5. NONPARAMETRIC METHODS FOR ANIMAL CARCINOGENESIS EXPERIMENTS

CHAPTER 5

NONPARAMETRIC METHODS FOR ANIMAL CARCINOGENESIS EXPERIMENTS

5.1 Introduction

This chapter examines the statistical evaluation of an animal carcinogenesis experiment with the goal of determining whether or not a test compound induces tumours. In most of the chapter, it is assumed that the animals were assigned to different exposure groups in a completely randomized design. Each group received a different dose level of a test compound or served as a control group, and the animals were examined for the presence of tumours either continuously (for observable tumours) or at necropsy (for occult tumours). As noted in Chapter 2, an evaluation of tumour occurrence data requires the examination of mortality patterns in the various groups. Accordingly, Section 5.3 describes the computation of nonparametric survival functions and nonparametric test statistics, which permit a comparison of mortality patterns among the different exposure groups; Section 5.4 describes methods for comparing the crude tumour rates of the different groups; Section 5.5 describes the method of Hoel and Walburg (1972) and other methods for nonfatal tumours; Section 5.6 describes the use of failure-time methods to analyse tumour incidence data for observable tumours or tumour mortality data for rapidly lethal tumours; and Section 5.7 discusses the method of Peto for analysing tumour data in which the context of observation of each tumour is known and tumours are observed in both the fatal and incidental contexts (Peto, 1974; Peto *et al.*, 1980). Many of the nonparametric test statistics presented in Sections 5.3 to 5.7 are closely related in functional form. Thus, Section 5.2 presents technical details common to the computation of nonparametric test statistics discussed in Sections 5.3–5.7.

5.2 Computation of nonparametric test statistics

Suppose that the animals have been randomized into $I+1$ experimental groups, and that the animals in the ith group are exposed to a dose level d_i of a test compound, with $d_0 < d_1 < \cdots < d_I$, for $i = 0, 1, \ldots, I$. Often, the group indexed by 0 will be a control group, with $d_0 = 0$. Suppose that observations of the experimental endpoint of interest (e.g., death or occurrence of a tumour) are made at K distinct times t_k, $k = 1, 2, \ldots, K$. The data corresponding to each experimental endpoint may be summarized in K $2 \times (I+1)$ contingency tables ($K \geq 1$). The kth contingency table takes the form of Table 5.1, where x_{ik} denotes the number of events (e.g., deaths or

Table 5.1 Summary of data corresponding to a particular experimental endpoint for all animals in risk set k

Dose level	d_0	d_1	...	d_i	...	d_I	Total
No. of events	x_{0k}	x_{1k}	...	x_{ik}	...	x_{Ik}	$x_{.k}$
No. of animals at risk	n_{0k}	n_{1k}	...	n_{ik}	...	n_{Ik}	$n_{.k}$

animals with tumour) observed in the ith group at t_k, and n_{ik} denotes the number of animals at risk in group i, for $k = 1, 2, \ldots, K$. The $n_{.k}$ animals for which data are summarized in the kth table will be referred to as risk set k. The definition of risk set will vary according to the experimental situation being considered.

The expected number of events in the ith exposure group for risk set k using indirect standardization is $E_{ik} = x_{.k}A_{ik}$, where $A_{ik} = n_{ik}/n_{.k}$. Thus, the observed and expected number of events in the ith group over the entire experiment are $O_i = \sum_{k=1}^{K} x_{ik}$ and $E_i = \sum_{k=1}^{K} E_{ik}$, respectively, for $i = 0, 1, \ldots, I$. Define

$$D_i = O_i - E_i = \sum_{k=1}^{K} (x_{ik} - E_{ik}) \tag{5.1}$$

and

$$V_{hi} = \sum_{k=1}^{K} \alpha_k A_{hk}(\delta_{hi} - A_{ik}) \tag{5.2}$$

where $\alpha_k = x_{.k}(n_{.k} - x_{.k})/(n_{.k} - 1)$ and δ_{hi} is defined as 1 if $h = i$ and 0 otherwise, for $h, i = 0, 1, \ldots, I$. Then, letting $\mathbf{D}' = (O_0 - E_0, O_1 - E_1, \ldots, O_I - E_I)$ be the vector of deviations of expected from observed values, and letting $\mathbf{V}$ be the $(I+1) \times (I+1)$ matrix with $(h+1, i+1)$ entry V_{hi}, a statistic to test for heterogeneity among the $I+1$ groups with respect to the rate of occurrence of the experimental endpoint in question may be calculated as

$$X_H^2 = \mathbf{D}'\mathbf{V}^{-}\mathbf{D}, \tag{5.3}$$

where $\mathbf{V}^{-}$ is a generalized inverse of $\mathbf{V}$. The statistic X_H^2 may be computed as $X_H^2 = \mathbf{D}_1'\mathbf{V}_1^{-1}\mathbf{D}_1$, where $\mathbf{D}_1$ is the vector of dimension I obtained by deleting $O_0 - E_0$ from the vector $\mathbf{D}$, and $\mathbf{V}_1$ is the $I \times I$ matrix of full rank obtained by deleting the first row and column of the matrix $\mathbf{V}$. If there is no difference among exposure groups with respect to the distribution of the occurrence of the endpoint in question, then X_H^2 will have an asymptotic chi-squared distribution with I degrees of freedom. A one-degree-of-freedom chi-squared test for an increasing or decreasing rate of occurrence of the endpoint in question with increasing dose level can be calculated as

$$X_T^2 = (\mathbf{d}'\mathbf{D})^2/(\mathbf{d}'\mathbf{V}\mathbf{d}), \tag{5.4}$$

where $\mathbf{d}' = (d_0, d_1, \ldots, d_I)$ is the vector of dose levels, and a test for departure from a monotone dose-response relationship can be based on

$$X_Q^2 = X_H^2 - X_T^2, \tag{5.5}$$

which has a chi-squared distribution with $I-1$ degrees of freedom under the null hypothesis that the dose-response relationship is linear.

In computing the above statistics, the deviations of expected from observed values from different risk sets are given equal weight. It is sometimes of interest to weight certain risk sets more heavily than others. Accordingly, define

$$D_{iW} = \sum_{k=1}^{K} w_k(x_{ik} - E_{ik}) \tag{5.6}$$

and

$$V_{hiW} = \sum_{k=1}^{K} w_k^2 \alpha_k A_{hk}(\delta_{hi} - A_{ik}), \tag{5.7}$$

where α_k, A_{ik}, E_{ik}, and δ_{hi} are defined above, and the w_k are non-negative weights. Then, letting $\mathbf{D}'_W$ denote the vector $(D_{0W}, D_{1W}, \ldots, D_{IW})$ and $\mathbf{V}_W$ denote the $(I+1)\times(I+1)$ matrix with $(h+1, i+1)$ entry V_{hiW}, a weighted statistic to test for heterogeneity among the $I+1$ groups with respect to the rate of occurrence of the experimental endpoint in question may be calculated as

$$X^2_{WH} = \mathbf{D}'_W \mathbf{V}_W^{-} \mathbf{D}_W, \tag{5.8}$$

where $\mathbf{V}_W^{-}$ is a generalized inverse of $\mathbf{V}_W$. The statistic X^2_{WH} may be computed as $X^2_{WH} = \mathbf{D}'_{W1}\mathbf{V}_{W1}^{-1}\mathbf{D}_{W1}$, where $\mathbf{D}_{W1}$ is obtained by deleting D_{0W} from $\mathbf{D}_W$ and $\mathbf{V}_{W1}$ is obtained by deleting the first row and column of $\mathbf{V}_W$. If there is no difference among exposure groups with respect to the distribution of the occurrence of the endpoint in question, and if the weights are chosen properly, then X^2_{WH} will have an asymptotic chi-squared distribution with I degrees of freedom. A one-degree-of-freedom test for an increasing or decreasing rate of occurrence of the endpoint in question with increasing dose level can be calculated as

$$X^2_{WT} = (\mathbf{d}'\mathbf{D}_W)^2/(\mathbf{d}'\mathbf{V}_W\mathbf{d}), \tag{5.9}$$

and a test for departure from a monotone dose-response relationship can be based on

$$X^2_{WQ} = X^2_{WH} - X^2_{WT}, \tag{5.10}$$

which has a chi-squared distribution with $I-1$ degrees of freedom under the null hypothesis that the dose-response relationship is linear.

In discussing X^2_H and X^2_{WH} above, it is stated that these statistics have asymptotic chi-squared distributions with I degrees of freedom under the null hypothesis. Exceptions may occur in the analysis of tumour data from experiments in which some groups have extremely high early mortality rates (e.g., due to toxicity of the test compound). For example, if all animals in the ith group die prior to observation of the first tumour in all $I+1$ groups, then $O_i = E_i = 0$, and the ith group makes no contribution to the above test statistics. In such cases, under the null hypothesis of homogeneity among groups, X^2_H and X^2_{WH} will have asymptotic chi-squared distributions with $I-r$ degrees of freedom, where r denotes the number of groups for which $O_i = E_i = 0$. Similarly, if the response is linear, X^2_Q and X^2_{WQ} will have asymptotic chi-squared distributions with $I-r-1$ degrees of freedom. The computation of the test statistics proceeds exactly as above, using only data from those groups for which

$E_i > 0$. When in all risk sets the n's are nearly zero for a particular dose, i.e., there is (almost) no animal at risk in that group, the asymptotic distributions may then not be valid. Thus, dose groups for which E_i is near zero should be omitted in calculating the test statistics, with a corresponding reduction in degrees of freedom for the tests of heterogeneity and of departures from a monotone dose-response relationship.

When results are available from several strata (see Section 2.5), an analysis combining the evidence from all strata can easily be obtained using the statistics described in this section. Let the strata be indexed by j, for $j = 1, \ldots, J$. Restricting the above methods to data from stratum j yields a vector of weighted observed minus expected, $\mathbf{D}_{Wj}$, and an associated covariance matrix, $\mathbf{V}_{Wj}$. Then, the combined analysis proceeds exactly as described in equations (5.8), (5.9) and (5.10), with $\mathbf{D}_W$ replaced by $\sum_{j=1}^{J} \gamma_j \mathbf{D}_{Wj}$, and $\mathbf{V}_W$ replaced by $\sum_{j=1}^{J} \gamma_j^2 \mathbf{V}_{Wj}$, where the γ_j depend on the choice of weights, w_k, in (5.6).

