8. ANALYSIS OF AUXILIARY DATA

CHAPTER 8

ANALYSIS OF AUXILIARY DATA

8.1 Introduction

In most long-term animal carcinogenicity studies, data are acquired on many variables other than neoplastic and non-neoplastic lesions. As noted in Chapter 5, for example, individual survival times may be of use both in establishing differences in mortality patterns among the various treatment groups and in adjusting for such differences in comparisons of tumour occurrence patterns between groups. Other key variables routinely monitored in long-term studies include body weight and feed consumption, clinical signs of toxicity, haematological parameters and organ weights taken at time of necropsy.

Statistical analysis of such auxiliary data may be broadly categorized into one of three general types. A variety of established procedures for the analysis of censored failure-time data can be used for survival data (Section 8.2). Continuous variables monitored at a particular point in time, such as terminal organ weights or body weight at 12 months on test, may be dealt with using analysis of variance procedures (Section 8.3). Variables observed at successive points in time, such as weekly body weight, are subject to repeated measures or growth-curve analyses (Section 8.4). The remainder of this chapter provides an overview of statistical techniques available within each of these three categories.

The analysis of concomitant information is not the main goal of the statistical analysis of a long-term experiment but assists in interpreting the findings with respect to carcinogenicity. The particular methods addressing carcinogenicity form the main part of this book and have been discussed in the preceding chapters. In this chapter, we shall give only a brief introduction to the variety of techniques available to analyse auxiliary data. It should be made clear that the methods mentioned in this chapter are generally not suitable for the analysis of carcinogenicity.

8.2 Analysis of survival data

It has become apparent throughout this monograph that mortality plays an important role in evaluating carcinogenicity in long-term animal experiments. Before any evaluation of the carcinogenic response is undertaken, a thorough examination of the underlying survival pattern should be performed. This will identify differential mortality patterns that can lead to bias in the assessment of the carcinogenic response. Thus, survival analysis can assist in the choice of appropriate methods to adjust for

differences in intercurrent mortality. The particular methods for the analysis of survival data have already been introduced in Chapter 5. In this section, we give a brief summary of these methods, with cross references to the appropriate sections.

The most common approach to the analysis of survival data in a long-term animal experiment is to estimate the survival curves in each experimental group. These are then displayed graphically, and statistical tests are performed to find whether there are significant differences in survival among the experimental groups or whether there is a significant trend in survival with increasing dose. Methods appropriate to these issues are outlined and illustrated in Section 5.3. The impact of different survival patterns on the assessment of the carcinogenic response is discussed at length in Chapter 2, specifically in Table 2.2. Methods adjusting for differences in survival in the analysis of carcinogenicity are discussed in detail in Sections 5.5, 5.6 and 5.7.

If survival as such appears to be an endpoint which merits more detailed analysis, for example by regression analysis to study the effect of other covariates apart from dose, the proportional hazards model introduced in Section 6.3 is the method of choice. This flexible regression model is a natural extension of the log-rank test for comparing survival in several groups, given in Section 5.3. Furthermore, the proportional hazards model allows the investigation of time-dependent covariates. For example, the influence on an animal's survival of its body weight (if monitored continuously during the experiment) could be analysed using the proportional hazards model. Proportional hazards methods are discussed by Kalbfleisch and Prentice (1980, Chapter 5), Miller (1981b, Chapter 6) and Cox and Oakes (1984, Chapter 8). These authors also provide a thorough treatment of all statistical issues of survival analysis, whereas Lee (1980) provides a more elementary text.

8.3 Analysis of variance

Consider a simple experiment in which there are I treated groups exposed to doses $d_1 < \cdots < d_I$ and an unexposed control, with dose $d_0 = 0$. The manner in which animals are assigned to various treatment groups and the manner in which the experiment is conducted will determine the appropriate analysis for the experiment at hand. As discussed in Chapter 3, the animals should be randomly assigned to each dose in accordance with the experimental design.

The simplest possible randomization scheme is to assign the available animals to the various treatment groups completely at random. With only one animal housed in each cage, this leads to the completely randomized design, discussed in Chapter 3. The familiar randomized block design means that the animals are grouped into a number of homogeneous blocks prior to randomization (for example, on the basis of initial body weight or litter status), with animals from each block randomly assigned to each treatment. In this case, the blocking factor (initial body weight or litter status) must be taken into account in the analysis of variance. Even with complete randomization, the conduct of the experiment is important in determining the method of statistical analysis. With two animals housed in each cage, for example, any cage effects are 'nested' within treatment effects and should be considered in the analysis-of-variance model employed.

