7. SPECIAL TOPICS

7.1 Introduction
7.2 Statistical inference at multiple sites
7.3 Multiplicity of tumours
7.4 Graded responses
7.5 Multifactorial designs: combining results
7.6 Litter effects
7.7 Association among tumour types
7.8 Historical control tumour rates

CHAPTER 7

SPECIAL TOPICS

7.1 Introduction

In the previous chapters we considered the broad aspects of the design and analysis of long-term animal experiments. There are, however, many special aspects that have not been considered in the previous chapters, but that merit mention, as they occur frequently in practice. These aspects are all concerned, in one way or another, with multiplicity; that is, in design or analysis, one factor or one variable is added or allowed to have multiple levels.

In Section 7.2, we shall deal with the problem that arises from the study of more than one tumour of interest in a long-term animal experiment, and, thus, the statistical inference has to be carried out on data from a variety of sites. In some experimental systems, the multiplicity of tumours at one site is viewed as a relevant biological endpoint, and the appropriate statistical methods are discussed in Section 7.3. A tumour response may be graded according to a fixed number of clinical or histological categories, and the methods applicable to these situations are outlined in Section 7.4. If a further dimension is added to the design of an experiment, either by stratification or by adding another exposure factor, it must be considered in the analysis, and methods for doing so are outlined in Section 7.5. In some experiments, the information about the common litter membership of the test animals may be kept in order to influence the design of the experiment. Statistical methods making use of this information are discussed in Section 7.6. Several tumours can occur simultaneously in one animal; in Section 7.7, we outline methods for assessing the association among tumour types in appropriate experiments. Finally, Section 7.8 deals with the incorporation of historical information on tumour incidence in untreated control animals into the statistical analysis.

7.2 Statistical inference at multiple sites

The false-positive and false-negative rates are of great importance in any screening procedure. In a carcinogenesis screening test, the false-positive rate is the percentage of noncarcinogenic compounds which are incorrectly classified as carcinogens, and the false-negative rate is the percentage of carcinogenic compounds which are incorrectly classified as noncarcinogens. In most animal carcinogenesis experiments, it is not possible to predict *a priori* potential target organs at which carcinogenic effects are

likely to occur. Thus, although the effect of a carcinogenic agent is likely to be concentrated in one or a few target organs, all organs which are examined histopathologically must be evaluated for evidence of carcinogenesis. Because of the multiple comparisons involved in the statistical evaluation of tumour incidence data from several organ and tissue sites, there is a danger of inflating the false-positive error rate. In particular, a simplistic decision rule that routinely labels a chemical as a carcinogen whenever a single tumour increase is significant at the 5% level for any exposed group at any of the organs examined can result in a false-positive rate considerably greater than 5% (Fears *et al.*, 1977; Salsburg, 1977; Fears & Tarone, 1977; Haseman, 1977; Elashoff *et al.*, 1979). Some authors have argued that the inflated false-positive rate associated with such a naive decision rule invalidates the use of animal experiments in screening chemicals for carcinogenesis (Salsburg, 1977). Others have noted that such a decision rule is not in fact used in practice, and that rules which attempt to model the actual decision process indicate that false-positive rates are close to the nominal level (Fears *et al.*, 1977; Fears & Tarone, 1977; Haseman, 1977; Gart *et al.*, 1979; Haseman, 1983b).

A major problem in trying to estimate the error rates of a carcinogenesis screening test, or to recommend explicit adjustments for the multiple comparisons involved, is the difficulty in modelling the interaction that takes place between the statistician and scientists of other disciplines included in the decision-making process. The evaluation of the carcinogenic potential of a test compound is not strictly a statistical decision. It is impossible to incorporate the totality of knowledge and experience of the pathologists, toxicologists, pharmacologists and other scientists involved in the decision-making process into a simple statistical model. Nevertheless, investigations based on the comparison of unadjusted tumour rates using the Fisher–Irwin exact test have led to statistical devices that can be used to keep false-positive rates under control. The most important finding of these investigations is that organs with low spontaneous tumour rates can be ignored effectively in the calculation of false-positive rates (Fears *et al.*, 1977; Gart *et al.*, 1979; Haseman, 1983b). In particular, for a given experimental design and nominal significance level, one can compute the minimum number of animals, in the combined control and exposed groups, which must be found with a given tumour in order for a significant result to be obtained using the Fisher–Irwin exact test. Accordingly, only those organs with spontaneous tumour rates for which this minimum number of tumours is likely to be obtained need be considered in determining an adjustment for multiple comparisons (Gart *et al.*, 1979). An alternative, but related, approach to control for multiple comparisons is to test for tumour increases using one nominal significance level, say $\alpha_1 = 0.05$, for organs with low spontaneous tumour rates (for example, tumour rates less than 2% for experiments with 50 animals per group), and a second, smaller nominal significance level, $\alpha_2 < 0.05$, for all other organs. The actual value of α_2 leading to a false-positive rate of 5% can be calculated for each species/strain/sex combination for which good estimates of spontaneous tumour rates exist. Various modifications of this approach are possible, for example, using a different nominal level for each value of the spontaneous tumour rate, the nominal level decreasing with increasing tumour rate. Of course, if there is evidence *a priori* that a test compound is likely to produce a carcinogenic effect at a

particular organ, then the nominal significance level for the suspected organ should not be reduced, regardless of the magnitude of the associated spontaneous tumour rate.

In order to avoid the multiple comparisons problem, Brown and Fears (1981) proposed a method of calculating a single overall significance level for a carcinogenesis experiment. Suppose that, in an experiment with one exposed group and a concurrent control group, T organs are examined for the presence of a tumour in each animal. Then, each animal may have tumours discovered in one of 2^T possible combinations of organ sites, ranging from 'no tumours found' to 'tumours found in all T organs'. For each of the T organs, a Fisher–Irwin exact test can be performed. For a fixed significance level, α, they provide a method for calculating the exact permutational probability of at least one significant Fisher–Irwin test, conditional on the 2^T marginal totals (each marginal total is the number of animals with tumours only in the organs represented by one of the 2^T possible combinations). An overall significance level can be calculated by applying the method with α set equal to the smallest of the T p-values observed in the individual Fisher–Irwin exact tests. Unlike previously discussed methods, the method of Brown and Fears requires no prior knowledge of the spontaneous tumour rates.

Meng (1985) has proposed a Bayesian approach to the multiple comparisons problem, incorporating historical data on spontaneous tumour rates in a manner suggested by Dempster *et al.* (1983). The method proposed by Meng has the disadvantage that an increased tumour rate at a single organ can be diluted by a general decrease in tumour rates at other organs (such general decreases have been observed due to the reduced food consumption in exposed animals). However, the development of related methods warrants further investigation.

7.3 Multiplicity of tumours

The methods described in Chapter 5 concentrate on the presence or absence of one or more tumours in an animal. This reflects the fact that the fundamental measure of carcinogenic effect is usually taken as the total number of *animals* which develop a tumour of a given type rather than the total number of such *tumours* (Peto *et al.*, 1980). The main reasons for this are, firstly, that multiple tumours in an animal are not independent events (a few animals often get a large number of tumours) and, secondly, that for tumours not observable until death it is impossible to determine whether treatment has caused tumours or has merely affected their progression. In theory, as noted by Peto *et al.* (1980), a chemical which inhibits metastatic spread of localized tumours might allow animals with tumours to live longer and have time to develop more tumours. Furthermore, it is the individual animal that is randomized among the dose groups, and thus the animal should be treated as the experimental unit.