5.3 Nonparametric analysis of survival curves

The first step in evaluating an animal carcinogenesis experiment is to determine the effect of exposure to the test substance on mortality. Suppose that deaths are observed at K distinct times t_k, $k = 1, 2, \ldots, K$. For the purposes of summarizing the effect of exposure to the test compound on mortality, the times of death for animals killed accidentally or in planned sacrifices are considered to be censored observations. Animals lost to observation are considered censored at the time they were last under observation. The mortality data at time t_k may be summarized as in Table 5.1, where x_{ik} is the number of deaths in group i at time t_k, and n_{ik} is the number of animals in group i at risk of dying at t_k (i.e., the number of animals that die at or after t_k). For any time t, let $R(t) = \{k : t_k \leq t\}$; that is, $R(t)$ is the set of all k with index times of deaths occurring at or before t. Then, the Kaplan–Meier estimator of the survival function for group i is the step function defined as (Kaplan & Meier, 1958)

$$\hat{S}_i(t) = \prod_{k \in R(t)} \left(1 - \frac{x_{ik}}{n_{ik}}\right),$$

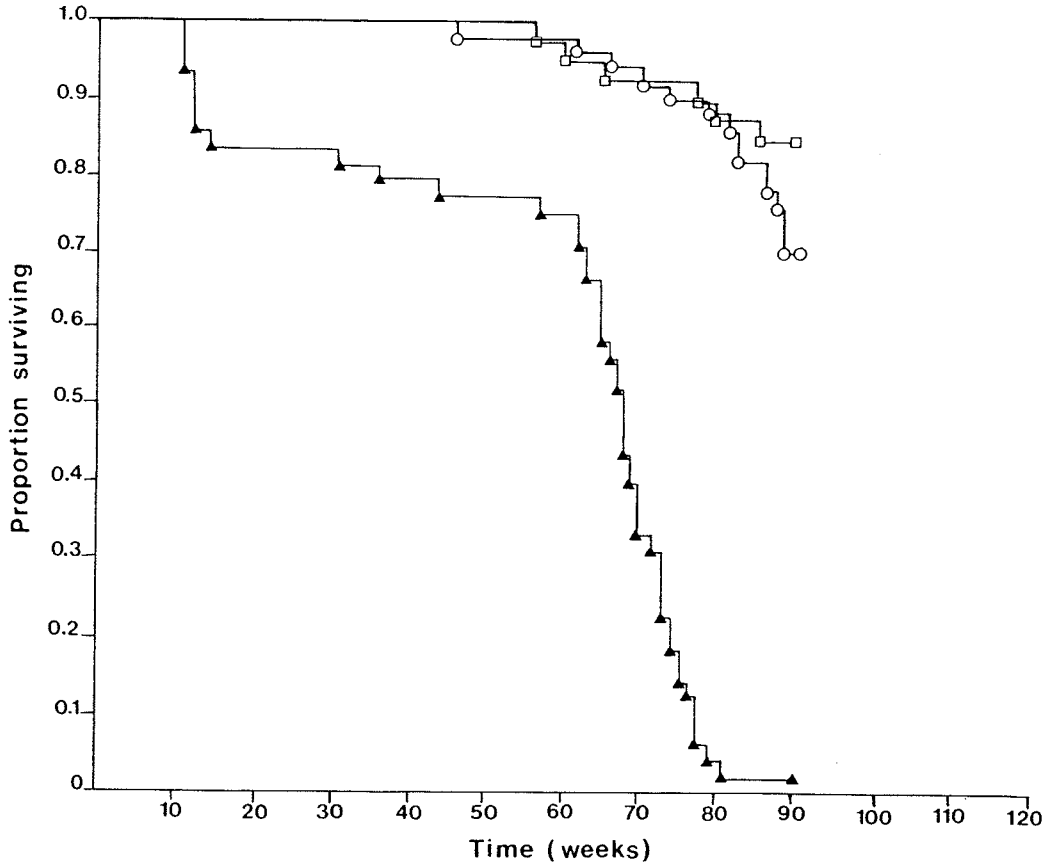
and the variance of $\hat{S}_i(t)$ may be estimated by

$$V\{\hat{S}_i(t)\} = \hat{S}_i^2(t) \sum_{k \in R(t)} \frac{x_{ik}}{n_{ik}(n_{ik} - x_{ik})}.$$

A plot of the $(I + 1)$ estimators, $\hat{S}_i(t)$ from the beginning of the experiment until terminal sacrifice reveals any effects of exposure to the test compound on mortality. Nonparametric estimates of percentiles can be obtained from the Kaplan–Meier survival curve (Miller, 1981b, pp. 74–75), and corresponding confidence intervals can be calculated. (See Slud *et al.*, 1984, for a review and comparison of several available methods.)

Consider the data presented in Table 4.1 from a bioassay of 1,2-dichlorethane using female B6C3F1 mice. All deaths were due to natural causes except for one accidental death at 22 weeks in the control group and the 69 deaths at terminal sacrifice after 90 weeks on study. The Kaplan–Meier survival curves for the control, low-dose and

Fig. 5.1 Kaplan–Meier estimates of survival curves for three groups of female mice (□, control; ○, low dose; ▲, high dose) treated with 1,2-dichloroethane



high-dose groups are given in Figure 5.1. Although there was only a slight increase in mortality in the low-dose group compared to the control group, there was a substantial increase in mortality in the high-dose group. Thus, it is clear that in comparing the proportions of animals with tumour in the high-dose group to those in the control or low-dose group, some consideration must be given to the possibility of bias due to the greater mortality in the high-dose group.

It may not always be necessary to test formally for differences in mortality patterns, as any difference in survival can lead to some degree of bias in the comparison of tumour rates and should, irrespective of its significance, be adjusted for. Nonetheless, formal comparisons can easily be made using generalized rank tests for censored data. The most widely used statistic for testing for survival differences is the generalized Savage statistic, often referred to as the log-rank statistic (Mantel, 1966; Cox, 1972), which is computed using (5.3). The corresponding trend statistic X_T^2, computed using (5.4), and departure from trend statistic X_Q^2, computed using (5.5), were presented by Tarone (1975).

For the data on female mice treated with 1,2-dichloroethane summarized in Figure 5.1, $X_H^2 = 127.8$, $X_T^2 = 85.3$ and $X_Q^2 = 42.5$, with degrees of freedom 2, 1 and 1, respectively. Thus, administration of 1,2-dichloroethane is clearly associated with increased mortality; however, the relationship is not strictly monotone in dose. The significance of X_Q^2 is due to the poor survival in the high-dose group relative to that in

the controls, while the survival in the low-dose and control groups is similar (comparison of the low-dose and control groups yields $p = 0.21$).

The Wilcoxon rank sum test has also been modified for censored survival data, with two proposed modifications, both based on statistics that can be computed using (5.8). For the modified Wilcoxon test of Breslow (1970), w_k is taken to be $n_{.k}$; while for the modified Wilcoxon test of Peto and Peto (1972) and Prentice (1978), w_k is taken to be $\tilde{S}(t_k)$, where $\tilde{S}(t_k)$ is an estimator of the survival function calculated from the pooled data of all $I+1$ groups. The corresponding trend test statistic X^2_{WT}, computed using (5.9), and departure from trend statistic X^2_{WQ}, computed using (5.10), were presented by Tarone and Ware (1977) and by Thomas *et al.* (1977). The modified Wilcoxon statistics are more sensitive than the generalized Savage statistics to differences in survival occurring early in an experiment, when a greater number of animals are at risk (Tarone & Ware, 1977; Thomas *et al.*, 1977). For data from experiments with heavy interim sacrifices, the Peto–Prentice-modified Wilcoxon statistic should be used (Prentice & Marek, 1979).

For the data on 1,2-dichloroethane, the Breslow-modified Wilcoxon statistics are $X^2_W = 110.3$, $X^2_{WT} = 77.2$ and $X^2_{WQ} = 33.1$, with degrees of freedom 2, 1 and 1, respectively. The Wilcoxon statistics give slightly lower values than the corresponding Savage statistics because differences in survival are more pronounced at the end of the experiment.

In combining results of the above linear rank tests from several strata as suggested in Section 5.2, the appropriate stratum weights are $\gamma_j = 1$ for all j for the log-rank statistic and the Peto–Prentice-modified Wilcoxon statistic, and $\gamma_j = (N_j + 1)^{-1}$ for the Breslow-modified Wilcoxon statistic, where N_j is the total sample size in stratum j.

5.4 Analysis of crude proportions

In a well-designed and executed experiment in which there is no great disproportion in survival among the groups, one can usually obtain a good first indication of the possible significance of the results from analysis of the crude proportions of animals with tumour (Gart *et al.*, 1979). This is, of course, the method traditionally used by toxicologists and pathologists. Although disproportionate survival may lead one astray, and such analyses are insensitive to differences in distributions of tumour occurrence or observation times, they are an instructive starting point for discussing the statistical analysis of tumour data.

Choice of denominator

The proportion consists of a numerator of the number of animals with the tumour of a specific site and/or type divided by a denominator of the number of animals at risk for that tumour. We consider three alternative ways of choosing the denominator:

(1) The number of animals initially put on test in each group.

(2) The number initially on test, less the numbers of animals which were not subjected to necropsy or for which the organ site in question was not submitted to or yielded tissue slides unsuitable for pathological examination. Thus, for instance, those

animals whose lungs were not available for pathological examination would be excluded from the denominators (and numerators) in the analyses of lung tumours, but if their livers were so examined, they would be included in the denominators of the analyses of liver tumours. In such cases, the denominators may vary from tumour site to tumour site within the same experiment.

(3) The initial number at risk, less those animals not subjected to necropsy or for which tissue slides for the organ in question are missing, and less, also, those animals 'dying early'. 'Dying early' may be defined in at least two ways:

(3a) Those dying before a pre-specified time on test, say, one year, before which time tumours almost never appear. This usually can be used only if the experimenter has reliable prior knowledge of the test animal and the tumour site. Of course, if a tumour is found before this time, one should use the next option.

(3b) Those dying before the first tumour at a specific site is found in any of the groups being compared. This again can lead to differing denominators for the various tumour sites within the same experiment.

The unadjusted denominator, although often used, is based on the tacit assumption that none of the missing animals had the tumour. One could also assume that all the missing animals had the tumour and adjust the numerators, rather than the denominators, accordingly. In some cases, these extreme possibilities for accounting for missing animals are analysed to determine if such extraordinary results could change the interpretation of the experiment. Such analyses have some polemic value but they are not usually presented in scientific publications. Therefore, we consider only the last two alternatives.

Alternatives (2) and (3) imply that the missing animals or those dying early are as likely to have had the tumour during their full lifetime as those that survived to terminal sacrifice and underwent a necropsy. The net effect of their deletion is to reduce the sample size, perhaps differentially among the groups. The typical outcome, particularly for the 'early death' correction, is that there is more early mortality among the higher-dose groups so that their denominators are reduced more than those of the control or lower-dose groups. Thus, although none of the numerators are changed, the proportions in the higher-dose groups are increased proportionately more. This may lead to a statistically significant positive dose-response or may erase an otherwise negative or inverse dose relation.

To illustrate these concepts and introduce some notation, consider again the data on 1,2-dichloroethane in Table 4.1, with lung as the target site. The two exposed groups consisted initially of 50 female animals, the design specifying equally spaced doses. The control group consisted initially of 40 female animals. The tumour under consideration (alveolar/bronchiolar adenoma) was first found in a high-dose animal dying at 62 weeks. The data may be summarized according to the various criteria in Table 5.2.

As the comparison of crude rates does not involve consideration of the time axis, we have simplified the notation for this Section 5.4 and use only the index for group ($i = 0, 1, \ldots, I$) and suppress the time index ($k = 1, \ldots, K$), which had been introduced in the general notation in Section 5.2.

Regardless of the choice of denominator, the statistical analyses may change in character depending on whether two or more than two groups are being compared and

Table 5.2 Number of animals with lung tumour and at risk (using different criteria) from data on 1,2-dichloroethane

Coded doses	$d_0 = 0$	$d_1 = 1$	$d_2 = 2$
No. with tumour	$y_0 = 2$	$y_1 = 7$	$y_2 = 15$
No. initially at risk	$n_0 = 40$	$n_1 = 50$	$n_2 = 50$
No. missing	0	0	2
No. at risk (Criterion 2)	$n_0' = 40$	$n_1' = 50$	$n_2' = 48$
No. dying before 62 weeks	3	2	12
No. at risk (Criterion 3b)	$m_0 = 37$	$m_1 = 48$	$m_2 = 36$

also on whether a large sample approximation may be legitimately employed rather than an exact or conditional analysis. We consider these cases in turn.

Comparison of two groups

Usually, this involves the comparison of a single-dose group and a control group. Consider the control group and the high-dose group in the data on 1,2-dichloroethane and lung tumours just presented. We use the denominator from the early-death criterion (3b). The notation and data are given in 2×2 tables in Table 5.3.

