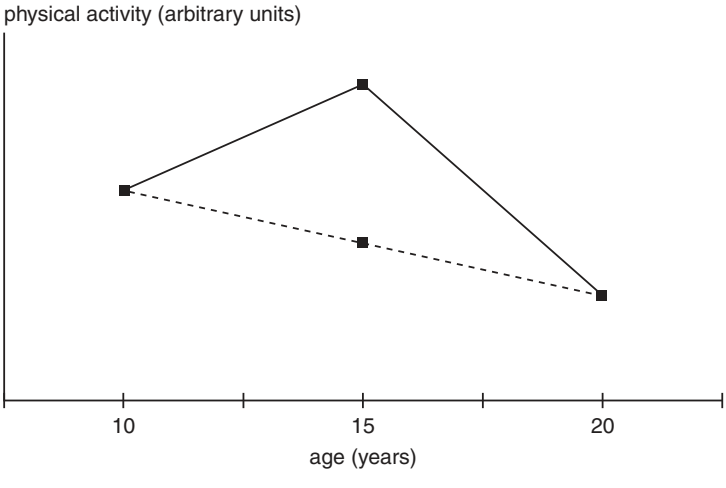

Figure 1.2 Illustration of a possible time of measurement effect (dotted line: real age trend, solid line: observed age trend).

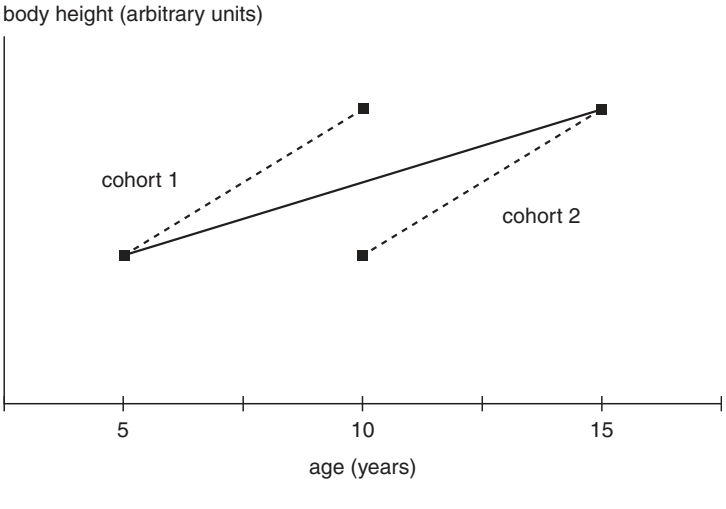

Figure 1.3 Illustration of a possible cohort effect (dotted line: cohort specific, solid line: observed).

longitudinal development of physical activity with age. Physical activity patterns were measured with a five-year interval, and were measured during the summer in order to minimise seasonal influences. The first measurement was taken during a summer with normal weather conditions. During the summer when the second measurement was taken, the weather conditions were extremely good, resulting in activity levels that were very high. At the time of the third measurement, the weather conditions were comparable to the weather conditions at the first measurement, and therefore the physical activity levels were much lower than those recorded at the second measurement. When all the results are presented in a graph, it is obvious that the observed age trend is highly biased by the period effect at the second measurement.

One of the most striking examples of a cohort effect is the development of body height with age. There is an increase in body height with age, but this increase is highly influenced by the increase in height of the birth cohort. This phenomenon is illustrated in Figure 1.3. In this hypothetical study, two repeated measurements were carried out in two different cohorts. The purpose of the study was to detect the age trend in body height. The first cohort had an initial age of five years; the second cohort had an initial age of 10 years. At the age of five, only the first cohort was measured, at the age of 10, both cohorts were measured, and at the age of 15 only the second cohort was measured. The body height obtained at the age of 10 is the average value of the two cohorts. Combining all measurements in order to detect an age trend will lead to a much flatter age trend than the age trends observed in both cohorts separately.