In some cases, however, experimental systems have been specifically developed to quantify response in terms of multiplicity, and it is useful to have methods available which take into account the number of tumours at a given site. The most widely used system of this type involves the mouse skin, where topical application of carcinogens can produce a sequence of multiple lesions – usually papillomas. Continuous surveillance of the animals is necessary to observe the course of the lesions accurately.

This involves observing the times of first occurrence of the lesions and the times when some disappear due to systemic regression, scratching, biting or other external reasons.

Another experimental model developed by Shimkin (Shimkin & Stoner, 1975) measures the development of lung adenomas in mice in a relatively short period of time. Animals are killed after seven or eight months and the number of lung adenomas are counted as a quantitative endpoint. In a long series of experiments with urethane, this same model was used to try to elucidate the mechanism of action of urethane carcinogenesis; in direct screening assays with this model, urethane is usually considered as a positive control.

The induction of multiple mammary tumours in female Sprague Dawley rats, mainly by 7,12-dimethylbenz[*a*]anthracene, is an animal model which has been developed to study the possible inhibitory effect of other chemicals such as, for example, vitamins, on carcinogenesis. Multiplicity of tumours in this model is considered a quantification of the response. This rat mammary model was developed as a quick model for direct screening of compounds, but it has lost favour for this purpose due to its limited specificity.

In general, none of these special animal models are considered to provide conclusive evidence on their own when used for screening the carcinogenicity of chemicals. They play a more useful role in the study of the mechanisms of carcinogenesis. Nevertheless, there is an interesting challenge in using the appropriate methodology for analysing such studies.

It is necessary to distinguish between the situation in which the number of tumours is counted at a fixed point in time in each experimental group and that in which the time of development of each individual tumour is accurately recorded. In the first case, the number of tumours seen in an individual animal represents the basic information. Let x_{li} denote this number for the lth animal in the ith experimental group ($l = 1, \ldots, n_i$; $i = 0, \ldots, I$). Analysis of variance methods can be applied to such data. Both parametric and nonparametric methods are available to test the null hypothesis that there is no difference between the experimental groups against either the unstructured alternative that the responses are different between groups, or the ordered alternative that the responses are increasing with increasing dose level.

In Section 8.3, detailed methods are given for the analysis of concomitant information by parametric or nonparametric one-way analysis of variance. These methods can be applied directly by treating the tumour counts x_{li} ($l = 1, \ldots, n_i$; $i = 0, \ldots, I$) as the basic observation per animal. In a parametric analysis of variance, the x_{li}'s are used directly for the calculation of the test statistics, whereas in the nonparametric methods they are converted into ranks. In the first case, one may also apply transformations, for example the square root or logarithm, to achieve a better fulfilment of the underlying assumptions (equal variances, normal distribution of observational errors).

One parametric approach has been proposed by Drinkwater and Klotz (1981). They suggest that the number of tumours per animal has a Poisson distribution, that is, the probability, $f(t)$, that an animal bears t tumours is

$$f(t) = e^{-\lambda}\lambda^t/t! \qquad (t = 0, 1, 2, \ldots).$$

Furthermore, in order to account for the empirically observed variation, they suggest that the parameter λ, which is the mean number of tumours per animal, is subject to further random variation modelled by a gamma distribution. This leads to a so-called negative binomial distribution which is frequently applied in the analysis of count data (Anscombe, 1949, 1950; Bliss, 1953). Drinkwater and Klotz (1981) outline the calculation of a likelihood ratio test statistic to compare the tumour counts in two experimental groups under this parametric model. Their comparison of this method with the t-test (parametric analysis of variance), the Wilcoxon test (nonparametric analysis of variance) and the chi-square test from a 2×2 table contrasting tumour incidence in the two groups is not fully conclusive as it is based on simulated data derived exactly from the model of a negative binomial distribution. However, they do provide some empirical support for this model. In the absence of any firm knowledge about a parametric model for the variation of the tumour counts, we recommend a nonparametric analysis of such tumour counts at a fixed point in time.

When the time of appearance of the multiple tumours is recorded for each animal, the methods developed by Gail *et al.* (1980) can be used. They consider that, in each animal, tumours are observed to appear at times $T_1 < T_2 < \cdots < T_K$. The notation of capital letters indicates that we introduce their approach in terms of the observable random variables. In addition, for any animal there is a censoring time C which is assumed to be independent of the sequence $T_1, T_2, \ldots, T_K$, K being the largest integer such that $T_K < C$. The jth gap is defined as $Z_j = T_j - T_{j-1}$, with $T_0 = 0$ for convenience. If the probability distribution of the jth gap depends only on j and on t_{j-1}, the value of T_{j-1}, but not on the earlier times $t_1, \ldots, t_{j-2}$, the sequence $T_1, T_2, \ldots$ is called a Markov sequence. The hazard function of Z_j is denoted by $h(z \mid j, t_{j-1})$ and is used as the basic element in developing inferential strategies. For this purpose, the authors consider that $h(z \mid j, t_{j-1})$ has a known parametric form or that $h(z \mid j, t_{j-1})$ is independent of t_{j-1}, in which case a so-called semi-Markov model results. In both situations the effect of the different treatment groups on the occurrence of tumours is modelled similarly to the proportional hazards model (Cox, 1972). To keep the notation simple, we consider two treatment groups, 1 and 2. Conditional on t_{j-1}, it is assumed that the gap Z_j has hazard

$$\exp(\alpha_j)h(z \mid j, t_{j-1}) \quad \text{or} \quad \mathrm{h}(z \mid j, t_{j-1})$$

according to whether the animal has been given treatment 1 or 2, respectively. Gail *et al.* (1980) discuss methods of estimating the parameters α_j, of testing their homogeneity, and of testing that the common value of α is zero. As mentioned above, various specializations of $h(z \mid j, t_{j-1})$ are used.

The models used in this approach can be derived from the so-called m-site model, which is often used in mathematical theories of carcinogenesis (Whittemore & Keller, 1978). The original paper should be consulted in detail when applying these methods. An alternative strategy is to apply the proportional hazards model (Cox, 1972) (see Section 6.3) with its feature of time-dependent covariates or strata (Kalbfleisch & Prentice, 1980) to adjust for the number of tumours already developed while comparing the hazards of developing the next tumour (Scribner *et al.*, 1983).

7.4 Graded responses

As the principal interest in carcinogenicity experiments is the presence or absence in an animal of a tumour of a given type, we have concentrated on techniques to analyse response as a 0–1 variable. Section 7.3 dealt with the multiplicity of tumours where counts of tumours occurring at a given site represent a quantification of the response. Another quantitative measure of the carcinogenic response is the grade of the lesion according to some pathological criteria. These may differ from site to site and between schools of experimental pathology, but it is possible to grade any lesion on a scale, such as: 0 = absent, 1 = minimal, 2 = slight, 3 = moderate, 4 = severe, 5 = very severe; or, 0 = absent, 1 = benign, 2 = malignant.