Table 5.3 2×2 table for comparison of control group and high-dose group from data on 1,2-dichloroethane

	Notation			Data		
Dose	d_0	d_2	Total	0	2	Total
Animals with tumour	y_0	y_2	s	2	15	17
Animals without tumour	$m_0 - y_0$	$m_2 - y_2$	$m_. - s$	35	21	56
	m_0	m_2	$m_.$	37	36	73

Approximate analyses of two groups

In our example, we consider the number of animals with tumour in the high-dose group as the observed quantity, $O = y_2$. Define the expected numbers of animals with tumour under the null hypothesis of no difference in tumour rates to be $E = (sm_2)/m_.$ for the high-dose group and $E' = (sm_0)/m_.$ for the control group. Define $D = O - E$. The variance of D under the null hypothesis is estimated by

$$V = \{s(m_. - s)m_0 m_2\}/\{m_.^2(m_. - 1)\}.$$

Alternatively, the reciprocal of this variance may be computed from the table of expected values,

$$1/V = \{(m_. - 1)/m_.\}\{1/E + 1/(m_2 - E) + 1/E' + 1/(m_0 - E')\}.$$

Many authors (Armitage, 1971, pp. 129 ff. and Snedecor & Cochran, 1980, pp. 124 ff.) use $m_.$ in place of $m_. - 1$ in the formula for V. Test statistics based on the above variance, however, have distributions better approximated by the normal or chi-square distribution in the unconditional sample space (Upton, 1982). The present formula is also better if one combines analyses for differing sexes and/or strains of test animals.

The approximately normal deviate test for equality of proportions is then

$$Z = D/\sqrt{V_{\cdot}}.$$

Large positive values of Z indicate a direct or positive relation with the application of the compound and tumour production, and large negative values indicate an inverse or negative relation between application of compound and tumour production. One-tailed p-values are then read from tables of the normal or Gaussian distribution. Two-tailed tests are conveniently performed by considering

$$X^2 = Z^2 = D^2\{(m_{\cdot} - 1)/m_{\cdot}\}\{1/E + 1/(m_2 - E) + 1/E' + 1/(m_0 - E')\},$$

which is an approximate chi-square variate with one degree of freedom.

If a continuity correction is used, we have

$$Z_c = (D \pm \tfrac{1}{2})/\sqrt{V},$$

where $-\frac{1}{2}$ is used for a one-tailed test of a positive or direct relationship and $+\frac{1}{2}$ is used for a one-tailed test of a negative or inverse relation. Note that the continuity correction is employed to make the p-value for the approximate test closer to that of the exact or conditional test, which we discuss later. It has little effect for large numbers.

The significance test depends not only on the relative magnitude of the differential effect of the exposure on tumour production in the two groups but also on sample size. A commonly-used measure of this effect, which does not depend on sample size, is the odds ratio. This is the ratio of the odds of a tumour in the treated group to the corresponding odds in the control group. In the notation of early-death criterion (3b) this is estimated by the cross-product ratio:

$$\hat{R} = \{y_{2\cdot}(m_0 - y_0)\}/\{y_0(m_2 - y_2)\},$$

where $\hat{R} > 1$ indicates a positive relation, $\hat{R} = 1$, no relation, and $\hat{R} < 1$, a negative or inverse relation of exposure with tumour production. Approximate confidence limits for this parameter can be computed by the method of Cornfield, which has been implemented in several computer programs (e.g., Thomas, 1975).

The question arises as to how large the numbers have to be to apply these approximate methods. The validity of these methods is not determined by the magnitude of the observed numbers themselves, but by the magnitude of the minimum of the expected values corresponding to the particular test or confidence interval method used. Thus, if

$$\min(E, m_2 - E, E', m_0 - E') \geq 1,$$

the Z_c-test should give a good approximation to the exact p-values. The accuracy of the approximate confidence interval also depends on the minimum expected values consistent with the marginal totals and the values of the odds ratios computed at the two confidence limits (see, e.g., Gart & Thomas, 1972). Thus, for instance, an experiment may be large enough to use approximate methods for a p-value but not for an upper 95% confidence limit.

Returning to our example, we have $O = 15$, and the table of expected values is shown in Table 5.4.

Table 5.4 Expected values in comparison of control group and high-dose group from data on 1,2-dichloroethane

	Dose		Total
	d_0	d_2	
With tumour	$E' = 8.62$	$E = 8.38$	$s = 17$
Without tumour	$m_0 - E' = 28.38$	$m_2 - E = 27.62$	$m_. - s = 56$
Total	$m_0 = 37$	$m_2 = 36$	$m_. = 73$

The minimum of the expected values, 8.38, is clearly large enough to apply the approximate test. Furthermore, we have

$$V = \{(17)(56)(37)(36)\}/\{(73)^2(72)\} = 3.3049,$$

or, alternatively,

$$1/V = (72/73)\{1/(8.38) + 1/(27.62) + 1/(8.62) + 1/(28.38)\} = 0.3026.$$

Thus,

$$Z = 6.62/\sqrt{(3.3049)} = 6.62\sqrt{(0.3026)} = 3.64,$$

for which the corresponding one-tailed $p = 0.00014$, indicating a highly significant positive difference between the high-dose and the control group. The two-tailed chi-square test yields

$$X^2 = Z^2 = 13.26, \qquad p = 0.00028.$$

The corresponding continuity corrected test is

$$Z_c = (6.12)/\sqrt{(3.3049)} = 3.37, \qquad p = 0.00038.$$

The cross-product ratio is $\hat{R} = \{(15)(35)\}/\{2(21)\} = 12.50$. The associated approximate 95% limits are (2.35, 88.28). Checking the validity of the approximation at the limits, we compute the expected values in the four-fold tables with fixed marginals for the lower and upper limits. These are given in Table 5.5.

Table 5.5 Expected values in tables corresponding to lower and upper confidence limits

Lower limit:	$R_l = 2.35$	
5.8998	11.1002	17
31.1002	24.8998	56
37	36	
Upper limit:	$R_u = 88.28$	
0.3569	16.6431	17
36.6431	19.3569	56
37	36	

In the table corresponding to the lower limit, the minimal entry 5.8998 is greater than 1, whereas in the table corresponding to the upper limit, the minimal entry 0.3569 is less than 1. Thus, although the significance test and the lower 95% limit would appear to be approximately correct, the upper limit 95% cannot be relied upon here.

If $m_.$ is substituted for $m_. - 1$ in V, then the Z_c test will always agree with the 95% confidence interval in excluding the odds ratio of $R = 1$ whenever the one-tailed p-value is less than 0.025 and *vice versa*. It will usually agree, as in this case, with the computation of Z_c.

Exact or conditional analyses of two groups

When the numbers are small, exact or conditional analyses are feasible and may be necessary. The theoretical basis of such analyses is the initial randomization of the animals into two groups (Gart *et al.*, 1979). Consider the 73 animals in our example to be randomly divided into two groups of 37 and 36. If exposure to the chemical does not change the risk of tumour, then, regardless of the outcome of the randomization, 17 animals are fated to have this tumour. This 'fixing' of 17 as the marginal total is the reason for calling this analysis 'conditional'. Now consider the actual outcomes of the experiment, i.e., in this particular randomization, $y_2 = 15$, and those possible outcomes 'more extreme' in the positive direction, in this case $y_2 = 16$, and $y_2 = 17$. It is a simple combinatorial exercise to count the numbers of ways in which these outcomes can occur relative to the total number of possible randomizations. The ratio of these numbers is the precise one-tailed p-value for the Fisher–Irwin exact test.

We put this argument in mathematical notation. The total number of possible randomizations of $m_.$ animals having s tumours is given by the binomial coefficient $\binom{m_.}{s}$. For any integers u and v, for which $0 \leq u \leq v$, $\binom{v}{u}$ is also referred to as the number of ways of choosing u objects from v objects, is given by

$$\binom{v}{u} = \frac{v(v-1)\cdots(v-u+1)}{1 \cdot 2 \cdots u}.$$

The number of ways in which there can be y animals with tumour in the dose group and $s - y$ in the control group is

$$\binom{m_0}{s-y}\binom{m_2}{y}, \qquad y = 0, 1, \ldots, s.$$

The conditional probability that y occurs is thus

$$P(y \mid s) = \binom{m_0}{s-y}\binom{m_2}{y}\Big/\binom{m_.}{s}, \qquad y = 0, 1, \ldots, s.$$

The exact one-tailed p-value for a possible increase in tumour incidence in the dose group is then

$$p = \sum_{y=y_2}^{s} P(y \mid s).$$

In applying this formula, note that $\binom{k}{j}$ is defined as zero when $j > k$.

Two-tailed exact tests are not so simply defined, since 'more extreme' does not have a unique meaning for unequal sample sizes. A reasonable procedure is to define p by accumulating all y such that $P(y \mid s) \leq P(y_2 \mid s)$, where y_2 is the observed outcome. When the sample sizes are equal, this rule leads to a p-value simply twice that for a one-tailed test. Otherwise, one may use the special tables of Armsen (1955).

Conditional point estimates of the odds ratio and exact confidence limits for this parameter have been described by Fisher (1935), Cornfield (1956) and Gart (1970), and their computation usually requires a computer program (see, e.g., Thomas, 1975). Programs that have been developed for pocket calculators can also be used (Rothman & Boice, 1979).

Consider again our example. We find

$$P(15 \mid 17) = \binom{37}{2}\binom{36}{15}\Big/\binom{73}{17} = 0.00021$$

$$P(16 \mid 17) = \binom{37}{1}\binom{36}{16}\Big/\binom{73}{17} = 0.00002$$

$$P(17 \mid 17) = \binom{37}{0}\binom{36}{17}\Big/\binom{73}{17} = 0.00000$$

and thus $p = 0.00023$.

Recall that the approximate one-tailed Z_c test yielded a comparable value of $p = 0.00038$.

The computer program of Thomas (1975) yields the conditional maximum likelihood estimator for the odds ratio of 12.08 *versus* the cross-product ratio of 12.50 noted previously. Similarly, the exact 95% confidence limits for the odds ratio are (2.44, 119.33). The lower limit is comparable to the approximate value, 2.35, but the upper limit is quite different from the approximate upper limit, 88.28. This confirms the previous finding that the approximate upper limit is not reliable because it depends on a very small expected value. Note also that the exact limits include the null value of 1 whenever the appropriate one-tailed exact p is greater than 0.025, and will exclude 1 when p is less than 0.025.

Comparison of several groups

The usual design has one control group and at least two dose groups of a compound under test. One is usually interested in testing whether the proportion of animals with tumour increases or decreases monotonically with dose; that is, if $p(d_i)$, $i = 0, \ldots, I$ are the true proportions of the tumour among the various groups, whether $p(d_i)$ is a monotonic function of dose. A convenient monotonic function is the logistic,

$$p(d_i) = \exp(\alpha + \beta d_i)/\{1 + \exp(\alpha + \beta d_i)\}, \qquad i = 0, 1, 2, \ldots, I,$$

where, typically, $d_0 = 0$, corresponding to the control group. This may be written in the

logarithmic scale as

$$\log\{p(d_i)/q(d_i)\} = \alpha + \beta d_i,$$

where $q(d_i) = 1 - p(d_i)$. This implies that the log odds (or logit) is a linear function of dose. Alternatively, this means that the odds ratio between two doses d_i and d_j is

$$R_{ij} = \{p(d_i)q(d_j)\}/\{q(d_i)p(d_j)\} = \exp\{\beta(d_i - d_j)\}.$$

Thus, if the doses are equally spaced, say at unit intervals, the model implies that odds ratios between adjacent doses are equal. This model has properties that enable the extension of simpler exact tests to more complex situations (Cox, 1958, 1970, Chapter 5). It should be pointed out, however, that most of the tests based on the logistic model are robust, that is, they are valid regardless of whether this model holds exactly, and many can also be justified from completely model-free considerations.

The data are usually arrayed in a $2 \times (I + 1)$ table, as in Table 5.6.