Both cohort and period effects can have an influence on the interpretation of results of longitudinal

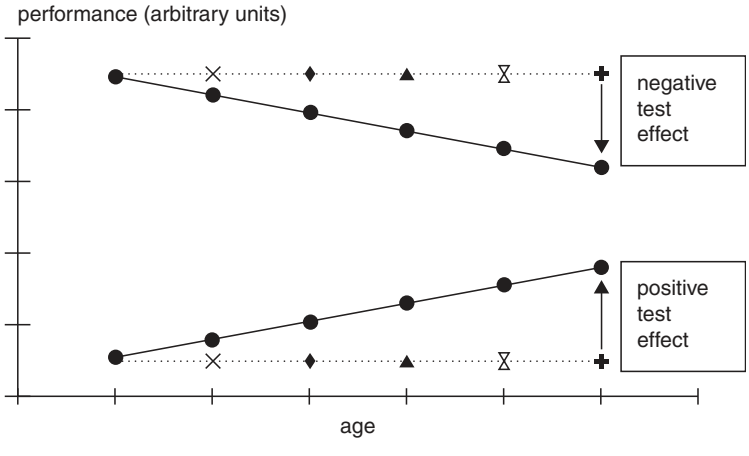

Figure 1.4 Test or learning effects; comparison of repeated measurements of the same subjects with non-repeated measurements in comparable subjects (different symbols indicate different subjects, dotted line: cross-sectional, solid line: longitudinal).

studies. An additional problem is that it is very difficult to disentangle the two types of effects. They can easily occur together. Logical considerations regarding the type of variable of interest can give some insight into the plausibility of either a cohort or a period effect. When there are (confounding) cohort or period effects in a longitudinal study, one should be careful with the interpretation of age-related results.

In studies investigating development, in which repeated measurements of the same subjects are performed, cohort and period effects are not the only possible confounding effects. The individual measurements can also be influenced by a changing attitude towards the measurement itself, a so-called test or learning effect. This test or learning effect, which is illustrated in Figure 1.4, can be either positive or negative.

One of the most striking examples of a positive test effect is the measurement of memory in older subjects. It is assumed that with increasing age, memory decreases. However, even when the time interval between subsequent measurements is as long as three years, an increase in memory performance with increasing age can be observed: an increase which is totally due to a learning effect (Dik et al., 2001).

1.3 General Approach

The general approach to explain the statistical methods covered in this book will be: the research question as basis for analysis. Although it may seem quite obvious, it is important to realise that a statistical analysis has to be carried out in order to obtain an answer to a particular research question. The starting point of each analysis will be a research question, and throughout the book many research questions will be addressed. The book is further divided into chapters regarding the characteristics of the outcome variable. Each chapter contains extensive examples, accompanied by computer output, in which special attention will be paid to the interpretation of the results of the statistical analyses.

1.4 Prior Knowledge

Although an attempt has been made to keep the (complicated) statistical methods as understandable as possible, and although the basis of the explanations will be the underlying research question, it will be assumed that the reader has some prior knowledge about (simple) cross-sectional statistical methods such as linear regression analysis, logistic regression analysis, and analysis of variance.

1.5 Example

In general, the examples used throughout this book are taken from the same longitudinal dataset. The dataset is taken from the Amsterdam Growth and Health Longitudinal Study, an observational longitudinal study investigating the longitudinal relation between lifestyle and health in adolescence and young adulthood (Kemper, 1995).

This dataset consists of a continuous outcome variable (serum cholesterol in mmol/liter) which is measured six times on the same subjects. In the examples, in general, two covariates are used. Body fatness, which is operationalised by the sum of the thickness of four skinfolds, is continuous

Table 1.2 Descriptive information[1] for the data used in most of the examples

Time-point	Cholesterol (mmol/liter)	Sum of skinfolds (cm)	Sex
1	4.43 (0.67)	3.26 (1.24)	69/78
2	4.32 (0.67)	3.36 (1.34)	69/78
3	4.27 (0.71)	3.57 (1.46)	69/78
4	4.17 (0.70)	3.76 (1.50)	69/78
5	4.67 (0.78)	4.35 (1.68)	69/78
6	5.12 (0.92)	4.16 (1.61)	69/78

[1] For cholesterol and sum of skinfolds, mean and between brackets standard deviation are given, while for sex the numbers (males/females) are given.