In the case of grading, a fixed, limited scale is applied to all animals, whereas for multiple tumour counting (Section 7.3) there is, at least in principle, no limitation on the number of tumours which could be observed.

Snedecor and Cochran (1980, pp. 146–148) have suggested that comparison of graded data from two groups may be carried out using Fisher's randomization test with small numbers and a continuity-corrected t-test with larger numbers. In the latter case, one could alternatively use the nonparametric techniques discussed in Section 8.3. Application of ranking procedures to graded responses can lead to a large number of ties, but it has been stated that this may not be crucial (Conover, 1980, p. 232).

To illustrate the suggestion of Snedecor and Cochran, consider the simple, fictitious experimental outcome (Table 7.1) of a control group of sample size six and a dose group of sample size four, graded on a three-point scale, 0, 1, 2.

Table 7.1 Example to illustrate Fisher's randomization test

n_i	Control group (6)	Dose group (4)
(x_{ij})	(0, 0, 0, 0, 1, 1)	(0, 1, 1, 2)
$x_{i.}$	2	4
$\bar{x}_{i.}$	$\frac{2}{6} = \frac{1}{3}$	$\frac{4}{4} = 1$

The randomization test is based on the fact that there are $(n_0 + n_1)!/(n_0!n_1!)$ possible divisions of the $n_0 + n_1$ animals into groups of n_0 and n_1. We also wish to find the number of such possible outcomes for which $\bar{x}_{1.} - \bar{x}_{0.}$ matches or exceeds the observed value. Equivalently, this is the number of outcomes for which $x_{1.}$ matches or exceeds the observed value, for example, 4. In this particular case, the number of outcomes is $10!/(4!6!) = 210$, the observed $\bar{x}_{1.} - \bar{x}_{0.} = \frac{2}{3}$ and $x_{1.} = 4$. Consider the numbers of ways in which $x_{1.} \geq 4$. These are listed in Table 7.2.

Thus, there are in total $30 + 1 + 4 = 35$ combinations with $x_1 \geq 4$, and the exact one-tailed p-value is $p = 35/210 = \frac{1}{6}$. The corresponding t-test with $n_0 + n_1 - 2$ degrees of freedom is

$$t_c = (\bar{x}_{1.} - \bar{x}_{0.} - c)/\{s_p\sqrt{(1/n_0 + 1/n_1)}\},$$

where

$$s_p^2 = \left\{\sum_{j=1}^{n_1}(x_{1j} - \bar{x}_{1.})^2 + \sum_{j=1}^{n_0}(x_{0j} - \bar{x}_{0.})^2\right\}\Big/(n_0 + n_1 - 2)$$

Table 7.2 List of extreme outcomes for computation of randomization test

	(x_{1i})	Numbers of combinations
$x_{1.} = 4$	(0, 1, 1, 2)	$\binom{5}{1}\binom{4}{2}\binom{1}{1} = 30$
	(1, 1, 1, 1)	$\binom{5}{0}\binom{4}{4}\binom{1}{0} = 1$
$x_{1.} = 5$	(1, 1, 1, 2)	$\binom{5}{0}\binom{4}{3}\binom{1}{1} = 4$

and c is a continuity correction. The value of c is one-half of the absolute value of the difference between the observed $\bar{x}_{1.} - \bar{x}_{0.}$, and the next highest value of this statistic among all the possible randomizations. Thus, for this example, $\bar{x}_{1.} - \bar{x}_{0.} = \frac{4}{4} - \frac{2}{6} = \frac{2}{3}$, and the next highest value is $\frac{3}{4} - \frac{3}{6} = \frac{1}{4}$. Thus $c = (\frac{2}{3} - \frac{1}{4})/2 = \frac{5}{24}$. We also find that $s_p^2 = \frac{5}{12}$. These give the approximate t-value with eight degrees of freedom:

$$t_c = \frac{\frac{2}{3} - \frac{5}{24}}{\sqrt{\{(\frac{5}{12})(\frac{1}{4} + \frac{1}{6})\}}} = \frac{11}{10} = 1.1.$$

This approximation yields $p = 0.1517$, only slightly less than the exact result of $\frac{1}{6}$.

It should be noted that Snedecor and Cochran recommend that with such small samples with very little overlap the exact p is easily calculated and should be reported. They note that, with more extreme results, t_c may yield too small a p-value. Consider the most extreme possible outcome in our example, that is, $x_{1.} = 5$ or $\bar{x}_{1.} - \bar{x}_{0.} = \frac{5}{4} - \frac{1}{6} = \frac{13}{12}$. The exact one-tailed $p = 4/210 = 0.0190$. In this case,

$$t_c = \frac{\frac{13}{12} - \frac{5}{24}}{\sqrt{\{(\frac{10}{97})(\frac{1}{4} + \frac{1}{6})\}}} = 3.047,$$

which, with 8 degrees of freedom, gives $p = 0.0079$, less than one-half the exact p-value. Randomization tests for trend and associated approximate t-tests may also be performed when more than two groups are to be compared.

The particular nature of a graded response may also be taken into account by using methods which have been developed for the analysis of ordinal data (McCullagh, 1980). Let the response be graded in G categories, $g = 1, \ldots, G$, and let there be $I + 1$ experimental groups, represented by dose levels $d_0, d_1, \ldots, d_I$. As above, let n_{gi} be the number of lesions in group i $(i = 0, 1, \ldots, I)$ at grade g $(g = 1, \ldots G)$. From these data, one can estimate γ_{gi}, the probability of a lesion in group i being graded at or below level g. Note that $\gamma_{Gi} = 1$, so that there are only $G - 1$ essential estimates γ_{gi} $(g = 1, \ldots, G - 1)$ for each group.

McCullagh (1980) proposes finding a suitable transformation of γ_{gi} and investigating how the transformed values depend on the dose levels d_i. Two particular models have been proposed for this purpose. One model is

(i) the proportional odds model:

$$\log[\gamma_{gi}/(1-\gamma_{gi})] = \theta_g - \beta d_i \qquad (1 \leq g < G).$$

This means that for any two dose levels i_1 and i_2 the odds ratio

$$[\gamma_{gi_1}/(1-\gamma_{gi_1})]/[\gamma_{gi_2}/(1-\gamma_{gi_2})] = \exp \beta(d_{i_2} - d_{i_1})$$

is independent of the grade g and depends on the difference between the dose levels only. The parameters θ_g, which are nuisance parameters, and β, which is the essential parameter relating the graded response to the dose levels, can be estimated by maximum likelihood methods. Another model is

(ii) the proportional hazards model

$$\log[-\log(1-\gamma_{gi})] = \theta_g - \beta d_i \qquad (1 \leq g < G).$$

This model is related to the proportional hazards model formulated by Cox (1972) for the analysis of survival data. Again, maximum likelihood methods allow estimation of the parameters θ_g and β.