Table 5.6 Notation for data from experiment with $I + 1$ groups

	Dose				Total
	d_0	d_1	...	d_I	
With tumour	y_0	y_1	...	y_I	s
Without tumour	$m_0 - y_0$	$m_1 - y_1$	...	$m_I - y_I$	$m_. - s$
Total	m_0	m_1	...	m_I	$m_.$

Our numerical example, with the elimination of the early-death criterion (3b), is given in Table 5.7.

Table 5.7 Data on lung tumour for three groups from study on 1,2-dichloroethane

	Dose			Total
	0	1	2	
With tumour	2	7	15	24
Without tumour	35	41	21	97
Total	37	48	36	121

Approximate analyses of several groups

Let us denote the observed numbers with tumour as $O_i = y_i$, $i = 0, 1, \ldots, I$, and their corresponding expected values under the null hypothesis of no difference as

$$E_i = (sm_i)/m_., \qquad i = 0, 1, \ldots, I,$$

where $m_. = \sum_{i=0}^{I} m_i$. Defining $D_i = O_i - E_i$, the test statistic for possible monotonic trend with dose is based on

$$T = \sum_{i=0}^{I} d_i D_i.$$

Under the null hypothesis, T has mean zero, and its variance is estimated by

$$V_T = [\{s(m_. - s)\}/\{m_.(m_. - 1)\}] \sum_{i=0}^{I} m_i(d_i - \bar{d})^2,$$

where $\bar{d} = (\sum_{i=0}^{I} m_i d_i)/m_.$. The Cochran–Armitage normal deviate test for trend (see, e.g., Armitage, 1971, pp. 363–365) is then

$$Z_T = T/\sqrt{V_T}.$$

If the doses are equally spaced, say, with interval Δ, a simple continuity correction may be easily employed to yield,

$$Z_{Tc} = (T \mp \Delta/2)/\sqrt{V_T},$$

where the minus or plus signs are used for one-tailed testing against a direct and inverse relation, respectively. In the case of unequally spaced doses, the continuity correction is less readily applied (Kendall & Stuart, 1961, p. 508). Although the actual level of the Cochran–Armitage trend test can deviate from the nominal level for asymmetric designs (Portier & Hoel, 1984b), this can be remedied by a Cornish–Fisher skewness correction (Tarone, 1986).

Two-tailed tests may be based on the squared value of Z_T, $X_T^2 = Z_T^2$, which is an approximate chi-square variate with one degree of freedom. For $I = 1$, these tests are exactly equivalent to the approximate tests for comparing two groups. Although the Cochran–Armitage test follows from the assumption of a logistic model, Tarone and Gart (1980) showed that it is asymptotically, locally fully efficient for testing the null hypothesis against any choice of a monotonic, locally linear function. The approximate tests should be adequate as long as the minimum expected value exceeds one.

The question arises whether a linear relation is an appropriate alternative. Testing for this possibility is facilitated by first computing the usual chi-square for heterogeneity in a $2 \times (I + 1)$ contingency table:

$$X_H^2 = \{(m_. - 1)/m_.\}\left[\sum_{i=0}^{I} D_i^2\{1/E_i + 1/(m_i - E_i)\}\right], \tag{5.11}$$

which is approximately distributed as a chi-square variate with I degrees of freedom if all of the $p(d_i)$ are equal. This statistic can also be used for testing for heterogeneity among groups in which there is no quantitative dose relationship in the treatment regimens, i.e., differing chemicals and/or vehicles or other differing control groups. For the dose-relation situation, the approximate chi-square statistic with $I - 1$ degrees of freedom for departure from linear trend is $X_Q^2 = X_H^2 - X_T^2$.

This computation is usually presented in a table analogous to an analysis-of-variance table. If $I = 2$, the chi-square statistic X_Q^2 is the appropriate statistic for testing the possibility of a quadratic relation (hence, the subscript Q). When $I \geq 3$, it is the statistic for an omnibus test of all nonlinear polynomial coefficients, i.e., quadratic, cubic, quartic, etc. When $I = 1$, $X_T^2 \equiv X_H^2$ and thus $X_Q^2 \equiv 0$, and no test of departure is possible.

The strength of the possible effect of dose can be estimated by fitting the linear logistic model. We may employ the method of maximum likelihood (see, e.g., Thomas

& Gart, 1983) to estimate β by $\tilde{\beta}$. The odds ratio, R_i, between any dose d_i and the control $d_0 = 0$, is then estimated by $\tilde{R}_i = \exp(\tilde{\beta} d_i)$. Alternatively, a much more simply computed estimator of R_i is found from the cross-product ratio. These estimators are $\hat{R}_i = \{O_i(m_0 - O_0)\}/\{O_0(m_i - O_i)\}$, $i = 1, \ldots, I$. This appears to be a reasonably good estimator for $\frac{1}{3} \leq R_i \leq 3$, but is otherwise biased towards unity.

Returning to our example, our preliminary calculations are illustrated in Table 5.8.

Table 5.8 Expected values for data on 1,2-dichloroethane (Table 5.7)

	Dose group				Total
	d_i	0	1	2	
No. with tumour	O_i	2	7	15	24
Expected no.	E_i	7.34	9.52	7.14	24
	$m_i - E_i$	29.66	38.48	28.86	97
Total	m_i	37	48	36	121
	$O_i - E_i = D_i$	−5.34	−2.52	+7.86	0.00

As the minimum expected value exceeds one, the approximate tests should be adequate.

$$T = (-5.34)(0) + (-2.52)(1) + (7.86)(2) = 13.20,$$

and

$$\sum_{i=0}^{2} m_i(d_i - \bar{d})^2 = \sum_{i=0}^{2} m_i d_i^2 - \left(\sum_{i=0}^{2} m_i d_i\right)^2 \Big/ m.$$
$$= 192 - (120)^2/121 = 72.9917.$$

Therefore,

$$V_T = [\{(24)(97)\}/\{(121)(120)\}](72.9917) = 11.7028.$$

The one-tailed test for positive trend yields

$$Z_T = (13.20)/\sqrt{(11.7028)}$$
$$= 3.86, \qquad p = 0.00006,$$

and the two-tailed chi-square test is

$$X_T^2 = Z_T^2 = (3.86)^2 = 14.90, \qquad p = 0.00011.$$

The corresponding continuity corrected test is

$$Z_{Tc} = (12.70)/\sqrt{(11.7028)} = 3.71, \qquad p = 0.00010.$$

Turning to the question of possible departure from linearity, we compute $X_H^2 = 16.33$ according to (5.11) and derive Table 5.9. Clearly, there is no evidence that a linear model does not fit.

If we fit the logistic model by maximum likelihood, we find $\tilde{\beta} = 1.32 \pm 0.37$. The

Table 5.9 Summary of analysis of crude proportions for the 1,2-dichloroethane example

Source of variation	Degrees of freedom	Chi-square	p
Linear trend	1	$X_T^2 = 14.90$	0.00011
Departure from linearity	1	$X_Q^2 = 1.43$	0.23
Total (heterogeneity)	2	$X_H^2 = 16.33$	0.00028

estimates of the odds ratio of the dosed to control groups, using this value and the cross-product ratios, are given in Table 5.10. The good agreement between these estimates reflects the fact that the linear logistic model fits these data well.

Table 5.10 Estimates of odds ratio from data on 1,2-dichloroethane

	Dose group		
$d_i = i$	0	1	2
$\hat{R}_i$	1.00	3.60	12.50
$\tilde{R}_i = \exp(\tilde{\beta} i)$	1.00	3.74	14.01

Exact or conditional analyses for linear trend

When the numbers are small, it may be necessary to use the exact test for trend (Cox, 1958), which is a generalization of the Fisher–Irwin test. This test statistic can be derived from the logistic model, but the null hypothesis and the distribution used to obtain a p-value hold very generally. Like the exact test for two groups, its theoretical basis is the randomization of the animals into several groups. Under the null hypothesis, it is assumed that the total number of animals with tumour in all the groups is fixed at s. The conditional distribution is then

$$P(y_0, y_1, \ldots, y_I \mid s) = \binom{m_0}{y_0}\binom{m_1}{y_1}\cdots\binom{m_I}{y_I}\Big/\binom{m.}{s}$$

where $\sum_{i=0}^{I} y_i = s$. The observed statistic for which probabilities of more extreme outcomes will be calculated is $\sum_{i=0}^{I} O_i d_i = A$. For tests of positive trend, the p value is computed from

$$p = \sum_{\Omega} P(y_0, y_1, \ldots, y_I \mid s).$$

where Ω consists of all possible values of $y_i \geq 0$ such that $\sum_{i=0}^{I} y_i = s$ and $\sum_{i=0}^{I} y_i d_i \geq A$. For a test of a negative trend, the sense of the last inequality is reversed. The application of this test usually requires a computer program (see, e.g., Thomas *et al.*, 1977). For $I = 1$, this test is identical to the exact test for two groups. Cox (1958)

showed that the Cochran–Armitage test is the normal approximation to the exact randomization trend test.

It is also possible to perform an exact test for departure from linearity. This requires further conditioning on the observed value of $\sum y_i d_i$, which results in a more complex distribution. Bayer and Cox (1979) have published a program that can be used for this purpose. Conditional maximum likelihood methods can be used in small numbers to estimate β, and these have been implemented, in a somewhat more general context, by Smith *et al.* (1981).

Returning briefly to our example, we find that the conditional test for linearity, fixing $s = 24$, yields the exact one-tailed $p = 0.00007$, which is quite close to the value found from the continuity-corrected Z_{Tc}, specifically, $p = 0.00010$.

Combination of results over sexes, strains or experiments

The approximate tests for trend can easily be combined over sexes, strains or experiments. In each analysis the doses must be uncoded or be coded in the same way. The combined normal deviate test statistic is calculated simply by adding the numerators and the squares of the denominators of the individual statistics and dividing the summed numerators by the square root of the summed squared denominators. Its mathematical formula is $Z_T = \sum T/\sqrt{(\sum V_T)}$, where the summation is over the different subexperiments.

The continuity corrected normal deviate test is $Z_{Tc} = (\sum T \pm \Delta/2)/\sqrt{(\sum V_T)}$, where the doses are equally spaced, Δ units apart in all experiments, and the minus is used for testing a direct relation and the plus for an inverse relation. Note that the $\Delta/2$ corrections are *not* summed over the several experiments in combining the test statistics. This is the so-called Mantel–Haenszel procedure (Mantel & Haenszel, 1959; Mantel, 1963), which is the optimal procedure for testing the common slope, β, of a stratified logistic model. For $I = 1$ it essentially reduces to Cochran's (1954) test for the combination of 2×2 tables. Radhakrishna (1965) and Tarone and Gart (1980) showed the asymptotic efficiency of these combined tests to be robust (or insensitive) to modest departures from this logistic model.

Combined approximate tests for departure from linearity or for homogeneity of slopes from different experiments, and the maximum likelihood estimation of a common β over several experiments, usually require the use of a computer program (see, e.g., Thomas & Gart, 1983). Exact combined tests also usually require such programs (for $I = 1$, see, e.g., Thomas, 1975; for $I \geq 2$, see, e.g., Bayer & Cox, 1979), as does calculation of the conditional maximum likelihood estimate (for $I = 1$, see Thomas, 1975; for $I \geq 2$, see Smith *et al.*, 1981).

The various combined tests may not be appropriate if the relative effect of treatment varies greatly over the various strata or experiments being combined. Such variation for logistic models is measured by differences in the odds ratio for one-dose experiments or by differences in logistic slope for multiple-dose experiments. Statistical tests for the homogeneity of odds ratios are given by Breslow and Day (1980, pp. 142–146; see also Tarone, 1985) and tests for homogeneity of logistic slopes by Thomas and Gart (1983).