There are two major categories of statistical methods available for the analysis of experimental data: parametric methods, which are based on specific assumptions (usually normally distributed data with equal variances in each group), and nonparametric methods, which are not based on such assumptions and often replace the actual observations by their ranks.

In describing these methods, we follow closely two main textbooks – one on parametric methods (Brownlee, 1965) and the other on nonparametric methods (Hollander & Wolfe, 1973). Many other textbooks also provide a good coverage of these methods and could be used when studying technical aspects in detail. The introductory nature of this section requires restriction to essential principles, and it should be borne in mind that the analysis of concomitant information should help in interpreting the findings of a long-term carcinogenicity study but does not, in general, play a central role.

Parametric methods

Let y_{ij} denote the response of animal j ($j = 1, \ldots, n_i$) at dose i ($i = 0, 1, \ldots, I$). As noted earlier, y_{ij} might represent body weight, feed consumption or any other continuous variable observed at a specified point in time. Let

$$\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$$

denote the mean of the n_i observations at dose i and

$$\bar{y} = \sum_{i=0}^{I} \sum_{j=1}^{n_i} y_{ij}/n$$

denote the mean of the

$$n = \sum_{i=0}^{I} n_i$$

animals in the experiment. The standard analysis-of-variance model for the completely randomized design is formulated as

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \tag{8.1}$$

where μ is a constant, τ_i denote the effects of treatment $i = 0, \ldots, I$, with $\sum_{i=0}^{I} \tau_i = 0$, and ε_{ij} are random error terms assumed to be independent, identically-distributed, normal random variables with a mean 0 and variance σ_i^2.

The assumption that the ε_{ij}'s have a normal distribution can be checked using a normal probability plot of residuals (Daniel & Wood, 1971) or using a goodness-of-fit test (Sokal & Rohlf, 1981, p. 696; Miller, 1986, p. 82). If the design is not badly unbalanced (i.e., if the n_i do not vary greatly), then moderate departures from normality have very little effect on the nominal significance levels of the analysis of variance methods presented in this section (Miller, 1986, pp. 80–82). Likewise, inequality of error variances has little effect on the analysis of variance tests unless the design is badly unbalanced (Miller, 1986, pp. 89–92). A preliminary test of homogeneity of error variances is not, in general, recommended. Rather, if visual inspection reveals obvious heterogeneity of error variances, then steps should be taken to try to

reduce that heterogeneity before applying analysis of variance methods (Miller, 1986, pp. 92–94).

Failure of the assumptions regarding normality and homogeneity of error variances could be due to the presence of anomalous values or outliers in the data. If these can be identified from the residual plots, they can be either corrected or eliminated prior to analysis of the data. In some cases, heterogeneity of variance can be avoided by using a suitable transformation of the data. If the variance σ_i^2 is proportional to the group mean $\bar{y}_i$, for example, the transformation $y' = \sqrt{y}$ will result in homogeneous error variances (Brownlee, 1965, p. 145). Generally, the Box–Cox power transformation (Box & Cox, 1964) can be employed in an attempt to achieve simultaneously both normality and homogeneity of variance.

In the usual one-way analysis of variance for the completely randomized design, the variability among the observed treatment group means is compared to the within-group variability using a standard F-test (Brownlee, 1965, p. 312). As indicated in Table 8.1, this involves calculation of a sum of squares, SS_D, between the treatment group means and a pooled within-treatment sum of squares for error, SS_E. After dividing by the degrees of freedom to form the corresponding mean squares MS_D and MS_E, the ratio $F = MS_D/MS_E$ follows a central F-distribution under the null hypothesis $H_0: \tau_i = 0\ (i = 0,1, \ldots, I)$.

Table 8.1 Analysis of variance for the completely randomized design

Source of variation	Degrees of freedom	Sum of squares	Mean square	Expected mean square	F statistic
Dose	I	$SS_D = \sum_{i=0}^{I} n_i(\bar{y}_i - \bar{y})^2$	$MS_D = SS_D/I$	$\sigma^2 + \sum_{i=0}^{I} n_i(\tau_i - \bar{\tau})^2$	MS_D/MS_E
Error	$n - I - 1$	$SS_E = \sum_{i=0}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$	$MS_E = SS_E/(n - I - 1)$	σ^2	
Total	$n - 1$				

If the between-group variation is significantly higher than the within-group variation, then there is evidence of significant differences between the treatment effects τ_i. However, no indication of what these differences are is provided. For this reason, tests for trend or multiple comparison procedures, which are described below, can be informative.