In general, any other monotone increasing function mapping the unit interval (0, 1) onto $(-\infty, \infty)$ can be used as a 'link'-function $l(\gamma)$ to postulate a model

$$l(\gamma_{gi}) = \theta_g - \beta d_i \qquad (1 \leq g < G).$$

The two examples given above, however, have the advantage of a straightforward interpretation of the parameters. This regression model can be generalized to allow for any set of covariates, not only one dose variable. The number of parameters will increase accordingly.

McCullagh has also pointed out that there is a direct theoretical relationship between his regression models for ordinal data and the nonparametric test discussed above, the latter, however, lacking simple descriptive parameters.

Graded response data can also be analysed in a stratified way, for example, when a tumour observed in an incidental context is graded. Generally, the grading is more severe in later time periods, and fewer animals from the higher-dose groups survive into the later intervals, due to toxicity. In such a case, the analysis could be stratified by time intervals, the θ_i's varying between time intervals, but the essential parameter β being the same for all.

7.5 Multifactorial designs: combining results

In the previous chapters, attention has centred on the design and analysis of experiments that test only one treatment of interest, often at different dose levels. Such one-factor experiments are commonly used to screen different exposures for their potential carcinogenicity. However, studies in which the experimental groups form a multiple-factor design are not uncommon. Sometimes, such designs are necessitated by practical reasons, so that, for example, the eight groups might form combinations of the main exposure of interest at four levels (control, low dose, middle dose, high dose) and two batches of animals, as it is impossible to obtain the number of animals required from one batch. On other occasions, there may be one main treatment of

interest, but one may wish to study simultaneously the effect of different methods of administering the treatment, for example, cigarette-smoke condensate at three dose levels dissolved in two alternative solvents. More interestingly, one may wish to study the effects of joint exposure to combinations of two or more carcinogens. It can be argued that such studies can often be more realistic than single carcinogen studies, since humans are frequently exposed to a variety of carcinogens simultaneously or in sequence. Often such studies are of value in investigating mechanisms of action; they also have direct public health implication, and they are often carried out where knowledge has already been accumulated about the dose-response of individual exposures.

In Chapter 3, the design of multifactorial experiments was briefly mentioned and some formal concepts outlined. For the purposes of this section, we limit ourselves to the study of two exposures, A and B, applied at all combinations of dose levels $a_0(=0), a_1, \ldots, a_i, \ldots, a_I$ of A and $b_0(=0), b_1, \ldots {}_{\Delta} b_j, \ldots, b_J$ of B so that there are $(I+1)(J+1)$ experimental groups. Such a design leads to two distinct questions:

(1) Given equivalent exposure to B, is the risk of tumour significantly related to exposure A?

(2) Is the joint effect of A and B different from what one would expect from the effect of A or B alone?

The first question is essentially aimed at avoiding bias due to the effect of B in assessing the effect of A, and of making a combined inference about A over the different levels of B. Seen in this light, A is the main exposure of interest, B being a secondary 'nuisance' variable that has to be standardized for.

There are two major techniques for answering this first question. Throughout Chapter 5, we have extensively described methods in which observed and expected values (as well as other statistics necessary for calculating significance levels of the observed/expected differences) from different time periods can be combined by accumulation. As long as the group structure remains the same in each stratum over which accumulation occurs, this method of combining can be used in an exactly analogous manner to combine results in dimensions other than time. Thus, in our example, we treat the data as consisting of $J+1$ subexperiments ('strata') defined by the levels of B. Each subexperiment has the same group structure ($I+1$ levels of A), and a combined result can be obtained in a straightforward manner. The same process, of course, can be used to make overall inferences for exposure B, adjusted for the effects of exposure A.

The second major technique for answering the first type of question would be applied where the response variable of interest can be related to the effects of the exposures A and B by a regression equation. For this purpose,we introduce the general notation that u_{ij} is the response in those animals exposed to level i of exposure A and level j of exposure B. For specific applications, the effect measure u_{ij} has to be defined very carefully, and this will have strong implications on the interpretation of the results. However, we discuss first the general concepts of absence or presence of interactions by denoting further μ to be an overall mean response, t_i a deviation from

the mean due to the ith level of A, and c_j a deviation from the mean due to the jth level of B (t_0 and c_0 are assumed to be zero to avoid overparameterization).

A test of the effect of A adjusted for the effect of B can be achieved by comparing the fit of the models

$$E(u_{ij}) = \mu + t_i + c_j$$

and

$$E(u_{ij}) = \mu + c_j,$$

where $E(u)$ denotes the expectation (i.e., mean) of u. This is, in general, a more appropriate test for the effect of A than the comparison of the two simpler models both ignoring c_j. The test recommended, on I degrees of freedom, is a test of overall variation in response with level of exposure A. It is of course possible to test for a linear effect of A by replacing t_i in the above formulation by a term γd_i, where γ is a parameter to be estimated and d_i is the dose applied at level i, although the full set of s parameters, representing effects of the confounding variable B, should be retained.

Implicit in both techniques for answering the first type of question is the no-interaction assumption, that is, that the effect of A does not vary significantly according to level of B. An overall conclusion that exposure A slightly increases tumour risk might be misleading, for example, if it considerably increased risk at high doses of B, while reducing risk somewhat at low doses. Statistical tests for interaction of the effects of the stratifying variable with those of the main variable of interest, when analysing stratified contingency tables, are given by Breslow and Day (1980). With the regression equation, a test of no interaction can be achieved by comparing the fit of the model

$$E(u_{ij}) = \mu + t_i + c_j$$

with that of the model

$$E(u_{ij}) = \mu + x_{ij},$$

where x_{ij} represent effects of each combination of treatments (x_{00} is assumed to be zero to avoid overparameterization).

In the second question, the interest is in the joint effect of both exposures. This, of course, is related to the test of no interaction, as lack of interaction implies in a sense that the joint effect of A and B does not differ from that of A or B, alone, since the data are well described by the model

$$E(u_{ij}) = \mu + t_i + c_j.$$

It is important to repeat that the particular type of model depends strongly on the scale on which the effects u_{ij} are measured. To consider this further let us for the moment take the response variable in experimental group (i, j) to be the probability of developing a tumour p_{ij}, which is estimated by the proportion of tumour-bearing animals. p_{i0} $(i = 0, 1, \ldots I)$ and p_{0j} $(j = 0, 1, \ldots J)$ then denote the probabilities of the dose-response patterns for the individual exposures.

There are two basic possibilities for modelling the response probability p_{ij} in the combination groups exposed to both A and B.

(i) In the first, *the additive model,* we consider the absolute increase over the background response probability, measured in the untreated control group, as the quantity which describes the effect of exposure, and assume that, in a group treated with both exposures, this effect should be the sum of the effects of both the respective individual effects

$$p_{ij} - p_{00} = (p_{i0} - p_{00}) + (p_{0j} - p_{00})$$

or

$$p_{ij} = p_{i0} + p_{0j} - p_{00}$$

for i and $j > 1$, p_{ij} being taken as 1 if the right-hand side of the second equation exceeds 1.