The issue of multiple comparisons among doses

Often, experimenters compare each of the I dosed groups in turn with the control group and report the results of these several tests in addition to the trend test. This raises the problem of multiple comparisons. If there is a strong direct relationship with dose, analyses of the results of the high dose and perhaps of some lower doses agree with the trend test in finding significance. A problem of interpretation develops when the lower-dose comparison is significant while the high-dose comparison and the trend test are not. In such cases, the chi-squares for homogeneity and departure from trend are usually large, if not significantly so. An adjustment in significance may be employed to allow for the possibility of finding significance by any one of two or more statistical tests. The simplest and most widely used correction employs the Bonferroni inequality (Miller, 1981a, pp. 6–10). If the desired overall significance level for the test of the chemical compound at I doses is α, the individual comparison of the ith dose to control is made at a significance level α_i, where $\sum_{i=1}^{I} \alpha_i = \alpha$. Alternatively, the observed p-value for comparison of the ith dose to control is multiplied by α/α_i. Because greater emphasis should be given to significance at the highest dose, α_I should be chosen to be larger than the remaining α_i, $i = 1, \ldots, I-1$.

The Bonferroni correction may be used to adjust multiple tests both in the previously considered survival analyses as well as in the following analysis of prevalent and rapidly lethal (or observable) tumour rates. Unless the target organ for a given test compound is known in advance, control of the overall experimental error rate is necessary. This more difficult question of multiple comparisons over organ sites is discussed in Section 7.2.

5.5 Prevalence analysis for nonlethal occult tumours

Hoel and Walburg (1972) pointed out the importance, when evaluating data on occult tumours, of making a distinction between those tumours that are lethal and those that are nonlethal. Nonlethal occult tumours are discovered at necropsy, either after terminal sacrifice or after an animal has died prior to terminal sacrifice because of illness unrelated to the presence of the tumour. In this section, tests for the equality of prevalence rates for nonlethal occult tumours are presented. An assumption underlying the derivation of these prevalence tests is that, at least with regard to the presence or absence of a nonlethal tumour, death is a random sampling mechanism. This is an extremely strong assumption, implying that a tumour-bearing animal is, in every way except for the presence of a tumour, as healthy as a tumour-free animal. Nonetheless, such statistical procedures can be useful in evaluating the carcinogenic potential of a test compound. In this section it is assumed that all tumours of a particular type observed in the carcinogenesis experiment under consideration are nonlethal.

Suppose the carcinogenesis experiment extends from time zero to time T, where T denotes the time at which the terminal sacrifice is scheduled. Now suppose that the interval $(0, T)$ is subdivided into $J-1$ subintervals, $\mathcal{I}_j$, where $\mathcal{I}_j = (T_{j-1}, T_j]$ for $j = 1, 2, \ldots, J-2$, and $\mathcal{I}_{J-1} = (T_{J-2}, T_{J-1})$, with $T_0 = 0$ and $T_{J-1} = T$. Then, let the number of animals dying in group i during subinterval $\mathcal{I}_j$ be denoted by N_{ij} and the

Table 5.11 Summary of tumour prevalence data for nonlethal tumours in interval $\mathcal{I}_j$

	Dose						Total
	d_0	d_1	⋯	d_i	⋯	d_I	
No. of animals with tumour	Y_{0j}	Y_{1j}	⋯	Y_{ij}	⋯	Y_{Ij}	$Y_{.j}$
No. of animals dying in $\mathcal{I}_j$	N_{0j}	N_{1j}	⋯	N_{ij}	⋯	N_{Ij}	$N_{.j}$

number of these animals in which a tumour is discovered at necropsy be denoted by Y_{ij}, $i = 0, 1, \ldots, I$. Then, for each subinterval $\mathcal{I}_j$, the tumour prevalence data may be summarized in a $2 \times (I + 1)$ table such as Table 5.11.

In addition, the tumour prevalence of the animals killed at terminal sacrifice can be summarized in a similar table, indexed by J, where N_{iJ} denotes the number of animals in group i surviving to terminal sacrifice, and Y_{iJ} denotes the number of these animals in which a tumour is found at necropsy. Note that, if there are any planned interim sacrifices, each time of such a sacrifice is treated as a distinct subinterval and contributes a separate $2 \times (I + 1)$ table of tumour prevalence data. Once the tumour prevalence data have been stratified into J strata, as described above, tests for equality of tumour prevalence rates can be derived using standard contingency table methods.

The prevalence test statistics can be computed using (5.3), (5.4) and (5.5), where D_i and V_{hi} are calculated as in (5.1) and (5.2), after substituting J for K, j for k, Y_{ij} for x_{ik} and N_{ij} for n_{ik}. Let X^2_{PH} denote the test for equality of tumour prevalence rates in the $I + 1$ groups using (5.3) (Armitage, 1966), X^2_{PT} denote the corresponding trend test statistic computed using (5.4) (Mantel, 1963), and X^2_{PQ} denote the corresponding departure from trend statistic computed using (5.5).

Consider now the data given in Table 4.1 on alveolar/bronchiolar adenomas in the experiment with 1,2-dichloroethane in female mice. In the opinion of a pathologist involved in evaluating this experiment, it was extremely unlikely that any of these adenomas contributed to the deaths of tumour-bearing animals. This claim is supported by the fact that, in the two groups (control and low-dose) with good survival, alveolar/bronchiolar adenomas were found only in animals surviving until terminal sacrifice. To demonstrate the prevalence methods for nonlethal tumours, let us first assume that the 90-week experiment was divided (prior to evaluation of the data) into three subintervals, (0, 52], (53, 72] and (73, 90), with terminal sacrifice planned at 90 weeks. No tumour was found in animals dying in the first 52 weeks, and, hence, the prevalence analysis is based on the 2×3 contingency tables presented in Table 5.12.

Applying the above methods, we find $X^2_{PH} = 15.10$, $X^2_{PT} = 13.51$ and $X^2_{PQ} = 1.59$, with 2, 1 and 1 degrees of freedom, respectively. Thus, administration of 1,2-dichloroethane is associated with increased tumour prevalence, and the increase is clearly dose-related.

The method of subdividing the length of the experiment into subintervals warrants further discussion. In the above analysis it was assumed that subdivisions were chosen *a priori*, without reference to the data on tumour prevalence. Peto *et al.* (1980) suggest an adaptive interval selection method in which subintervals are determined by the tumour prevalence data. This method is illustrated using the tutorial example of Peto *et*

Table 5.12 Contingency tables for preselected time intervals for prevalence analysis of data on 1,2-dichloroethane

$\mathcal{I}_1 = (53, 72)$

d_i	0	1	2
Y_{i1}	0	0	8
$N_{i1} - Y_{i1}$	3	3	14

$\mathcal{I}_2 = (73, 90)$

Y_{i2}	0	0	6
$N_{i2} - Y_{i2}$	3	11	8

$\mathcal{I}_3$: terminal sacrifice

Y_{i3}	2	7	1
$N_{i3} - Y_{i3}$	31	28	0

al. (1980) (Table 5.13), and is then applied to the data on 1,2-dichloroethane and lung adenoma.

The first step in the adaptive interval selection method, shown in row 1 of Table 5.13, is to list, in increasing order, the times of death of all animals (pooling times of death from all $I + 1$ exposure groups) for which necropsies were performed. The list starts with the first time at which an animal died and was subjected to necropsy – week 50. The times of death for animals in which a tumour was found are underlined. In the case of ties (i.e., times at which some animals had tumours but others did not), the animals with tumours are listed first. The second row of the table contains asterisks which separate the times into what Peto *et al.* refer to as 'ad-hoc runs'. Each ad-hoc run is a sequence of consecutive underlined times followed by a sequence of consecutive times without underlining. The times before that at which the first animal with a tumour was found do not play any further role in the analysis, as the prevalence is zero for that period. The third row gives the proportion of underlined times in each run; that is, this row gives the estimated tumour prevalence within each time interval

Table 5.13 Calculation of subintervals for prevalence analysis using the adaptive method of Peto *et al.* (1980)

(1)	50	67	67	67	67	94	97	105	110	110	115	115	120	121	124	124	128	130	134	137	139	
(2)		*				*		*			*			*					*			*
(3)		0.25				0.50			0.67			0.67				0.20				0.67		
(4)		*				*		*			*								*			*
(5)		0.25				0.50			0.67					0.375						0.67		
(6)		*				*		*											*			*
(7)		0.25				0.50								0.455						0.67		
(8)		*				*													*			*
(9)		0.25										0.462								0.67		

defined by a run. The remaining rows summarize the process of merging adjacent runs for which the estimated prevalence rates decrease with increasing time, and deleting asterisks separating merged runs. This process of 'pooling adjacent violators' is equivalent to maximum likelihood estimation of prevalence rates assuming nondecreasing prevalence (Ayer *et al.*, 1955). For the data in the table, the first decrease in row 3 occurs between the fourth run (estimated prevalence, 0.67) and the fifth run (estimated prevalence, 0.20). Merging of these runs forms a new fourth interval with an estimated prevalence of 0.375 (row 5). The first decrease in row 5 occurs between the third run (estimated prevalence, 0.67) and the new fourth interval. Merging of the third run and the fourth interval results in a new third interval with an estimated prevalence 0.455 (row 7). The first decrease in row 7 occurs between the second run (estimated prevalence, 0.50) and the new third interval. Merging of the second run with the third interval results in three intervals with increasing estimated prevalence rates (row 9). Thus, the adaptive method gives a subdivision into three time intervals: the first subinterval consists of week 67, with an estimated prevalence of 0.25; the second subinterval consists of weeks 94–130, with an estimated prevalence of 0.462; and the third subinterval consists of weeks 134–139, with an estimated prevalence of 0.67.

For the data on the effects of 1,2-dichloroethane in female mice on adenoma incidence, the above adaptive method leads to a single interim subinterval from 62 weeks (when the first adenoma was found) to 88 weeks. Thus, the prevalence tests using the adaptive interval selection method are based on the 2×3 contingency tables in Table 5.14.

For these tables, $X^2_{PH} = 14.49$, $X^2_{PT} = 13.23$ and $X^2_{PQ} = 1.25$, with degrees of freedom 2, 1 and 1, respectively. These values are smaller, but quite similar, to the values obtained previously using the prevalence analysis after a-priori subdivision into three subintervals.

Whatever the method of interval selection, it is possible to find subintervals in which deaths are observed in only one exposure group. Tumours found in such subintervals will be ignored in calculating the prevalence test statistics. All such subintervals occurring after the first tumour has been observed may be merged with adjacent intervals containing deaths in additional exposure groups, although care should be

Table 5.14 Contingency tables after time partition by ad-hoc runs for prevalence analysis of data on 1,2-dichloroethane

$\mathcal{I}_1 = [62, 90)$

d_i	0	1	2
Y_{i1}	0	0	14
$N_{i1} - Y_{i1}$	4	13	21

$\mathcal{I}_2$: terminal sacrifice

Y_{i2}	2	7	1
$N_{i2} - Y_{i2}$	31	28	0

taken not to merge intervals with widely disparate baseline prevalence rates. One instance in the adaptive interval selection method in which this situation will arise routinely is when the last animal dying prior to terminal sacrifice has a tumour but the penultimate animal dying has no tumour. In this case, the last subinterval will contain only the animal which died last. Accordingly, the last time of death should be included in the immediately preceding subinterval, provided this preceding subinterval includes deaths from another exposure group.

For both methods of interval selection described above, when applied to the data for 1,2-dichloroethane, the resulting prevalence method chi-squared statistics for the effect of exposure on tumour prevalence are smaller than the corresponding chi-squared statistics based on the crude tumour rates after eliminating animals dying prior to observation of the first tumour (Table 5.9). For an experiment the size of that with 1,2-dichloroethane, such a finding is not unusual. Regardless of the interval selection method, some efficiency may be lost because of the small number of animals dying in control (and sometimes low-dose) groups prior to terminal sacrifice. This is better illustrated by considering what would have happened if there had been no low-dose group in the 1,2-dichloroethane experiment. In order to compare the control group and the high-dose group, the prevalence test using the adaptive interval selection method is based on the 2×2 contingency tables in Table 5.15.