For the following, we denote by

$$s_i^2 = \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2/(n_i - 1)$$

the error mean square for treatment group $i = 0, 1, \ldots, I$ with $E(s_i^2) = \sigma_i^2$. The pooled error mean square is then given by

$$s^2 = \sum_{i=0}^{I} (n_i - 1)s_i^2 \Big/ \sum_{i=0}^{I} (n_i - 1).$$

In what follows, we assume that $\sigma_i^2 = \sigma^2$ for $i = 0, 1, \ldots, I$, so that $E(s^2) = \sigma^2$. Note also that $s^2 = MS_E$ from Table 8.1.

Tests for trend

These have already been discussed in the framework of tumour data and survival (see Chapters 2 and 5) and can also be used with concomitant information. Monotonicity is represented by formulating the alternative $H_1: \tau_i \leq \tau_j$ $(0 \leq i < j \leq I)$, with at least one strict inequality. Monotone decreasing or two-sided alternatives may also be specified.

Armitage (1955) proposed a test for linear trend with equally spaced doses (represented here by the group index i) based on the regression model

$$y_{ij} = a + b \cdot i + \varepsilon_{ij}, \tag{8.2}$$

where a and b are parameters to be estimated. The null hypothesis $H_0: b = 0$ is rejected in favour of the alternative $H_1: b > 0$ if

$$\frac{\sum_{i=0}^{I} n_i(i-\bar{i})\bar{y}_i}{\sqrt{\sum_{i=0}^{I} n_i(i-\bar{i})^2}} \geq t_{\alpha,n-I-1}s,$$

where $\bar{i} = \sum_{i=0}^{I} in_i/n$, and $t_{\alpha,n-I-1}$ denotes the $100(1-\alpha)$ percentile of the t-distribution with $n-I-1$ degrees of freedom. This test is identical to the test for significance of the slope in a linear regression except that s^2 is used to estimate σ^2 rather than the residual mean square error. This is done to eliminate any bias which would be included in the residual mean square for error if the true dose-response curve were not linear. Abelson and Tukey (1963) noted that for $n_i \equiv n_0$ $(i = 1, \ldots, I)$, that is, equal group sizes, Armitage's test is of the form

$$\frac{\sum_{i=0}^{I} c_i\bar{y}_i}{\sqrt{\sum_{i=0}^{I} c_i^2}} \geq t_{\alpha,n-I-1}\frac{s}{\sqrt{n}},$$

where $\sum_{i=0}^{I} c_i = 0$. Although it is impossible to choose the weights c_i to be uniformly most powerful against all possible monotone increasing functions, Abelson and Tukey suggested the weights

$$c_i = \left\{(i-1)\left(1-\frac{i-1}{I+1}\right)\right\}^{1/2} - \left\{i\left(1-\frac{i}{I+1}\right)\right\}^{1/2}.$$

Although Armitage's procedure will be more powerful if the dose-response curve is linear, there exist nonlinear alternatives for which the power of Armitage's test is smaller than the power of the Abelson–Tukey test. Extensions of this procedure to the case of unequal n_i are discussed by Barlow *et al.* (1972) and by Miller (1986, pp. 78–80).

Pairwise group comparisons

Although tests for trend are usually of greatest relevance and interest, it can sometimes be informative to carry out certain pairwise group comparisons (for example, comparing each of the I treatment groups to the control).

A test for the difference between any two groups (indexed, say, by h and i) can be

performed by declaring the difference to be significant if

$$|\bar{y}_i - \bar{y}_h| \geq t_{\alpha/2,n-I-1} s \left(\frac{1}{n_i} + \frac{1}{n_h}\right)^{1/2}, \tag{8.3}$$

where $t_{\alpha/2,n-I-1}$ denotes the $100(1-\alpha/2)$ percentile of the t-distribution with $n-I-1$ degrees of freedom and where α is the nominal significance level. This provides a valid test for the single comparison of group i to group h. However, if several such tests, say $M>1$, are carried out (for example, comparing each treatment group in turn to the control, for which $M=I$), then the overall significance level (that is, the probability of finding at least one of the M tests significant under $H_0: \tau_i = 0$ for all i) using the criterion in (8.3) will exceed the nominal level α. If, for example, two independent comparisons are performed with $\alpha = 0.05$, then the probability of declaring at least one of the two differences significant under H_0 is $1.0-(1.0-0.05)^2 = 0.0975$, which is substantially higher than the nominal significance level, 0.05, of the two separate tests.