(ii) In the second, *the multiplicative model,* the effect of an exposure is measured by the proportional increase over the background response, so that in a combination group the effect should be equal to the product of the effects of individual exposures

$$p_{ij}/p_{00} = (p_{i0}/p_{00})(p_{0j}/p_{00})$$

for i and $j > 1$. This is also equivalent to

$$p_{ij}/p_{0j} = p_{ik}/p_{0k}$$

for $j, k = 0, 1 \ldots, J$ $(j \neq k)$, which means that the effects attributed to exposure A indicated by the subscript i are the same at any level j or k of exposure B.

The two models introduced above represent simple statistical models based on the choice of different effect measures. More refined models incorporating mechanistic considerations, usually with reference to the multistage action of the carcinogenic process, have been proposed (for example, Siemiatycki & Thomas, 1981). Note that, under a multistage hypothesis, one would normally expect two carcinogens that act on different stages to act multiplicatively, whereas two carcinogens acting on the same stage might act additively.

The additive and multiplicative models outlined above have the advantage that they can also be formulated in terms of relative risks. Let $R_{ij} = p_{ij}/p_{00}$ $[(i, j) \neq (0, 0)]$ be the relative risk of group (i, j) compared with the control group, where neither exposure is present. The additive model then predicts that $R_{ij} = R_{i0} + R_{0j} - 1$, whereas the multiplicative model predicts that $R_{ij} = R_{i0}R_{0j}$.

Formulating these models in terms of relative risks enables utilization of the basic methods for the analysis of long-term animal experiments described in Chapter 5. These methods, which account for intercurrent mortality and consider the context of observation of tumours, allow one to describe the differences in tumour yield between two or several groups in terms of relative risks. We illustrate this by an example, based on our study of all combinations of $I + 1$ levels of A and $J + 1$ levels of B, in which we investigate the multiplicative model for the joint action of the two exposures.

If a_i is a fixed level of exposure A, then the groups receiving the combination $(a_i, b_0) \cdots (a_i, b_J)$ form a subexperiment which can be analysed by the methods described in Chapter 5. The particular analysis may depend on the context of observation of the tumours and on whether time to death is available, but in all cases

one should calculate for each level of factor B an expected number of tumour-bearing animals E_{a_ij} which can be contrasted with the respective observed numbers O_{ij}.

$$R_{a_ij} = (O_{ij}/E_{a_ij})/(O_{i0}/E_{a_i0})$$

then represent the relative risks at the different dose levels of factor B with reference to the baseline $b_0 = 0$. Such an analysis can be carried out at all $r + 1$ levels of factor A, and when the effect of factor B is the same at all these levels, it is justified to summarize it by adding the O's and E's over the different levels of factor A.

$$O_{.j} = \sum_{i=0}^{I} O_{ij} \quad \text{and} \quad E_{.j} = \sum_{i=0}^{I} E_{a_ij}$$

then denote the so-derived summary values, and the relative risks at the different levels of factor B, averaged over factor A, would then be

$$R_{.j} = (O_{.j}/E_{.j})/(O_{.0}/E_{.0}).$$

This process of averaging the effect of one factor over the different levels of the other assumes that the relative risks are the same, irrespective of which level of factor A is considered. This means that the assumption of a multiplicative model (as formulated above) is made implicitly.

In exactly the same way as we have summarized the effect of factor B averaged over factor A, we can derive a summary description of the effect of factor A averaged over factor B. This yields the observed and expected numbers $O_{i.}$ and $E_{i.}$ and hence the summary relative risks

$$R_{i.} = (O_{i.}/E_{i.})/(O_{0.}/E_{0.}).$$

Finally, we can also calculate the relative risks of developing a tumour in any of the single combination groups compared to the untreated control group. This is done by conducting an analysis only with the particular group of interest (a_i, b_j) and the untreated control group (a_0, b_0). If we denote the observed and expected numbers in this analysis as O_{ij}, O_{00}, E_{ij} and $E_{00}^{(i,j)}$, the resulting relative risks are then

$$R_{ij} = (O_{ij}/E_{ij})/(O_{00}/E_{00}^{(i,j)}).$$

Under the multiplicative model, one would expect to observe that

$$R_{ij} = R_{i.}R_{.j}.$$

This can be inspected in an informal way, with the calculated relative risks. Either the model is reasonably fulfilled for all $r \cdot s$ combination groups, or the pattern of deviation will provide an indication of whether the multiplicative model applied fits the data or not. It should be made clear that this approach does not represent a full-scale fitting of a statistical model, as the random variation behind the relative risk estimates is not considered in this descriptive approach. Also, the method is likely to be useful only when the observed number of tumours in the untreated control group is not too small, otherwise the variation in relative risk will be very large. However, this descriptive approach can give an indication of possible underlying models (Métivier *et al.*, 1984). It can be used to investigate whether a multiplicative model for relative risks of life-time

development of tumours can explain the joint effect. Adjustment for intercurrent mortality and context of observation is achieved by utilizing the methods given in Section 5.7.

For the analysis of observable tumours or rapidly lethal tumours, survival methods can be applied, and the probability of tumour-free survival to the end of the experiment may be a relevant endpoint to consider. A multiplicative model for this effect measure would imply that the age-specific hazard in a combination group is the sum of the age-specific hazards of the respective single-exposure groups. This was shown by Wahrendorf *et al.* (1981), and Korn and Liu (1983) proposed likelihood ratio tests for this purpose.

To assume that the age-specific hazard rates are multiplicatively related in the absence of an interaction has been used in the framework of Weibull models (see Section 6.3). Assuming common values for w and κ, the parameters β_{ij} fitted for each group can be viewed as relative hazard rate parameters. A formal test of the multiplicative model for the hazards can be achieved by using maximum likelihood methods (Peto & Lee, 1973) to compare the models $\log \beta_{ij} = \mu + t_i + c_j$, which may be called the multiplicative main effects model, and $\log \beta_{ij} = \mu + x_{ij}$, which is a saturated model. This results in a chi-squared statistic on IJ degrees of freedom testing the overall fit of the multiplicative main effects model. Departures from the fit can be investigated further by comparing the observed number of animals with tumours in each group O_{ij} with that expected under the first model

$$\hat{E}_{ij} = \hat{\beta}_{ij} v_{ij},$$

where v is defined in Section 6.3.

An analogous analysis of time to tumour or death from tumour can be based on the proportional hazards model (Section 6.3). In this case one would assume that the hazard function satisfies the equation

$$\lambda_{ij}(t) = \exp(\Delta_{ij})\lambda_0(t)$$

for (i, j), where postulating $\Delta_{ij} = \mu + t_i + c_j$ would lead to a multiplicative model for the hazard functions, and where analogous comparisons between the multiplicative main effects model and the saturated model can be performed. More formal tests of the multiplicative model can be carried out in a straightforward manner when time to death can be ignored and the data can be expressed as simple counts of tumour-bearing animals in a $2 \times (I + 1) \times (J + 1)$ contingency table (Bishop *et al.*, 1974; Baker & Nelder, 1978).

Finally, it should be noted that studying quantal responses to mixtures of drugs has a long tradition in investigations of acute toxic effects. The different models considered in this framework have been reviewed by Hewlett and Plackett (1979, Chapter 7).