The prevalence test for equality of tumour rates gives $X^2_{PH} = X^2_{PT} = 6.23$. The analysis of crude tumour rates after eliminating animals dying prior to observation of the first tumour (see Section 5.4) gave an approximate chi-squared test statistic for equality of tumour rates of $X^2 = 13.26$. Even though survival in the high-dose group is quite poor, the simpler analysis of adjusted crude tumour rates gives a much more significant result than the prevalence analysis. This is because only four control animals died prior to terminal sacrifice when all but one of the high-dose animals died, while only one high-dose animal survived to terminal sacrifice when the majority of the

Table 5.15 Contingency tables for comparison of control group and high-dose group from data on 1,2-dichloroethane

$\mathcal{I}_1 = [62, 79)$

d_i	0	2
Y_{i1}	0	13
$N_{i1} - Y_{i1}$	3	21

$\mathcal{I}_2 = [80, 85)$

Y_{i2}	0	1
$N_{i2} - Y_{i2}$	1	0

$\mathcal{I}_3$: terminal sacrifice

Y_{i3}	2	1
$N_{i3} - Y_{i3}$	31	0

control animals contribute to the prevalence analysis. For larger experiments, the efficiency of the prevalence analysis should improve relative to the crude tumour rate analysis. It is important to note that the inefficiency of the interval prevalence method in this example is due to the large difference in intercurrent mortality rates. If intercurrent mortality rates are equal in all groups, then the interval prevalence method will, in general, be more efficient than the crude tumour rate analysis (McKnight, 1981).

Dinse and Lagakos (1983) proposed a logistic regression method for analysing nonlethal tumour data. Their method does not require selection of time intervals but, rather, makes use of the time of death of each animal. Dinse and Lagakos assume that the tumour prevalence rate at time t for animals in the group exposed to dose level d_i of the test compound is given by

$$P(d_i; t) = \exp\{\gamma(t) + \delta_i\}/[1 + \exp\{\gamma(t) + \delta_i\}], \tag{5.11}$$

where $\gamma(t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_r t^r$. The carcinogenic potential of the test compound is assessed by testing the null hypothesis $H_0: \boldsymbol{\delta} = \mathbf{0}$, where $\boldsymbol{\delta}' = (\delta_0, \delta_1, \ldots, \delta_I)$ is a vector of group-specific parameters. The validity of the test of H_0 rests on the assumption that the prevalence function under H_0 can be represented adequately by the logistic function

$$P(t) = \exp\{\gamma(t)\}/[1 + \exp\{\gamma(t)\}]. \tag{5.12}$$

The polynomial $\gamma(t)$ will be referred to as the prevalence log-odds function. As in Section 5.3, denote the kth time at which animal deaths are observed by t_k, and let x_{ik} denote the number of deaths observed in group i at time t_k, $k = 1, \ldots, K$; $i = 0, 1, \ldots, I$. Similarly, let Y_{ik} denote the number of animals in which a tumour is found at necropsy among the x_{ik} animals from group i dying at t_k. Let $\hat{\gamma}(t) = \hat{\beta}_0 + \hat{\beta}_1 t + \hat{\beta}_2 t^2 + \cdots + \hat{\beta}_r t^r$, where $\hat{\boldsymbol{\beta}}$ denotes the maximum likelihood estimator of $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_r)'$ under $H_0: \boldsymbol{\delta} = 0$. It follows that score tests of H_0 can be based on the $(I + 1)$ statistics

$$\hat{D}_i = \sum_{k=1}^{K} (Y_{ik} - x_{ik}\hat{P}_k),$$

where $\hat{P}_k = \exp\{\hat{\gamma}(t_k)\}/[1 + \exp\{\hat{\gamma}(t_k)\}]$. Note that $\hat{D}_i$ can be written as $O_i - \hat{E}_i$, where O_i is the observed number of animals in group i in which tumours were discovered, and $\hat{E}_i$ is the expected number calculated on the basis of the estimated polynomial prevalence log-odds function. The covariance matrix $\hat{\mathbf{V}}$ of the vector $\hat{\mathbf{D}} = (\hat{D}_0, \hat{D}_1, \ldots, \hat{D}_I)'$ can be obtained using standard score test methodology, and a test of $H_0: \boldsymbol{\delta} = \mathbf{0}$ can be based on

$$\hat{X}^2_{PH} = \hat{\mathbf{D}}'\hat{\mathbf{V}}^{-}\hat{\mathbf{D}},$$

which will have an asymptotic chi-squared distribution with I degrees of freedom under H_0, provided the prevalence function under H_0 can be described by (5.12) with $\gamma(t)$ an r degree polynomial in time. Similarly, writing $\delta_i = \Delta d_i$ for all i, a test for monotone trend in response can be derived as a score test of $H_0: \Delta = 0$, which leads to

the test statistic

$$\hat{X}^2_{PT} = (\mathbf{d}'\hat{\mathbf{D}})^2/(\mathbf{d}'\hat{\mathbf{V}}\mathbf{d}).$$

Provided the null prevalence function can be described by (5.12) with $\gamma(t)$ an r degree polynomial, the statistic $\hat{X}^2_{PT}$ will be asymptotically distributed, under H_0, as a chi-squared random variable with one degree of freedom.

Although the method of Dinse and Lagakos (1983) does not require selection of time intervals, an appropriate degree polynomial must be selected to estimate the prevalence function. The importance of the choice of r is illustrated by the data from the 1,2-dichloroethane experiment. Using the data from all three exposure groups and assuming a linear prevalence log-odds function (i.e., $r = 1$ in $\gamma(t)$), one finds that $\hat{X}^2_{PT} = 17.67$. For the same data, assuming a quadratic prevalence log-odds function (i.e., $r = 2$ in $\gamma(t)$), one finds that $\hat{X}^2_{PT} = 9.93$. Similarly, when comparing the high-dose group to the control group, deleting the data from the low-dose group, one finds $\hat{X}^2_{PT} = 17.44$ with a linear prevalence log-odds function and $\hat{X}^2_{PT} = 6.04$ with a quadratic prevalence log-odds function. It should be noted that the disparity in the results obtained with linear and quadratic prevalence log-odds functions in this example is due primarily to the large differences in intercurrent mortality rates among groups. In such a situation, different choices of interval can similarly lead to widely disparate results using analysis based on $\hat{X}^2_{PT}$. McKnight (1985) has noted that, in cases of extreme differences in intercurrent mortality rates, all methods of time adjustment eventually break down.

This example raises the important issue of the need for further research on methods for choosing the degree of the polynomial, $\gamma(t)$. Using all three exposure groups and fitting (5.12) to the 1,2-dichloroethane data, the model with parameters β_0, β_1 and β_2 provides a significantly better fit than the model with only parameters β_0 and β_1 ($p = 0.0023$). Thus, selection of the best fitting polynomial, in the absence of information on exposure level, would lead to selection of a quadratic prevalence log-odds function. Letting $\delta_i = \Delta d_i$ for all i and fitting (5.11) to the 1,2-dichloroethane data, the model based on β_0, β_1, β_2 and Δ provides little improvement in fit over the model based on β_0, β_1 and Δ ($p = 0.65$). Thus, selection of the best fitting polynomial, in the presence of information on exposure level, would lead to selection of the linear prevalence log-odds function, and to the corresponding finding of a stronger association between exposure to the test compound and tumour prevalence.

In a simulation study comparing $\hat{X}^2_{PT}$ with linear and quadratic prevalence log-odds functions to X^2_{PT} with a variety of interval selection methods, Dinse (1985) simulated tumour prevalence functions based on Weibull distributed times to tumour. These prevalence functions are clearly not linear in time on the logistic scale. Under the null hypothesis, results of the test based on $\hat{X}^2_{PT}$ with a quadratic prevalence log-odds function agreed more closely with those of tests based on X^2_{PT} using the various interval selection methods than did those of the test based on $\hat{X}^2_{PT}$ using a linear prevalence function. All tests considered in the simulation study tended to reject too often in cases when mortality increased with dose level, the test based on $\hat{X}^2_{PT}$ with a linear prevalence log-odds function rejecting most often. Hence, on the basis of the above examples and of limited simulation results, it would seem prudent to select the degree

of the polynomial, $\gamma(t)$, by fitting the model in (5.11) with $\boldsymbol{\delta} = \mathbf{0}$, and to test the significance of successively higher-degree polynomials at a moderate significance level, say 10–20%.

Although the availability of a method for nonlethal tumours which avoids the need for interval selection is desirable, unqualified recommendation of the method of Dinse and Lagakos (1983) must await further investigation of issues related to polynomial selection. Hitchcock (1966) showed that logistic regression tests such as that based on $\hat{X}^2_{PT}$ usually offer only slight gains in efficiency over comparable stratification methods such as tests based on X^2_{PT}, and Gart (1977) verified this finding in a specific application. The simulation study of Dinse (1985) offers further verification. Whatever the slight gain in efficiency of the test based on $\hat{X}^2_{PT}$, it may be offset by the invalidity of the test when the assumed logistic relation over time does not hold (see Cox, 1966). Provided that the tumour prevalence does not change rapidly in any of the selected time intervals, the test based on X^2_{PT} will be valid regardless of whether a linear relation or some higher polynomial in time holds for tumour prevalence rates. Nevertheless, the logistic regression method of Dinse and Lagakos provides an attractive alternative to interval-based methods, particularly since it provides a statistical framework within which the potential ambiguities introduced by the need to choose intervals or polynomials can be resolved. A recently proposed method based on weighted prevalence estimators (Selwyn *et al.*, 1985) requires neither interval nor polynomial selection, and thus warrants further investigation.

5.6 Analysis of rapidly lethal occult tumours and of observable tumours

In this section we consider statistical methods that use the information on times to tumour, or times to death because of tumour, more precisely. We discuss two kinds of experiments:

(1) studies in which the specific tumour under study is found at necropsy, i.e., is occult, but is assumed to be rapidly lethal, and

(2) studies of easily observable tumours in living animals, such as those in skin-painting experiments, or experiments in which the endpoint is a palpable tumour.

For such studies, we show how to calculate the curves for survival without apparent tumour and give methods for comparing such curves.

In the first kind of study, the tumour when found at death is usually assumed to cause the death, even if the animal died accidentally or was sacrificed because it was moribund. It may be questioned whether animals found with a tumour at scheduled sacrifice should also be assumed to have a lethal tumour. If the tumour is truly rapidly lethal, very few should be found at sacrifice. However, for the terminal sacrifice the fatal/incidental distinction is not essential, since all animals that are killed at this terminal sacrifice were the only ones at risk of dying of a lethal tumour, and considering these tumours as either lethal or incidental will not alter the results. Such questions are discussed more fully in the next section. In any case, the analyses require knowledge of the times of death of all the animals and a categorization of animals into those bearing the tumour of interest and those not bearing the tumour of interest.

Studies in which the tumour is directly observable are simpler to analyse. Here, it is necessary to know the time at which the tumour is first seen in any animal; the times of any subsequent appearance, disappearance, reappearance, appearance of additional tumours, or death are not required for analyses of this type. We do require knowledge of the times of death of all animals that die without ever getting the tumour.

The data are usually recorded in time units of days or weeks and are divided into sets labelled 'uncensored' or 'censored' for animals with and without tumours, respectively. Our analyses require specification of all time points at which tumours are found in any group, denoted, as in Section 5.2, by $t_1, \ldots, t_k, \ldots, t_K$. The number of animals with tumour found at t_k in group i is denoted by y_{ik}. The total number of animals with tumour found over the course of the experiment is then $y_{i.} = \sum_{k=1}^{K} y_{ik}$. Note that $y_{i.}$ corresponds to y_i from Section 5.4, where the time index was suppressed. In general, the number at risk at t_k for group i is m_{ik}. Thus, m_{i1} corresponds to m_i in Section 5.4. Successive values of $m_{i,k+1}$ are found to be $m_{ik} - y_{ik}$ less the number dying in group i in the interval $[t_k, t_{k+1})$. That is, those dying at t_{k+1} are still included in $m_{i,k+1}$. Note that for experiments on lethal tumours, m_{ik} consists of all animals alive at the beginning of t_k, while for observable tumours m_{ik} excludes those living animals that already have a tumour.