The goal of multiple-comparisons procedures is to allow several comparisions of interest to be made while maintaining the overall significance level at a fixed α. The methods presented require that the M comparisons to be made be chosen *a priori*. The simplest multiple comparisons method is the Bonferroni method (Miller, 1981a). The test criterion for the Bonferroni method is identical to that in (8.3) above, except that α is replaced by $\alpha' = \alpha/M$. A slight improvement on the Bonferroni method, particularly for large M, is provided by the Dunn–Sidák method (Dunn, 1974; Miller, 1981a), for which the test criterion is again identical to that in (8.3), but with α replaced by $\alpha'' = 1-(1-\alpha)^{1/M}$.

The comparisons that are likely *a priori* to be of general interest are of each treatment group in turn to the control. When $n_i = n_0$ for all i, the method of choice is the many-to-one t-test (Dunnett, 1955), which is performed by declaring a difference to be significant if

$$|\bar{y}_i - \bar{y}_0| \geq |d|_{\alpha,I,n-I-1} s\sqrt{2/n_0},$$

where $|d|_{\alpha,I,n-I-1}$ is tabulated by Dunnett (1955, Table 2). The many-to-one t-test for unequal sample sizes and tabulated critical values for the general case are described by Dunnett (1964) and Dutt *et al.* (1975, 1976).

Extensions of the one-way analysis of variance

Extension of the one-way analysis of variance may be required, since, as indicated in Chapter 3, many experiments do not follow the simple structure of a completely randomized design. For example, consider a two-generation study in which the parent or F_0 generation has been assigned to treatment groups in accordance with a completely randomized design, and the males and females in the same groups were mated on a one-to-one basis. Suppose now that a fixed number $m \geq 2$ of pups of each sex was selected from each litter to continue on test in the second or F_1 generation. Two animals from the same litter may be expected to have similar characteristics because of their common genealogy (see Chapter 3). Thus, this experiment has two levels of randomization, since the litters are first randomly assigned to treatments and then the pups are randomly selected from the litters. Here, the litter effect is considered to be nested within the main treatment effect.

As a second example, consider a single-generation experiment in which animals are assigned to treatments using a completely randomized design, but two or more animals are caged together. In this case, it is possible that animals housed together respond more similarly than animals housed in different cages. Thus, the cage effect is nested within the main treatment effect. These experiments can be analysed using the nested analysis of variance model

$$y_{ijk} = \mu + \tau_i + \lambda_{j(i)} + \varepsilon_{ijk}, \tag{8.4}$$

where y_{ijk} denotes the response for animal $k = 1, 2, \ldots, m_{ij}$, in litter or cage $j = 1, 2, \ldots, n_i$, in treatment group $i = 0, 1, \ldots, I$. Here τ_i denotes the effect of treatment i, $\lambda_{j(i)}$ denotes the random effect of the jth level of the nesting factor within treatment i, and ε_{ijk} denotes a random error term. The ε_{ijk}'s are assumed to be independent normal random variables with mean 0 and variance σ_ε^2 and the $\lambda_{j(i)}$ are independent normal random variables with mean 0 and variance σ_λ^2.

If the m_{ij} are all equal, the analysis is straightforward. The litter or cage averages can be calculated and analysed using the procedures discussed previously for the randomized design. In addition, the significance of litter or cage effects can be assessed using standard analysis-of-variance procedures. If the number of animals varies from litter to litter, then procedures for analysis of the experiment are more complicated. Healy (1972) has proposed an analysis based on weighted averages of litter means, where the weights are estimated from the variance components. This procedure ignores the uncertainty in the estimation of the weights which may invalidate the technique for small sample sizes. Tietjen (1974) has examined a test for treatment effects based on a Satterthwaite approximation (Searle, 1971). However, the conventional F-test ignoring the imbalance appears to perform better than the Satterthwaite approximation.