7.6 Litter effects

The statistical methods discussed in Chapter 5 are based on the assumption that the responses for different animals are statistically independent. As noted in Chapter 3, however, the assumption of independence may be violated with experimental designs

involving the use of littermates, since pups within the same litter may tend to respond similarly (Gaylor *et al.*, 1985b). There are three main possibilities for distributing littermates among the experimental groups. First, animals from each complete litter may be assigned to the same group, as is necessary in two-generation experiments when the parental generation is exposed to different doses of a substance. Second, littermates may be distributed among different treatment groups in blocked designs with blocks defined in terms of equal-sized litters. Third, animals may be allocated to different treatment groups regardless of litter membership as in the completely randomized design.

In the first two cases, special methods of statistical analysis are required. In the first situation, it is necessary to take into account the within-litter correlation by using the variation between litters rather than variation between animals as the basis for between-group comparisons. Below, we discuss the statistical consequences of this in general terms and outline specific methods of analysis that may be used with such data. The third case, in which littermates are distributed across experimental groups, does not have as great an impact on the statistical inference and will be discussed at the end of this section.

Complete litters assigned to different groups

In the presence of positive intralitter correlation, standard methods of statistical analysis which ignore the litter structure will tend to underestimate the standard error of the difference between the overall response rates in two different treatment groups and hence overstate the statistical significance of any observed differences (Haseman & Hogan, 1975). Gladen (1979) showed that the use of standard chi-squared or likelihood ratio tests which ignore intralitter correlation can result in inflated false-positive rates. In this regard, Rao and Scott (1981) have shown that the correct asymptotic null distribution of the usual chi-squared statistic for comparing several treatment groups is in fact a weighted sum of independent chi-square random variables with weights related to the intralitter correlation within each group. The use of the Fisher–Irwin exact test for comparing two treatment groups, ignoring the litter structure, has also been shown to result in somewhat inflated false-positive rates in the presence of positive intralitter correlation. With negative intralitter correlation, however, the standard tests would be valid in the sense that the actual false-positive rate would tend to be less than the nominal rate.

In order to avoid these problems, statistical methods which take litter structure into account should be employed. Let n_{ij} denote the size of the jth litter in the ith treatment group ($j = 1, \ldots, m_i; i = 1, \ldots, I$), and let x_{ij} denote the number of these animals developing tumours. Conditional on the n_{ij}, Cochran (1943) assumed that the sample proportions $\hat{p}_{ij} = x_{ij}/n_{ij}$ have mean

$$E(\hat{p}_{ij} \mid p_{ij}) = p_{ij}$$

and variance

$$V(\hat{p}_{ij} \mid p_{ij}) = p_{ij}(1 - p_{ij})/n_{ij},$$

where the p_{ij} are considered to be held constant. If the p_{ij} are actually independent

random variables with mean

$$E(p_{ij}) = \mu_i$$

and variance

$$V(p_{ij}) = \sigma_i^2 > 0,$$

we have

$$E(\hat{p}_{ij}) = \mu_i$$

and

$$V(\hat{p}_{ij}) = \frac{\mu_i(1-\mu_i)}{n_{ij}} + \sigma_i^2\left(1 - \frac{1}{n_{ij}}\right).$$

Note that the first term in the above formula represents the variation that would occur in the absence of any litter effects, whereas the second term reflects the between-litter variation associated with such effects. Cochran proposed a weighted analysis of variance of the observed proportions $\hat{p}_{ij}$ as a means of assessing treatment differences, with estimates of the litter-specific variances σ_{ij}^2 used to obtain the weights (see also Kleinman, 1973).

Gladen (1979) used a more general model with $V(\hat{p}_{ij}) = f(n_{ij})$ for some general function f. Although the natural estimator

$$\hat{p}_i = x_{i.}/n_{i.} = \sum_{j=1}^{m_i} x_{ij} \Big/ \sum_{j=1}^{m_i} n_{ij}$$

of μ_i is unbiased, Gladen proposed the unbiased jackknife estimator

$$\hat{p}_{Ji} = m_i\hat{p}_i - \frac{(m_i - 1)}{m_i}\sum_{j=1}^{m_i} \hat{p}_{i(-j)},$$

where $\hat{p}_{i(-j)} = (x_{i.} - x_{ij})/(n_{i.} - n_{ij})$ denotes the estimator of μ_i omitting the jth litter in group i. The jackknife estimator of the variance of each $\hat{p}_{Ji}$ is given by

$$v_{Ji} = \frac{m_i - 1}{m_i}\sum_{j=1}^{m_i}\left[\hat{p}_{i(-j)} - \frac{1}{m_i}\sum_{j=1}^{m_i}\hat{p}_{i(-j)}\right]^2.$$

To compare two groups with m_1 and m_2 litters, respectively, the statistic

$$t_J = \frac{\hat{p}_{J1} - \hat{p}_{J2}}{(v_{J1} + v_{J2})^{\frac{1}{2}}}$$

will then approximate a t-distribution with $m_1 + m_2 - 2$ degrees of freedom under the null hypothesis $H_0: \mu_1 = \mu_2$, provided m_1 and m_2 are sufficiently large.

Frangos and Stone (1984) investigated, among other approaches, the estimator

$$\bar{p}_i = \sum_{j=1}^{m_i} \hat{p}_{ij}/m_i,$$

which is the average of the litter-specific proportions in group i. The variance of $\bar{p}_i$ can be estimated by

$$\bar{v}_i = \sum_{j=1}^{m_i} (\hat{p}_{ij} - \bar{p}_i)^2/\{m_i(m_i - 1)\}.$$

For comparison of two groups with m_1 and m_2 litters, this would lead to a standardized test statistic

$$z = (\bar{p}_1 - \bar{p}_2)/(\bar{v}_1 + \bar{v}_2)^{\frac{1}{2}},$$

which asymptotically follows a standard normal distribution. The small sample behaviour of z has not been investigated. For a single group, however, Frangos and Stone (1984) demonstrated that confidence intervals based on $\bar{p}_i$ and on a modification of an estimator proposed by Southward and Van Ryzin (1972) outperform the jackknife confidence intervals.

Another nonparametric technique has been considered by Haseman and Soares (1976). In particular, they showed that comparing the litter-specific proportions in two treatment groups using the nonparametric Wilcoxon test (Hollander & Wolfe, 1973, p. 68) is in many circumstances a sufficiently accurate and powerful statistical procedure.

A more parametric approach to modelling litter effects was used by Williams (1975), who assumed that p_{ij} follows a beta distribution with parameters $\alpha_i, \beta_i > 0$. In this case, $\mu_i = \alpha_i/(\alpha_i + \beta_i)$ and $\sigma_i^2 = \mu_i(1-\mu_i)\rho_i/(1+\rho_i)$, where $\rho_i = (\alpha_i + \beta_i)^{-1} > 0$ provides a measure of the degree of association between litter mates. Williams proposed the use of beta-binomial likelihood ratio methods to test the null hypothesis $H_0: \mu_1 = \mu_2$; $\rho_1 = \rho_2$ against a general alternative. Williams also considered $H_0: \mu_1 = \mu_2 = \mu$ with $\rho_1 = \rho_2 = \rho$ fixed. Based on a simulation study, Shirley and Hickling (1981) concluded that, for typically encountered litter sizes, the nonparametric Wilcoxon test was preferable to the beta-binomial likelihood ratio test. (See also Haseman & Kupper, 1979.)