Table 1.3 Illustration of two different data structures

Broad data structure							
Id	Y_{t1}	Y_{t2}	Y_{t3}	$X1_{t1}$	$X1_{t2}$	$X1_{t3}$	*X2*
1	3	5	8	10	14	16	1
2	2	4	9	13	15	15	1
3	4	6	7	12	13	16	0
Long data structure							
Id		*Y*		*X1*	*X2*		**Time**
1		3		10	1		1
1		5		14	1		2
1		8		16	1		3
2		2		13	1		1
2		4		15	1		2
2		9		15	1		3
3		4		12	0		1
3		6		13	0		2
3		7		16	0		3

and also measured six times on the same subjects and sex, which is dichotomous and which is measured only once and has the same value at all six repeated measurements.

In the chapter dealing with dichotomous outcome variables (i.e. Chapter 7), the continuous outcome variable cholesterol is dichotomised (i.e. the highest tertile versus the other two tertiles) and in the chapter dealing with categorical outcome variables (i.e. Chapter 8), the continuous outcome variable cholesterol is divided into three equal groups based on tertiles. Table 1.2 shows descriptive information for the variables used in the example.

All the example datasets used throughout the book are available on request by jwr.twisk@amsterdamumc.nl.

1.6 Software

Most of the example analyses performed in this book are performed in STATA (version 17).

However, SPSS (version 26) is also used for some of the example analyses. STATA is chosen as the main software package for the longitudinal data analyses, because almost all statistical analyses can be performed in STATA and because of the simplicity of the syntax and the output. In Chapter 13, an overview (and comparison) will be given of other software packages such as R (version 4.0.3) and SAS (version 8). In all these packages, algorithms to perform longitudinal data analysis are implemented in the main software. Both syntax and output will accompany the overview of the different software packages.

1.7 Data Structure

It is important to realise that different statistical software packages need different data structures in order to perform longitudinal data analyses. In this respect a distinction must be made between a long data structure and a broad data structure. In a long data structure, each subject has as many data records as there are measurements over time, while in a broad data structure each subject has one data record, irrespective of the number of measurements over time (see Table 1.3).

1.8 What is New in the Third Edition?

In addition to changes made throughout the book to update the material and to make some of the explanations clearer, some new chapters have been added. In the new Chapter 5, hybrid models are introduced. Hybrid models are used to disentangle the between- and within-subjects interpretation of the regression coefficient obtained from a longitudinal data analysis. The new Chapter 6 contains a discussion regarding causality in observational longitudinal studies, while in the new Chapter 9, the analysis of outcome variables with floor or ceiling effects is discussed. In Chapter 10, 'Analysis of Longitudinal Intervention Studies', three new sections have been added: one section about an alternative repeated measures analysis to take into account regression to the mean; one section about the analysis of data from a stepped wedge trial design; and one section about the difference in difference method.

Chapter 2

Continuous Outcome Variables

2.1 Two Measurements

The simplest form of longitudinal study is that in which a continuous outcome variable is measured twice in time. With this simple longitudinal design, the following question can be answered: Does the outcome variable change over time? Or, in other words: Is there a difference in the outcome variable between two time-points?

To obtain an answer to this question, a paired *t*-test can be used. Consider the hypothetical dataset presented in Table 2.1. The dataset consists of 10 subjects, who were measured on two occasions. The paired *t*-test is used to test the hypothesis that the mean difference between Y_{t1} and Y_{t2} equals zero. Because the individual differences are used in this statistical test, the longitudinal problem of the dependency of the repeated observations within the subjects is reduced to a cross-sectional problem. The test statistic of the paired *t*-test is the average of the differences divided by the standard deviation of the differences divided by the square root of the number of subjects (Equation 2.1).