Nested designs can be further generalized to the case where there are more than two levels of randomization. For example, consider a two-generation study in which males and females are assigned to treatments under a completely randomized design, with each male randomly paired with two females. This experiment may be viewed as having three levels of randomization, with sires randomly assigned to treatments, dams randomly allocated to sires for mating and pups randomly selected within the litters. The litters are thus nested within sires which are in turn nested within treatments.

Consider an experiment in which n animals are to be assigned to each of $I + 1$ experimental groups. Suppose that the $(I + 1)n$ animals are divided into n groups of size $I + 1$ so that all animals in the same group have similar weights, and that one animal from each weight group is randomly assigned to each treatment. The groups in this experiment are referred to as 'blocks', and the experiment is referred to as following a 'randomized complete block design'. The blocks do not have to be defined in terms of animal weight. For example, one could consider litters to be blocks and assign one animal from each litter to each treatment (see Section 7.5). The randomized complete block design can be analysed using the analysis of variance model

$$y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}, \tag{8.5}$$

where y_{ij} denotes the response of the animal from block j given treatment i, μ is a

constant, τ_i denotes the effect of treatment i, β_j denotes the effect of block j, and ε_{ij} is a random error term.

The above model is based on the assumption that each treatment has the same effect on each block. The effect of blocking is to remove from the error sum of squares a term which measures the variation in the observed response among blocks. It also reduces the experimental error of the estimated differences between treatments. If the inter-block variation is large, the randomized complete block design will provide more sensitive tests for treatment effects than the completely randomized design.

Nonparametric methods

The methods outlined above rely on the parametric assumption that the error terms in the respective analysis of variance models – (8.1), (8.2), (8.4) and (8.5) – are distributed according to a normal distribution with mean zero and some unknown variance. This assumption is frequently not met by the data one is analysing. Therefore, methods that make less stringent assumptions about the underlying distribution have been developed which can easily be employed, as they are simply based on ranks. When compared to the methods based on assumptions of normality, these methods have been shown to lose only slightly in efficiency when the assumptions are valid, but can be considerably more efficient when they do not hold.

We shall give a brief introduction to nonparametric methods which can be used to analyse continuous variables monitored at a particular point in time. This introduction follows the description of these methods in the textbook by Hollander and Wolfe (1973).

Let y_{ij} denote the response of animal j ($j = 1, \ldots, n_i$) to dose i ($i = 0, 1, \ldots, I$). We deal again with model (8.1), $y_{ij} = \mu + \tau_i + \varepsilon_{ij}$, but assume that the error terms ε_{ij} are mutually independent and follow some continuous random distribution. To test the null hypothesis, that all treatment effects τ_i ($i = 0, 1, \ldots, I$) are equal, against the alternative that they are not all equal, all n observations y_{ij} are ranked in ascending order, giving rank 1 to the lowest value and rank n to the largest. Let r_{ij} be the rank of observation y_{ij} in this joint ranking. The sum of ranks for observations in group i is

$$R_i = \sum_{i=1}^{n_i} r_{ij} \qquad (i = 0, 1, \ldots, I),$$

the average rank in group i being denoted by $R_{i.} = R_i/n_i$. The average rank of all n observations is $R_{..} = (n+1)/2$. In order to assess whether the ranks in the individual groups differ from the overall average, the statistic

$$H = \frac{12}{n(n+1)} \sum_{i=0}^{I} n_i (R_{i.} - R_{..})^2$$

is computed. Under the null hypothesis of no difference between the treatment groups, H follows asymptotically a chi-square distribution with I degrees of freedom. This test is usually referred to as the Kruskal–Wallis test. For small sample sizes and limited numbers of groups ($I = 2$, $n_i \leq 5$), tables of exact critical values for H have been published (Hollander & Wolfe, 1973).

In the case of ties among the observations, the mean of the respective ranks, or midranks, may be assigned to all the tied observations. Consider g sets of tied observations and let $t_j\,(j=1,\ldots,g)$ be their respective size. Then H should be corrected by dividing its value by

$$1-\left\{\sum_{j=1}^{g}(t_j^3-t_j)\Big/(n^3-n)\right\}.$$

In the case of two groups only ($I=1$), the Kruskal–Wallis test is identical to the Mann–Whitney or Wilcoxon test.