A different approach has been suggested by Kupper and Haseman (1978). They assume that 'fetuses in the same litter tend to have an inherent relationship to one another.' Thus, the assumption of mutual independence of the outcomes within a litter, which usually leads to a binomial within-litter model, is altered by applying a correction factor which depends on the covariance between two Bernoulli trials within one litter. Kupper and Haseman (1978) demonstrate that this correlated binomial model is in a sense an extension of the beta-binomial model, as discussed by Williams (1975), in that it also allows, to some degree, negative correlations between responses within a litter. A likelihood ratio test is again proposed to test for differences between the experimental groups. For one example, Kupper and Haseman demonstrated a better fit of their correlated binomial model than of the beta-binomial model.

All of the procedures proposed above are based on asymptotic approximation and rely on large m_i for their validity. This is particularly important when the response probabilities μ_i are near zero or one. Because many rodent lesions occur with frequencies of 1% or lower, exact permutation tests of the null hypothesis $H_0: \mu_1 = \mu_2$ against the alternative $H_2: \mu_1 > \mu_2$ based on the observed difference $d = \hat{p}_1 - \hat{p}_2$ may be considered (Crump & Howe, 1980). These tests are based on the fact that, under the null hypothesis, each of the $s = \binom{m_1 + m_2}{m_1}$ possible assignments of m_1 litters to group one and m_2 litters to group two are equally likely. The significance level for the exact randomization test against the alternative $H_1: \mu_1 > \mu_2$ is then given by r/s, where r is the number of permutations leading to a value of d at least as large as the observed

value. When s is large, Crump and Howe suggest that the significance level may be estimated on the basis of a random sample from the permutation distribution. In the special case $n_{ij} \equiv n$, the algorithm given by Soms (1977) may be used to obtain the randomization significance level. Although these procedures are exact, they are conservative for small values of μ in the sense that the false-positive rate can be notably less than the nominal level (Krewski *et al.*, 1984a).

Several approaches to modelling dose-response data which take into account intralitter correlation have also been proposed. Under the beta-binomial model considered by Williams (1975), the marginal distribution of x_{ij} (the number of animals developing tumours in the jth litter in the ith treatment group) is a beta-binomial, so that the likelihood for the parameters $\mu_i = \alpha_i/(\alpha_i + \beta_i)$ and $\rho_i = (\alpha_i + \beta_i)^{-1} > 0$ $(i = 1, \ldots, I)$ is a product of I independent beta-binomial terms. Segreti and Munson (1981) then proposed that the effects in dose d_i administered to the ith group be modelled as

$$\mu_i = \lambda + (1 - \lambda)F(\alpha + \beta \log d_i),$$

where $0 < \lambda < 1$ and $\beta > 0$ as in (6.2) and (6.8). The beta-binomial likelihood may then be used to obtain estimates of the parameters α, β and λ in the presence of dose-specific litter effect parameters $\rho_1, \ldots, \rho_I$. (Segreti and Munson also consider a simpler but less realistic model in which $\rho_1 = \cdots = \rho_I = \rho$.) The former three estimates then provide a fitted dose-response curve

$$\hat{\mu} = \hat{\lambda} + (1 - \hat{\lambda})F(\hat{\alpha} + \hat{\beta} \log d).$$

Another approach to this problem has been studied by Ochi and Prentice (1984). In general terms, they consider a correlated probit regression model in which the binary responses within the same litter are defined as indicators of whether or not the corresponding components of a multivariate normal regression vector with common mean and variance exceed some threshold value. Although the likelihood calculations are somewhat more complex than in the Segreti–Munson model, the Ochi–Prentice model provides for multiple covariates as well as flexibility in modelling changes in intralitter correlation with dose.

Litters distributed across groups in blocked designs

The situation in which littermates are distributed across litters has been investigated by Mantel *et al.* (1977) and by Mantel and Ciminera (1979). For a discussion of this issue see also Mantel (1980). Basically, they suggest comparing the tumour incidence in different treatment groups by stratifying over the litters. This could be done with the methods discussed in Chapter 5. However, as litters are usually not very large, it may be that, towards the end of an experiment in certain litters, animals in only one experimental group are at risk, with no surviving littermates for comparison. Thus, the information from these animals may remain unused. To avoid this, Mantel *et al.* (1977) suggest several special devices for combining remaining animals into new strata. Mantel and Ciminera (1979) outline a different approach, in which so-called 'Savage-

scores' are assigned to each animal, irrespective of litter and group, based on when or whether the animal developed a tumour. Using these scores, which are common in life-table analyses, litter-adjusted comparisons between control and treated groups were proposed.

Problems with this approach were pointed out by Michalek and Mihalko (1983), with a discussion by Mantel (1983). It was demonstrated that the attempt to utilize remaining information can confound litter and treatment effects. It should also be noted that use of the hypergeometric variance, as in Chapter 5 [formula (5.2)], is preferable for stratified comparisons to the permutational variance used by Michalek and Mihalko (1983) because the permutational variance is invalid when treatment influences mortality. Therefore, a simple litter-stratified analysis, as outlined by Mantel *et al.* (1977) and Michalek and Mihalko (1984), but without recovery of interlitter information, appears to be the most advisable approach. In any case, complete randomization of experimental animals into all the experimental groups, irrespective of their litter membership, represents the preferable experimental design.

7.7 Association among tumour types

It is obvious from the preceding chapters and sections that several tumour types are investigated in a long-term animal experiment. Statistical analysis is usually performed for each of these tumour types individually. This may lead to problems of multiple comparisons in making statistical inferences from the study, as discussed in Section 7.2. The association of a given tumour type with another represents in this context a nuisance factor which is manifested in the intercurrent mortality. Methods accounting for this are discussed at length in Chapter 5.

However, the association among tumour types also represents an interesting aspect of studies on the mechanism of action of the exposure in the entire biological system (for example, animal) investigated. For the moment we shall neglect the role of treatment and consider only animals from one group. Looking at two tumour types, A and B, say, one could define the association between these tumour types by the odds ratio in the resulting 2×2 table if one categorizes the animals according to the occurrence of the two tumours of interest:

		Tumour A: absent	Tumour A: present
Tumour B	absent	a	b
	present	c	d

The association would be defined as $\hat{\psi} = (ad)/(bc)$, and standard statistical methods for odds ratios (see Chapter 5) could be employed. However, before doing so, careful attention must be given to the way the tumours have been found in the animals. The intercurrent mortality, probably influenced by the presence of one or both of the two tumours, plays a crucial role. Consider the simple model of Breslow *et al.* (1974).