Table 2.1 Hypothetical dataset for a longitudinal study with two measurements

Id	Y_{t1}	Y_{t2}	Difference (d)
1	3.5	3.7	−0.2
2	4.1	4.0	0.1
3	3.8	3.5	0.3
4	3.8	3.9	−0.1
5	4.0	4.4	−0.4
6	4.1	4.9	−0.8
7	4.0	3.4	0.6
8	5.1	6.8	−1.7
9	3.7	6.3	−2.6
10	4.1	5.2	−1.1

$$t = \overline{d} \Big/ \left(\frac{s_d}{\sqrt{N}}\right) \tag{2.1}$$

where t is the test statistic, $\overline{d}$ is the average of the differences, s_d is the standard deviation of the differences, and N is the number of subjects.

This test statistic follows a *t*-distribution with $(N - 1)$ degrees of freedom. The assumptions for using the paired *t*-test are twofold, namely (1) that the observations of different subjects are independent and (2) that the differences between the two measurements are approximately normally distributed. In research situations in which the number of subjects is quite large (say above 25), the paired *t*-test can be used without any problems. With smaller datasets, however, the assumption of normality becomes important. When the assumption is violated, the non-parametric equivalent of the paired *t*-test can be used (see Section 2.2). In contrast to its non-parametric equivalent, the paired *t*-test is not only a testing method. With the paired *t*-test the average of the paired differences with the corresponding 95% confidence interval can also be estimated.

It should be noted that when the differences are not normally distributed and the sample size is rather large, the paired *t*-test provides a valid result regarding the *p*-value, but interpretation of the average differences can be complicated, because the average is not a good indicator of the mid-point of the distribution even when the sample size is large.

2.1.1 Example

One of the limitations of the paired *t*-test is that the method is only suitable for two measurements over time. It has already been mentioned that the example dataset used throughout this book consists of six repeated measurements. To illustrate the paired *t*-test in the example dataset, only the first and last measurement of this dataset are used. The question to be answered is: Is there a difference in cholesterol between $t = 1$ and $t = 6$?

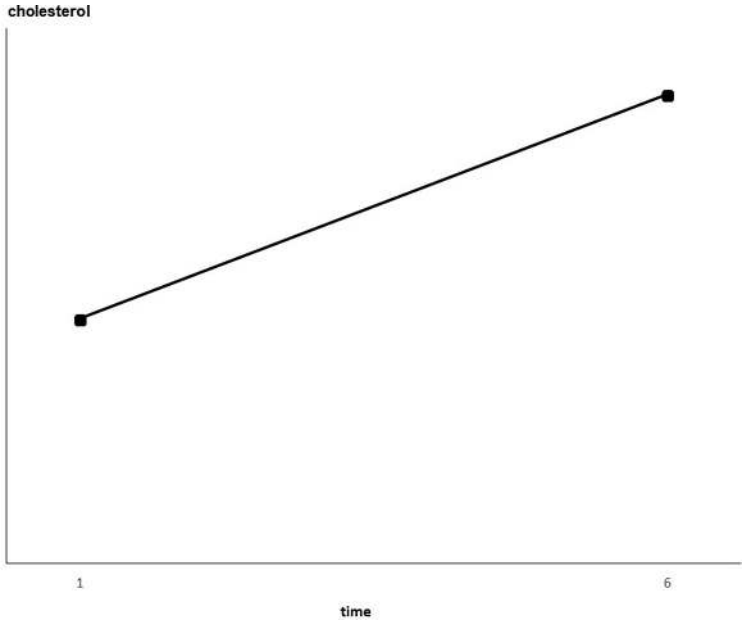

Figure 2.1 Longitudinal development of cholesterol between $t = 1$ and $t = 6$.

Figure 2.1 shows the graphical representation of the data, while Output 2.1 shows the result of the paired *t*-test.