Test for monotone trend

A test for monotone trend will be indicated in most experiments involving a series of increasing dose levels. The analysis of concomitant information must assess whether the observations of interest follow a corresponding trend. In addition, when multiplicity of tumours (Section 7.3) or graded responses (Section 7.4) are considered, nonparametric approaches to the analysis of these endpoints are indicated.

From the overall ranking used for the Kruskal–Wallis test, one test statistic for the presence of a positive trend with increasing dose can be derived as follows. The rank sum R_0 of the first group can be viewed as a two-sample Wilcoxon statistic, comparing the responses in group 0 with the pooled responses in groups 1 to I. Similarly, R_0+R_1, the rank sum of groups 0 and 1 combined, can serve as a test statistic to compare these to groups jointly against the combined group 2 to I.

With $I+1$ groups, I such two-sample comparisons can be considered. The sum of all their test statistics will be considered to test for the presence of a positive trend

$$L=R_0+(R_0+R_1)+\cdots+(R_0+R_1+\cdots+R_{I-1})=\sum_{i=0}^{I-1}(I-i)R_i. \tag{8.6}$$

Under the null hypothesis of no difference between the $I+1$ groups the expectation of L is

$$E(L)=\frac{n+1}{2}\sum_{i=0}^{I-1}s_i,$$

where

$$s_i=\sum_{j=0}^{i}n_j$$

is the cumulative sample size up to, and including, group i. The variance of L is

$$\operatorname{var}(L)=\frac{n+1}{12}\left\{\sum_{i=0}^{I-1}s_i(n-s_i)+2\sum_{i=0}^{I-2}\sum_{j=i+1}^{I-1}s_i(n-s_j)\right\}. \tag{8.7}$$

Small values of L, that is, values below the expectation, are indicative of a positive trend. This leads to the following standardized test statistic

$$T_L=[E(L)-L]/[\operatorname{var}(L)]^{1/2}. \tag{8.8}$$

Asymptotically, T_L follows a standard normal distribution. If the value of T_L exceeds z_α, the upper $(1-\alpha)$ percentile of the standard normal distribution, a positive trend can be concluded with a significance level of α.

In the case of ties, midranks, r_{ij}^* say, are assigned and the test statistic T can be corrected by replacing, in formula (8.7), the term $(n+1)/12$ by

$$\frac{1}{n(n-1)} \sum_{i=0}^{I} \sum_{j=1}^{n_i} \left(r_{ij}^* - \frac{n+1}{2} \right)^2.$$

This test, which is similar to the test proposed by Page (1963) for complete block designs and also described by Hollander and Wolfe (1973), has been proposed by Wahrendorf *et al.* (1985) for complete randomized designs in the framework of mutagenicity data. However, it is also perfectly applicable to the analysis of concomitant information or special responses in long-term animal experiments. Here, the consistency and strength of the trend can also be estimated by some nonparametric measure of the stochastic ordering between two populations.

As can be seen in (8.6), this nonparametric trend test weights the rank sums of all treatment groups by an integer score. For the many experiments conducted on a multiplicative dose scale, these scores correspond to the logarithms of the dose levels. Marascuilo and McSweeney (1967) have proposed a second test for trend where the actual dose levels are used as scores, and the construction of such a general rank test has also been noted by Cuzick (1985). These tests can easily be performed in a stratified situation by summing the differences $E(L) - L$ and the variances calculated according to (8.7) over the strata and then forming a standardized test statistic according to (8.8).

The above tests for trend are based on an overall ranking of the observations in all $I+1$ groups. Another nonparametric test of trend, which is based on all $I(I+1)/2$ pairwise comparisons of two groups, is the Jonckheere test (Jonckheere, 1954), also described by Hollander and Wolfe (1973). Two groups, u and v say $(u, v = 0, 1, \ldots, I; u \neq v)$, can also be compared by a Wilcoxon test by counting the number of pairs (α, β) for which $y_{u\alpha} < y_{v\beta}$. If $\phi(a, b) = 1$ if $a < b$, and 0 otherwise, this is

$$U_{uv} = \sum_{\alpha=1}^{n_u} \sum_{\beta=1}^{n_v} \phi(y_{u\alpha}, y_{v\beta}),$$

frequently referred to as Mann–Whitney counts. Summing these U_{uv} from all $I(I+1)/2$ pairwise comparisons gives the Jonckheere statistic

$$J = \sum_{u=0}^{I-1} \sum_{v=u+1}^{I} U_{uv}.$$

Under the null hypothesis of no difference between the $I+1$ experimental groups, this follows asymptotically a normal distribution with expectation

$$E(J) = \left\{ n^2 - \sum_{i=0}^{I} n_i^2 \right\} \Big/ 4$$

and variance

$$\text{var}(J) = \left\{n^2(2n+3) - \sum_{i=0}^{I} n_i^2(2n_i+3)\right\}\Big/72.$$