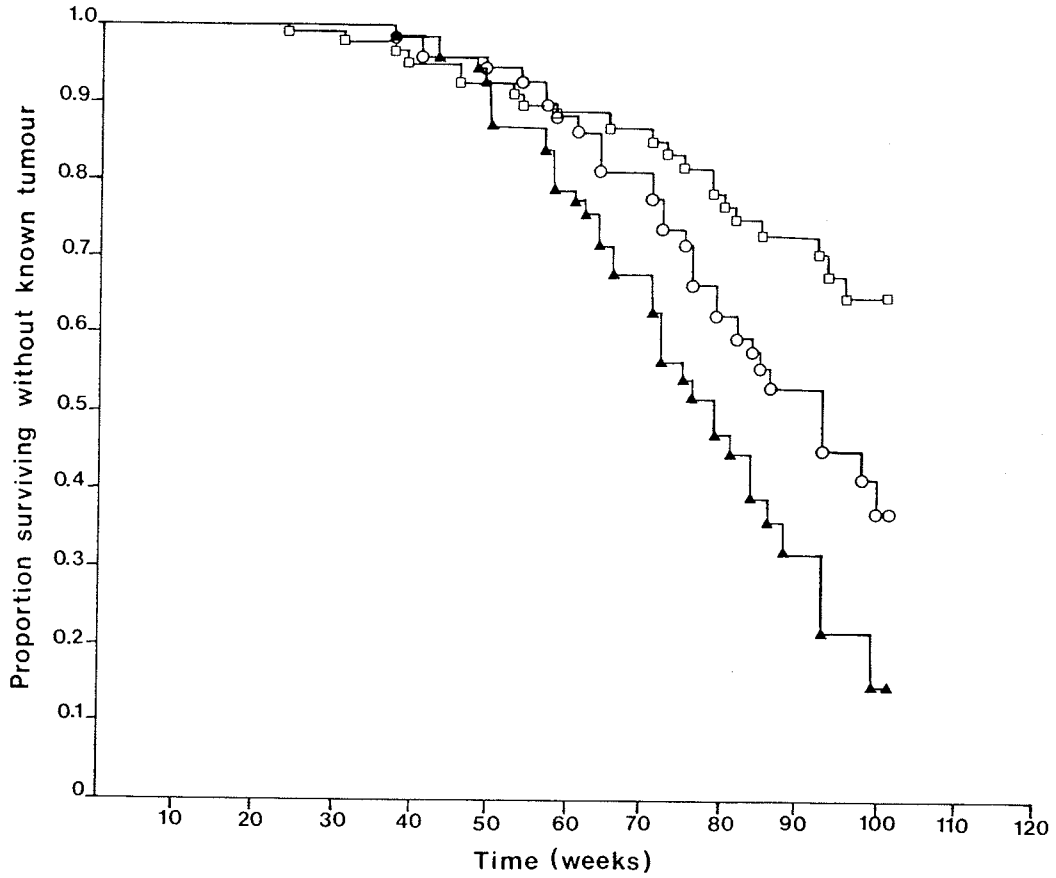
We illustrate this notation by considering the data from Section 4.2 on observable tumours induced by painting cigar-smoke condensate in groups D, E and F from Table 4.2. No untreated control group, which would naturally be indexed with 0, is present in this example. We therefore index the three groups 0, 1 and 2, for increasing dose, giving index 0 to the group receiving 59 mg cigar-smoke condensate per week (group F). For numerical convenience, we subtract 59 from the actual doses and code them as 0, 19 and 44. Note that $I = 2$ in this example. Table 5.16 gives a partial listing of the data that have $K = 37$ distinct time points with tumour.

Calculation of the Kaplan–Meier estimator of the survival function without known tumour follows easily from the formulae given in Section 5.3. (It is interesting to note that this procedure was originally used in the same framework for analysing the occurrence of tumours in skin-painting experiments by Miescher *et al.*, 1941.) If y_{ik} is substituted for x_{ik} and m_{ik} for n_{ik}, the formulae for $\hat{S}_i(t)$ and $V\{\hat{S}_i(t)\}$ apply to the

Table 5.16 Summary of ages at detection of observable tumours in data on cigar-smoke condensate

Tumour time point (weeks)		y_{ik}/m_{ik}			Total
		i:	0	1	2	
k	t_k	d_i:	0	19	44	$y_{.k}/m_{.k}$
1	24		1/83	0/81	0/79	1/243
2	31		1/79	0/77	0/73	1/229
3	38		1/76	1/74	1/69	3/219
⋮	⋮		⋮	⋮	⋮	⋮
36	99		0/19	1/9	1/3	2/31
37	100		0/18	0/9	1/2	1/29
Total: $y_{i.}$			21	31	37	$y_{..} = 89$

Fig. 5.2 Kaplan–Meier estimates of tumour-free survival curves for three groups of female mice (□, 59 mg/week; ○, 78 mg/week; ▲, 103 mg/week) painted with cigar-smoke condensate



present analysis. The curves are plotted similarly. (See Figure 5.2 for the plot of the cigar-smoke condensate study.)

The statistical tests for possible differences in these curves of survival without known tumour are also analogous to the tests of survival curves and of prevalent tumour rates. Log-rank test statistics may be computed using (5.3), (5.4) and (5.5), where D_i and V_{hi} are calculated as in (5.1) and (5.2) after substituting y_{ik} for x_{ik} and m_{ik} for n_{ik}. Let X^2_{LH} denote the test statistic for equality of tumour mortality or incidence rates in the $I+1$ groups computed using (5.3) (Mantel, 1966; Cox, 1972), X^2_{LT} denote the corresponding trend test statistic computed using (5.4) (Tarone, 1975), and X^2_{LQ} denote the corresponding departure from the trend statistic computed using (5.5) (Tarone, 1975). Furthermore, Z_{LT} may denote the corresponding one-tailed normal deviate for trend.

Alternatively, modified Wilcoxon rank sum tests can be employed. Wilcoxon test statistics may be computed using (5.8), (5.9) and (5.10), where D_{iW} and V_{hiW} are calculated as in (5.6) and (5.7) after substituting y_{ik} for x_{ik} and m_{ik} for n_{ik}. As discussed previously, the modified Wilcoxon test is more sensitive for detecting differences in the curves early in the experiment. For experiments with heavy intercurrent mortality, the Peto–Prentice-modified Wilcoxon statistics should be used (Prentice & Marek, 1979).

We turn now to the estimation of the strength of association of tumour effect with dose. First, consider the assumptions that Cox (1972) made in deriving the log-rank

test. Let $\lambda_i(t)$, the hazard rate function in group i, be the (instantaneous) probability of a tumour in the time interval $(t, t+\Delta t)$. Cox assumed proportional hazard rates, i.e., $\lambda_i(t)/\lambda_0(t) = \rho_i$ for all t and $i = 1, \ldots, I$. That is, the incidence of tumours may vary over time, but the pairwise relative risks among all the groups are constant. The homogeneity chi-square tests whether $\rho_i = 1$ for all i, while the trend test is sensitive to log-linear alternatives of the form $\rho_i = \exp(\beta d_i)$. Various estimators have been suggested for ρ_i. The simplest to use is the ratio of ratios of observed and expected values (Pike, 1972),

$$\hat{r}_i = (O_i/E_{iL})/(O_0/E_{0L}), \qquad i = 1, 2, \ldots, I,$$

where E_{iL} $(i = 0, 1, \ldots, I)$ is the expected number of lethal or observable tumours, calculated as described in general terms in Section 5.2.

Breslow (1975) and Bernstein *et al.* (1981) found that this estimator is valid for $\frac{1}{2} \leq \rho \leq 2$, but that it may be biased towards unity for other values. Another rather simply applied estimator is the so-called Mantel–Haenszel estimator, which is essentially a weighted combination of the cross-product ratios:

$$\tilde{r}_i = \left\{ \sum_{k=1}^{K} y_{ik}(m_{0k} - y_{0k})/m_{.k} \right\} \Big/ \left\{ \sum_{k=1}^{K} y_{0k}(m_{ik} - y_{ik})/m_{.k} \right\}.$$

The results of Bernstein *et al.* (1981) indicate that this estimator has small bias away from unity for $\frac{1}{3} \leq \rho_i \leq 3$, but may have large bias otherwise.

Finally, as suggested by Gart (1972), the methods of logistic regression for contingency tables can be applied to either the $2 \times 2 \times K$ or $2 \times (I+1) \times K$ table. From the former, an estimator of ρ_i for each $i = 1, 2, \ldots, I$ can be found from the maximum likelihood estimator, $\hat{\rho}_i$ of the odds ratio (see, e.g., Thomas, 1975). Alternatively, we can fit a stratified logistic model to the $2 \times (I+1) \times K$ table and obtain the maximum likelihood estimator of the common slope, $\tilde{\beta}$ (see, e.g., Thomas & Gart, 1983). The estimators of ρ_i based on the linear logistic model are then $\tilde{\rho}_i = \exp(\tilde{\beta} d_i)$. Either of these estimators tends to be biased away from unity with small sample sizes.

Let us return now to the example of painting with cigar-smoke condensate. We find the one-tailed test for trend is $Z_{LT} = 4.49$, $p = 0.000004$. The comparable test statistic using the Wilcoxon form of the test yields a normal deviate Z_{LT} of 3.46 with $p = 0.00027$. The Wilcoxon form is less significant, as it gives greater weight to the comparisons early in the experiment, where the curve of survival without tumour of the control group is actually lower than that of the dosed groups.

The full set of chi-square analyses by both tests is given in Table 5.17. Chi-squares for heterogeneity are highly significant by both methods. Almost all of this variation is accounted for by the linear trend chi-squares. Thus, neither method finds any evidence of departure from linearity.

Consider now the question of estimating the strength of association between dose and tumour effect, illustrated for this example in Table 5.18. For this example, $\tilde{\beta} = 0.0271 \pm 0.0062$. The good agreement between $\tilde{\rho}_i$, which assumes linearity on the logistic scale, and the other estimates, which do not, reflects the low chi-squares for departure from linearity.

Table 5.17 Summary of analysis of observable tumours from data on cigar-smoke condensate

Source of variation	Degrees of freedom	Log-rank test		Wilcoxon test	
		Chi-square	p	Chi-square	p
Linear trend	1	20.15	0.0000	11.96	0.0005
Departure from linearity	1	0.01	0.9187	0.06	0.8124
Total (heterogeneity)	2	20.16	0.0000	12.02	0.0025

Table 5.18 Estimates of odds ratios from data on cigar-smoke condensate

	Dose d_i		
	0	19	44
$O_i = y_{i\cdot}$	21	31	37
E_{iL}	37.92	29.74	21.35
O_i/E_{iL}	0.554	1.042	1.733
$\hat{r}_i$	1.00	1.88	3.13
$\tilde{r}_i$	1.00	1.88	3.12
$\hat{\rho}_i$	1.00	1.91	3.34
$\tilde{\rho}_i = \exp(\tilde{\beta} d_i)$	1.00	1.67	3.29

Comparison of life-table analyses and analyses of crude proportions

It is not uncommon for analyses of crude proportions to yield substantially the same interpretation as that reached by the more elaborate analyses using the life-table techniques we have just described. If there is more intercurrent mortality among the high-dose groups (see Chapter 2, Table 2.2), the more sophisticated analyses will probably yield a more significant positive or direct association with dose than will the crude analysis. Under these circumstances, existing relationships not found by the crude analysis may become apparent in the life-table analysis.