The first lines of the output show descriptive information (i.e. mean values, standard deviation (SD), number of pairs, etc.), which is not really important in light of the research question. The second part of the output provides the more important information. First of all, the mean of the paired differences is given (i.e. −0.68687), and also the 95% confidence interval around this mean (−0.81072 to −0.56302). A negative value indicates that there is an increase in cholesterol between $t = 1$ and $t = 6$. Furthermore, the result of the actual paired *t*-test is given: the value of the test statistic $(t = -10.961)$, with $(N - 1)$ degrees of freedom (146), and the corresponding *p*–value (0.000). The result indicates that the increase in cholesterol is statistically significant ($p < 0.001$). The fact that the increase over time is statistically significant was already clear in the 95% confidence interval confidence interval of the mean difference, which did not include zero.

2.2 Non-parametric Equivalent of the Paired *t*-test

When the assumptions of the paired *t*-test are violated, it is possible to perform the non-parametric equivalent of the paired *t*-test, the (Wilcoxon) signed rank sum test. This signed rank sum test is based on the ranking of the individual difference scores, and does not make any assumptions about the distribution of the outcome variable. Consider the hypothetical dataset presented in Table 2.2. Again, the dataset consists of 10 subjects, who were measured on two occasions.

The signed rank sum test evaluates whether the sum of the rank numbers with a positive difference is equal to the sum of the rank numbers with a negative difference. When those two are equal, it suggests that there is no change over time. In the hypothetical dataset, the sum of the rank numbers with a positive difference is 11.5 (i.e. 1.5 + 4 + 6), while the sum of the rank numbers with a negative difference is 43.5. The exact calculation of the level of significance is complicated, and goes beyond the scope of this book. All statistical handbooks contain tables in which the level of significance can be found (see for instance Altman, 1991), and with all statistical software packages the level of significance can be calculated. For the hypothetical example, the *p*-value is between 0.2 and 0.1, indicating no significant change over time.

The (Wilcoxon) signed rank sum test can be used in all longitudinal studies with two measurements. It is a testing method which only provides *p*-values,

Output 2.1 Results of a paired t-test performed to analyse the difference in cholesterol between $t = 1$ and $t = 6$

Paired samples statistics

	Mean	N	Std. deviation	Std. error mean
cholesterol at $t = 1$	4.435	147	.6737	.0556
cholesterol at $t = 6$	5.1216	147	.92353	.07617

Paired samples test

	Paired differences					t	df	Sig. (2-tailed)
	Mean	Std. deviation	Std. error mean	95% confidence interval of the difference				
				Lower	Upper			
cholesterol at $t = 1$ − cholesterol at $t = 6$	−.68687	.75977	.06266	−.81072	−.56302	−10.961	146	.000

Table 2.2 Hypothetical dataset for a longitudinal study with two measurements

Id	Y_{t1}	Y_{t2}	Difference (d)	Rank number
1	3.5	3.7	−0.2	3
2	4.1	4.0	0.1	1.5[1]
3	3.8	3.5	0.3	4
4	3.8	3.9	−0.1	1.5[1]
5	4.0	4.4	−0.4	5
6	4.1	4.9	−0.8	7
7	4.0	3.4	0.6	6
8	5.1	6.8	−1.7	9
9	3.7	6.3	−2.6	10
10	4.1	5.2	−1.1	8

[1] The average rank is used for tied values.

without effect estimation. In real life situations, it will only be used when the sample size is very small (i.e. less than 25).

2.2.1 Example

Although the sample size in the example dataset is large enough to perform a paired t-test, in order to illustrate the method, the (Wilcoxon) signed rank sum test will be used to test whether or not the difference between cholesterol at $t = 1$ and at $t = 6$ is significant. Output 2.2 shows the result of this analysis.