In this case, large values of J are indicative of a positive trend, leading to the standardized test statistic

$$T_J = [J - E(J)]/[\text{var}(J)]^{1/2}$$

which, under the null hypothesis, follows a standard normal distribution.

Multiple comparisons

For the purpose of multiple comparisons, say M pairwise comparisons among groups (for example, comparison of each exposed group in turn to the control group), the average ranks of each group $R_{i.}$ $(i = 0, 1, \ldots, I)$ are used. These correspond to the group means used earlier in the parametric approach, and the underlying arguments regarding the logic of adjusting tests (because of the multiplicity of comparisons) are exactly the same as described there. We shall outline the approximation procedure given by Dunn (1964) and described by Hollander and Wolfe (1973).

Maintaining an overall significance level of α, one can decide that the response in group i is different from the response in group h, that is, $\tau_i \neq \tau_h$, if

$$|R_{i.} - R_{h.}| \geq z_{\alpha'}\left\{\left(\frac{n(n+1)}{12}\right)\left(\frac{1}{n_i} + \frac{1}{n_h}\right)\right\}^{1/2},$$

where $z_{\alpha'}$ is the $100(1-\alpha')$ percentile of the standard normal distribution and $\alpha' = \alpha/2M$. It has to be noted that the above comparison between a group i and a group h is based on the overall ranking of all observations, and, thus, it depends on the observations in the other groups. For detailed discussions, see Hollander and Wolfe (1973) and Miller (1981a).

8.4 Repeated measures and growth curves

A long-term study often involves repeated measurements of the same parameter in the same subject over a period of time. For example, blood samples may be taken from a subsample of animals at specified points in time and subjected to detailed haematological evaluation. Body weights are generally recorded for all animals in a study on a regular basis in order to establish growth profiles. Similar records of food consumption are also maintained.

Since early work on the analysis of such data by Box (1950), statistical methods for repeated measures of growth-curve data have undergone extensive development (Geisser, 1980; Woolson & Leeper, 1980). The essential difference between these procedures and those discussed in Section 8.2 is that repeated measures are taken on the same individual, and that one needs to allow for the possibility of correlation among these observations. In the first part of this section, we consider the use of

multivariate linear models for this purpose. Nonparametric or related approaches to this same problem are then considered in the second part. Finally, some biologically-based growth models are described briefly.

Multivariate linear models

Let Y_{it} denote the measured value of a particular variable for individual $i = 1, \ldots, N$ at time $t = 1, \ldots, T$. These data may be conveniently summarized in matrix form as

$$\mathbf{Y} = \begin{bmatrix} Y_{11} & \cdots & Y_{1T} \\ \vdots & & \vdots \\ Y_{N1} & \cdots & Y_{NT} \end{bmatrix},$$

where each row corresponds to the set of results for one individual obtained during the course of the study period, and each column represents the results for all individuals at a particular point in time. A multivariate linear model for the data $\mathbf{Y}$ is then

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}, \tag{8.9}$$

where $\mathbf{X}$ is an $N \times P$ design matrix consisting of zeros and ones indicating the treatment assigned to each individual and $\boldsymbol{\beta}$ is a $P \times T$ matrix of unknown parameters reflecting both treatment and time effects, with each column representing linear model regression coefficients for that time period. The rows $\mathbf{E}_i$ of the error matrix $\mathbf{E}$ are assumed to be independent, multivariate normal random variables with mean $\mathbf{0}$ and covariance matrix $\mathbf{S}_i$. Under this model, the rows $\mathbf{Y}_i$ are independent, multivariate normal random variables with mean $\mathbf{M}_i = (\mathbf{X}\boldsymbol{\beta})_i$ and covariance matrix $\mathbf{S}_i$. As for the univariate analysis of variance procedures discussed in the previous Section 8.3, a multivariate analysis of variance based on the model in (8.9) may be carried out under the homoscedasticity assumption that $\mathbf{S}_i = \mathbf{S}$ for all i. Details of this analysis are given by Morrison (1976) and in other texts on multivariate methods. The case of heteroscedasticity has been discussed by Chakravorti (1974). Procedures for handling missing values have been proposed by Kleinbaum (1973) and Leeper and Woolson (1982).