Assume X_A and X_B to be the times of clinical onset of tumours A and B, defined operationally as the earliest time the tumours would be detected by necropsy. Let Y_A and Y_B be the times from onset until death from tumours A or B, and Z the time of death due to an unrelated cause, including serial sacrifice. An animal would be classified in one of the cells of the above 2×2 table if

(a) $Z < \min(X_A, X_B)$: neither tumour present at necropsy;
(b) $X_A < X_B$ and $\min(Z, X_A + Y_A) < X_B$: only tumour A present;
(c) $X_B < X_A$ and $\min(Z, X_B + Y_B) < X_A$: only tumour B present;
(d) $\max(X_A, X_B) < \min(X_A + Y_A, X_B + Y_B, Z)$: both tumours present.

Assume that there is no association between tumours A and B, that is, X_A and X_B are independent random variables, but one tumour, say A, is rapidly lethal, that is, Y_A is very small. Then animals with both tumours present are very unlikely to be found among those dying. Based on such data the estimated association would appear to be unjustifiably negative, that is, $\hat{\psi} < 1$. However, if one constructed a 2×2 table, as above, on the basis only of animals which died from causes not related to the tumour(s), such as by serial sacrifice, the rapid lethality of one tumour would lead to the general finding of a low proportion of animals with this tumour, either alone or in combination with the other tumour. Therefore, the resulting estimate of the association should not have a systematic error.

The method for the analysis of carcinogenicity data proposed by Turnbull and Mitchell (1978) and Mitchell and Turnbull (1979), as already discussed in Section 6.3, includes prevalence models. Tumour prevalence can be observed directly only by serial sacrifice of animals. Therefore, such designs are needed to obtain an unbiased assessment of the association among tumour types. The log-linear prevalence model (see Section 6.3) can depend on various factors. These factors may include aspects of the experimental design, such as treatment group or time period of scheduled sacrifice, but also relate to the presence or absence of certain tumours. The resulting interaction terms between different tumour types can be used to evaluate possible associations. In addition, the interaction terms of each individual tumour with the treatment group will indicate whether the occurrence of this tumour depends on the treatment of the animals. Also, the interaction terms with the time periods will give indications of the temporal pattern of the tumour prevalence. This log-linear model for prevalence is combined with a logistic model for lethality which potentially depends on the same set of factors and interactions, but for which a different subset may prove to be significant. Interpretation of the results from both models has to be made jointly and the results have to be checked carefully for biological consistency.

In the search for jointly best-fitting prevalance and lethality models, the same approaches used for the fitting of multiplicative models for discrete data (Bishop *et al.*, 1974) can be applied. The operational criteria by which interaction terms are successively inserted or deleted may vary from application to application. Usually, likelihood ratio statistics are employed to judge whether a significant change in the goodness of fit is observed when altering the model in one direction or another. An example using this approach, but also including consideration of the simple 2×2 tables above, has been given by Wahrendorf (1983). Data from a long-term carcinogenicity

study with DDT using CF-1 mice were used. In this study, which was originally reported by Tomatis *et al.* (1974), mice were fed 250 ppm of DDT for 15 or 30 weeks, after which exposure ceased. An untreated control group was also used. Scheduled sacrifices were conducted at 15, 30, 65, 95 and 120 weeks of exposure, though not equally frequently in all groups. Consequently, only two time intervals were available to define the corresponding factor in the prevalence and lethality model.

Three tumour types were investigated: lymphomas, liver and lung tumours. In the prevalance model, liver tumours showed a significant interaction with the factor treatment group, demonstrating the well-known hepatocarcinogenic effect of DDT. The prevalence of all three tumours showed a significant interaction with time, indicating increased occurrence of all tumours in the later time interval. However, there also remained significant negative interaction terms between lymphomas and liver tumours and between lymphomas and lung tumours. These were inspected further by calculating coefficients of association in a simple 2×2 table contrasting two tumours. Such tables were derived by using only those animals which were sacrificed in each group and each time interval. This showed a consistent pattern of negative association between lymphomas and liver tumours. Counteracting this negative association, by including among those animals with both tumours also those who died naturally, did not alter this conclusion. As the prevalence of lymphomas was not related to treatment group, it was concluded that the hepatocarcinogenic activity of DDT may have an influence on the development of lymphomas in CF-1 mice.

7.8 Historical control tumour rates

In the evaluation of a chemical carcinogenesis experiment, knowledge of the spontaneous tumour rates obtained from control groups of previous experiments can often provide insight into the possible carcinogenicity of a test compound (Gart *et al.*, 1979; Tarone *et al.*, 1981; Haseman, 1983a). The most appropriate and important comparison of an exposed group is with the control group randomized from the same source. However, historical control tumour rates can be helpful in evaluating experiments for which the statistical analysis based on matched control tumour rates indicates equivocal evidence of carcinogenicity. One situation in which historical rates are likely to be particularly helpful is in the evaluation of small nonsignificant tumour increases at tissue sites with very low spontaneous tumour rates. When historical control rates are used to evaluate an equivocal experiment, both the magnitude and variability of these rates must be considered (Tarone *et al.*, 1981; Haseman, 1983a).

Although informal, ad-hoc comparisons with historical control data can often provide some insight into the carcinogenic potential of a test chemical (Fears *et al.*, 1977; Tarone *et al.*, 1981), methods have recently been developed which permit the incorporation of historical control information in a formal framework. Tarone (1982) modelled historical control rates using a beta-binomial model and derived a test for dose-related trends which is a modification of the Cochran–Armitage test. The modification depends both on the magnitude and variability of the historical rates. Hoel (1983) proposed an exact test based on the beta-binomial model. When the parameters of the beta-binomial distribution are known, Hoel's exact test is valid, and

Tarone's test is asymptotically valid (Krewski *et al.*, 1985; Hoel & Yanagawa, 1986). Problems arise, however, when the parameters must be estimated from the available historical data (Tamura & Young, 1986). Bias in the estimates of the beta-binomial parameters causes the methods to give too much weight to the historical control data. Thus, methods based on the beta-binomial model should be used with caution until unbiased estimators of the beta-binomial parameters are developed.

Dempster *et al.* (1983) assume that the logits of the historical control rates are normally distributed, and evaluate the evidence of a dose-response relationship using a Bayesian analysis, again incorporating information about the magnitude and variability of the historical rates. Dempster *et al.* also discuss diagnostic methods to assess the sensitivity of their analysis to different prior distributions and to assess the goodness of fit of the various models (including the beta-binomial model). The small-sample performance of the method of Dempster *et al.* has not been investigated, but, because of the tractability of estimation procedures for normal models, it is unlikely that their method will share the problems associated with those based on the beta-binomial model.

In making a formal analysis based on historical data, care must be taken to ensure that the historical control rates used in the analysis come from experiments which are similar to the current experiment in factors known to affect the magnitude of spontaneous tumour rates. Such factors may include the length of time on study, housing conditions, type of food, and possibly the year of birth of the test animals (Gart *et al.*, 1979; Tarone, 1982; Haseman, 1983a). Certainly, some initial screening is necessary to determine which historical rates may be used in the analysis of a particular experiment. In cases where the historical control data are informative with respect to the concurrent control response rate, their use may greatly strengthen the inferences made concerning the hypothesis of carcinogenicity. In contrast to their value in hypothesis testing, however, historical control data seem to provide little additional information when modelling dose-response relationships (Smythe *et al.*, 1986).