The first part of the output provides the mean rank of the rank numbers with a negative difference and the mean rank of the rank numbers with a positive difference. It also gives the number of cases with a negative and a positive difference. A negative difference corresponds with the situation that cholesterol at $t = 6$ is less than cholesterol at $t = 1$. This corresponds with a decrease in cholesterol over time. A positive difference corresponds with the situation that cholesterol at $t = 6$ is greater than cholesterol at $t = 1$, i.e. corresponds with an increase in cholesterol over time. The last line of the output shows the z-value. Although the (Wilcoxon) signed rank sum test is a non-parametric equivalent of the paired t-test, in many software packages a normal approximation is used to calculate the p-value. This z-value corresponds with a highly significant p-value (0.0000), which indicates that there is a significant change (increase) over time in cholesterol. Because there is a highly significant change over time, the p-value obtained from the paired t-test is the same as the p-value obtained from the signed rank sum test. In general, however, the non-parametric tests are less powerful than the parametric equivalents and will therefore give slightly higher p-values.

2.3 More than Two Measurements

In a longitudinal study with more than two measurements performed on the same subjects (Figure 2.2),

Output 2.2 Results of a (Wilcoxon) matched pairs signed rank sum test to analyse the difference in cholesterol between $t = 1$ and $t = 6$

Wilcoxon matched-pairs signed-ranks test			
cholt1	**cholesterol at t1**		
with cholt6	**cholesterol at t6**		
Mean rank	Cases		
34.84	29	− Ranks	(cholt6 Lt cholt1)
83.62	118	+ Ranks	(cholt6 Gt cholt1)
	0	Ties	(cholt6 Eq cholt1)
	147	Total	
Z = −8.5637	Two-tailed p = 0.0000		

the situation becomes somewhat more complex. A design with only an outcome variable, which is measured several times on the same subjects, is known as a one-within design. This refers to the fact that there is only one factor of interest (i.e. time) and that this factor varies only within-subjects. In a situation with more than two repeated measurements, a paired t-test cannot be carried out. Consider the hypothetical dataset, which is presented in Table 2.3.

The question: Does the outcome variable change over time? can be answered with a generalised linear model (GLM) for repeated measures. The basic idea behind this statistical method, which is also known as multivariate analysis of variance (MANOVA) for repeated measures is the same as for the paired t-test. The statistical test is carried out for the $T-1$ differences between subsequent measurements. In fact, GLM for repeated measures is a multivariate analysis of these $T-1$ differences between subsequent time-points. Multivariate refers to the fact that $T-1$ differences are used simultaneously as outcome variables. The $T-1$ differences and corresponding variances and covariances form the test statistic for the GLM for repeated measures (Equation 2.2).

$$F = \left(\frac{N-T+1}{(N-1)(T-1)}\right)H^2 \tag{2.2a}$$

$$H^2 = N \times Y_d^t \times \left(S_d^2\right)^{-1} \times Y_d \tag{2.2b}$$

where F is the test statistic, N is the number of subjects, T is the number of repeated measurements, Y_d^t is the row vector of differences between subsequent measurements, Y_d is the column vector of differences between subsequent measurements, and S_d^2 is the variance/covariance matrix of the differences between subsequent measurements.

The F-statistic follows an F-distribution with $(T-1),(N-T+1)$ degrees of freedom. For a detailed description of how to calculate H^2 using Equation 2.2b, reference should be made to other textbooks (Crowder and Hand, 1990; Hand and Crowder, 1996; Stevens, 1996).[1] As with all statistical methods, GLM for repeated measures is based on several assumptions. These assumptions are more or less comparable to the assumptions of a paired t-test: (1) observations of different subjects at each of the repeated measurements need to be independent, and (2) the observations need to be multivariate normally distributed, which is comparable but slightly more restrictive than the requirement that the differences between subsequent measurements are normally distributed. The calculation described above is called the multivariate approach because several differences are analysed together. However, to answer the same research question, a univariate approach can also be used. This univariate approach is comparable to a simple analysis of variance (ANOVA) and is

[1] H^2 is also known as Hotelling's T^2, and is often referred to as T^2. Because throughout this book T is used to denote the number of repeated measurements, H^2 is the preferred notation for this statistic.