This multivariate approach to data on repeated measurements requires that sufficient data be available in order to estimate the many unknowns involved in the mean vector $\mathbf{X}\boldsymbol{\beta}$ and the dispersion matrix $\mathbf{S}$ of the data $\mathbf{Y}$. To reduce the dimensionality of $\boldsymbol{\beta}$, Potthoff and Roy (1964) suggested the use of polynomials in time t to provide for longitudinal effects (see also Khatri, 1966). The use of a low-order polynomial of degree $Q < T$, for example, would reduce drastically the number of parameters to be estimated when T is large. (This approach has been employed in the analysis of growth-curve data from a long-term bioassay of *ortho*-toluenesulfonamide conducted by Arnold *et al.*, 1980.) Similarly, further assumptions could be made concerning the form of $\mathbf{S}$ (Grizzle & Allen, 1969), although oversimplification may result in an increase in false-positive rates (Boik, 1981; Elashoff, 1981; Schwertman *et al.*, 1981). A parametric approach in which the regression coefficients of the growth curves are considered as random variables and provide the basis for statistical inference has been proposed by Schach (1982).

Nonparametric methods and related approaches

As with the nonparametric methods discussed in Section 8.3 for the univariate linear model, there exist multivariate nonparametric methods which are again based on ranks and can be applied to the situation of repeated measurements (Bhapkar & Patterson, 1977). Koch *et al.* (1980) provide a comprehensive overview of the methods available for different situations. A specific application suitable to situations of partially incomplete observations is given by Koziol *et al.* (1981). Both these papers provide many further references.

Another approach to the analysis of growth curves would be to fit a certain parametric model to the shape of each individual growth curve, and to extract certain parameters or functionals from the fitted curves which are particularly relevant to the biological aspects of the assay. For example, the slope of the growth curve or an estimate of its second derivative (acceleration of growth), the area under the curve, the location or value of a maximum or minimum, or a categorization of the curve's profile may represent such measures derived for each animal from its growth curve. These measures for the different treatment groups can then be compared using the nonparametric techniques for one-way analysis of variance as given in Section 8.3. Thus, this approach, initially suggested by Wishart (1938) and considered by Prestele *et al.* (1979) and Haux (1985), reduces the multivariate data of repeated measurements to univariate comparison. This approach is very promising for the analysis of auxiliary data in carcinogenicity studies, since it is based on easily interpreted parameters, it can allow for different numbers of observations per animal and it does not rely on strong parametric assumptions.

Robust estimation procedures (Huber, 1981) may also be considered as a means of avoiding the parametric assumptions required in the multivariate linear model. Pendergast and Broffitt (1985), for example, considered the use of M-estimation for growth curve data. This approach is based on an arbitrary loss function chosen to be less sensitive to outlying values than the quadratic loss function on which the multivariate analysis of variance is based. It appears to provide a robust alternative to the latter analysis. Like the nonparametric methods based on ranks, however, this robustness is achieved only with considerably more computational effort.

Biological growth-curve models

The multivariate linear models discussed previously are purely statistical in nature and, while often providing an adequate description of growth-curve data, do not have an underlying biological basis. Sandland and McGilchrist (1979) consider models which allow for an initial period of rapid growth followed by a period of slower growth and then a levelling off or even a decline in body weight. The logistic model, for example, is based on the differential equation

$$\frac{dW(t)}{dt} = aW(t)[b - W(t)],$$

where $W(t)$ denotes the expected body weight at time t. These models are specified in

terms of biologically meaningful parameters and predict the anticipated shape, although this does not necessarily imply that the model represents the correct underlying growth mechanism (Kowalski & Guire, 1974).

Other stochastic models may be based on autoregressive processes in which successive errors may be correlated (Glasbey, 1979), or on stochastic differential equations (Sandland & McGilchrist, 1979).

Models which relate body weight to food consumption have also been proposed (Daniel, 1983). This last approach has been used to distinguish between weight changes attributable to changes in food consumption and those resulting from alterations in metabolism.