If the intercurrent mortality is minimal or roughly equal in the various groups, the crude proportion and life-table analyses will often be quite similar. Cuzick (1982) confirmed this impression theoretically. He found that crude proportion analysis is over 95% efficient relative to Cox's life-table analysis when less than 50% of the animals have tumours.

It is instructive to compare the results of the crude proportion analysis for the example of cigar-smoke condensate. Note that the first tumour occurred at 24 weeks and the last at 100 weeks. The summary of the data on early and intercurrent mortality is given in Table 5.19. We see that, even after adjusting for early mortality, there remain substantial differences in the percentages of intercurrent mortality among the groups of those at risk at 24 weeks and not getting a tumour subsequently. These range from 74% to 98%.

Table 5.19 Summary of mortality from data on cigar-smoke condensate

	Dose d_i			Total
	0	19	44	
n_i	100	100	100	$n_. = 300$
Death before $t = 24$	17	19	21	57
m_{i1}	83	81	79	$m_{.1} = 243$
Interim deaths (24–100 weeks)	46	46	41	133
y_i	21	31	37	
$m_{i1} - y_i$	62	50	42	
Interim deaths (% of $m_{i1} - y_i$)	74%	92%	98%	

For both the life table and the crude proportions tests, we use the identical total observed numbers of animals with tumours, but the expected values are calculated differently (see Section 5.4). Table 5.20 presents the results.

Table 5.20 Observed and expected values calculated by two methods (life-table and crude proportions) from data on cigar-smoke condensate.

	Dose d_i			Total
	0	19	44	
$O_i = y_i$	21	31	37	$y_{..} = 89$
E_{iL}	37.42	29.74	21.35	89
$O_i - E_{iL}$	−16.92	1.26	15.65	0
$E_i = (m_{i1} y_{..})/m_{.1}$	30.40	29.67	28.93	89
$O_i - E_i$	−9.40	1.33	8.07	0

Because of the high intercurrent mortality in the high-dose group, its expectation is much lower in the life-table analyses than it is in the crude proportion analysis, which does not take the differential mortality into account. The inverse relation holds for the low-dose group, which has lower intercurrent mortality. Thus, deviations of the expected from the observed values are larger in the life-table analysis. The results of the crude proportion analysis are shown in Table 5.21.

Table 5.21 Summary of crude proportion analysis of data on cigar-smoke condensate

Source of variation	Degrees of freedom	Crude proportion tests	
		Chi-square	p
Linear trend	1	7.88	0.0050
Departure from linearity	1	0.31	0.5771
Total (heterogeneity)	2	8.19	0.0166

Table 5.22 Estimates of odds ratio from crude proportion analysis of data on cigar-smoke condensate

Dose	d_i	0	19	44
Odds ratio	$\hat{R}_i$	1.00	1.83	2.60
	$\tilde{R}_i$	1.00	1.49	2.52

The trend and homogeneity tests are still significant but less so than by either of the life-table analyses. Recalling the discussion of Section 5.4, the associated measures of association are shown in Table 5.22. For this analysis, we found that $\tilde{\beta} = 0.0210 \pm 0.0075$. These estimates are somewhat lower, particularly at the high dose, than the estimates found from the life-table analysis. Thus, although both analyses imply similar qualitative interpretations, the life-table analyses yield more highly significant tests with larger measures of association. This is a quite common result and represents an example of outcome type *A* of Table 2.2.

Alternative life-table and exact tests

When the experiment is small, the question arises whether the life-table tests are valid. Simulations performed by Tarone (1975) and Latta (1981) indicate that the fit of the test statistics to the appropriate chi-square distributions is quite good in large sample sizes. However, when the numbers are small, these tests may reject the null hypothesis too often. For small expected values, say, the minimum of $E_{iL} < 5$, a conservative version of the life-table chi-square tests has been suggested (Peto & Pike, 1973). However, Gart (1975) and Haybittle and Freedman (1979) point out circumstances in which application of the conservative test must be used with caution. If only one of the groups, say, i, has animals at risk beyond time t', then this should be made the final time, t_K, for the analyses based on the conservative test. If tumours occur for $t > t'$ in the ith group, then their time points are omitted from calculation of the observed and expected. Thus, $O_i < y_{i.}$ in such instances.

For very small numbers, exact versions of these tests can, in principle, be done. However, this involves an additional assumption requiring that the hazard rate for deaths without tumour in each group be proportionally related to its hazard rate for tumour incidence. Such tests are described by Tarone (1975) and Cox (1959).

A test for acceleration (as defined in Section 2.2) has been developed by Breslow *et al.* (1984). Their test statistic is of the form given in equation (5.6), with w_k taken to be the estimated cumulative hazard function (using the pooled data from control and exposed groups) at t_k, the kth ordered time of death due to tumour. Because acceleration is unlikely in experiments with inbred strains, this test should be used in conjunction with a test such as the log-rank test which has power against more general alternatives. Accordingly, appropriate adjustment for multiple comparisons is required in those cases in which the acceleration test is employed (Breslow *et al.*, 1984).

Combination of tests

The approximate test for trend can easily be combined over experiments, as with the analysis of crude proportions. One simply adds the numerators of the Z_{TL} and divides by the square root of the sums of squares of the individual denominators. The resulting statistic is an approximate normal deviate. Combination of the homogeneity and departure chi-squares usually requires a computer program (see, e.g., Thomas & Gart, 1983).

Issues of multiple comparisons

When the statistical analysis is done by comparing in turn each dose group, d_i, to the control, d_0, the question of multiple testing can be handled by the Bonferroni inequality, as it was for the analysis of crude proportions (see Section 5.4). In addition, because we have suggested two life-table adjusted tests for trend, another question of multiple testing arises. It is invalid to compute routinely both tests and report only the one that yields the higher significance. Tarone (1981) has given a method for adjusting the p-value for the more extreme of these two tests. The adjustment method is derived explicitly for the Breslow-modified Wilcoxon; however, the method applies also to the Peto–Prentice-modified Wilcoxon, after substituting the Peto–Prentice weights for the Breslow weights throughout.

5.7 Analysis of occult tumour data when contexts of observation are known

Although the analysis described in Section 5.5 is valid if all tumours of a particular type are nonlethal and the analysis described in Section 5.6 is valid if all tumours of a particular type are rapidly lethal, the relationship between the presence of a tumour and death of host animals often lies between these two extremes. Some tumours of a particular type may be, for all practical purposes, nonlethal, while other tumours of the same type may contribute to the death of their host animals. As noted in Chapter 2, differences among exposure groups with respect to intercurrent mortality can cause serious bias in tests for equality of tumour rates. For data on occult tumours, analysis assuming that all tumours are lethal when, in fact, some are nonlethal, and analysis assuming that all tumours are nonlethal when, in fact, some are lethal, can lead to incorrect inferences if intercurrent mortality rates differ among exposure groups (Peto *et al.*, 1980; Lagakos, 1982). In the common situation in which intercurrent mortality rates increase with increasing dose level, the tumorigenic effect will be overstated in the first case (i.e., assuming all tumours are lethal when some are nonlethal) and understated in the second case (i.e., assuming all tumours are nonlethal when some are lethal).

In order to avoid biases due to differences in intercurrent mortality and at the same time to make some use of data on time of death, Peto (1974) and Peto *et al.* (1980) recommended that pathologists assign a context of observation to each observed tumour (Section 2.10). A tumour that either directly or indirectly kills its host is said to be observed in a fatal context. A tumour that is observed at necropsy of an animal that

died of some cause unrelated to the tumour is said to be observed in an incidental context. Although it may be difficult to determine the contexts of observation for some tumours, it is assumed in this section that all tumours of a particular type have been classified as either fatal or incidental.

The analysis of data on occult tumours using contexts of observation is based on the methods given in Sections 5.5 and 5.6. The analysis of incidental tumours is a straightforward modification of the analysis of nonlethal tumours presented in Section 5.5. A modification is necessary, because those animals killed by the tumour in question (i.e., animals for which the tumour is observed in a fatal context) should not enter into the analysis of incidental tumours. As in the analysis of Section 5.5, the length of the experiment is subdivided into distinct time intervals. Within each time interval, the data on incidental tumours may be summarized in a table such as Table 5.3, with N_{ij} corresponding to the number of animals in group i dying during interval j from causes unrelated to the presence of the tumour in question, and Y_{ij} corresponding to the number of these animals in which the tumour was observed in the incidental context, for $i = 0, 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$. All tumours found in animals killed in planned sacrifices are classified as incidental. Using the methods of Section 5.3, we form a vector $\mathbf{D}_P$ of differences of expected from observed values for the data on incidental tumours and compute the corresponding covariance matrix $\mathbf{V}_P$.

Analysis of tumours observed in the fatal context is based on the methods in Section 5.6. At each time t_k at which the tumour in question is observed in the fatal context, a contingency table like Table 5.16 can be formed, where y_{ik} corresponds to the number of animals in group i for which the tumour was observed in the fatal context at t_k, and m_{ik} corresponds to the number of animals in group i surviving (and thus still at risk of being killed by a tumour) to time t_k. Note that, in the analysis of fatal tumours, animals in which the tumour is observed in the incidental context are treated exactly as all other animals not killed by the tumour. As in Section 5.6, a vector $\mathbf{D}_L$ of differences of expected from observed values is formed using the fatal tumour data, and the corresponding covariance matrix $\mathbf{V}_L$ is computed.

The analysis of data on occult tumours using contexts of observation is based on the vector $\mathbf{D}_C = \mathbf{D}_P + \mathbf{D}_L$, with covariance matrix $\mathbf{V}_C = \mathbf{V}_P + \mathbf{V}_L$. Test statistics for heterogeneity X^2_{CH}, trend X^2_{CT}, and departure from trend (X^2_{CQ}) may be calculated as in (5.3), (5.4) and (5.5), respectively, with $\mathbf{D}_C$ substituted for $\mathbf{D}$ and $\mathbf{V}_C$ substituted for $\mathbf{V}$.

As noted earlier, it may be difficult to determine the contexts of observation of some tumours. Accordingly, Peto *et al.* (1980) suggest that tumours be classified on an ordinal scale, namely, 1 if a tumour is definitely incidental, 2 if a tumour is probably incidental, 3 if a tumour is probably fatal and 4 if a tumour is definitely fatal. Usually, the above analysis would be performed with tumours classified as 1 or 2 taken to be incidental and tumours classified as 3 or 4 taken to be fatal. With the ordinal classification, however, the analysis can be repeated using different cut points (e.g., classifications 1, 2 and 3 taken as incidental and classification 4 taken as fatal) to determine if inferences are influenced by possible misclassification of tumour context. For further discussions of the problems associated with the assignment of causes of death, see Lagakos (1982) and Kodell *et al.* (1982b).

Consider the data on pituitary tumours given in Table 4.3 for 16 groups of male Colworth rats exposed to increasing dose levels of *N*-nitrosodimethylamine. As noted by Peto *et al.* (1980), treatment-induced, fatal liver tumours caused a marked, dose-related increase in mortality. Pituitary tumours were observed in both fatal and incidental contexts and were classified on the ordinal scale described in the preceding paragraph. The first part of Table 5.23 gives the observed and expected numbers of animals with tumours found in a fatal context; these are tumours recorded as fatal, or probably fatal, with the codes 4 or 3 in Table 4.3. The observed and expected numbers were derived using the methods of Section 5.6, and the vector of their difference is $\mathbf{D}_L$, the corresponding covariance matrix $\mathbf{V}_L$ not being displayed.

The code -3 in Table 4.3 was used for animals that were totally cannibalized or autolysed; their cause of death was not ascertainable. These animals were considered in the analysis of fatal tumours as if they had died on day 1 without a tumour and were excluded from the incidental tumour analysis.

The code -2 in Table 4.3 was used for animals whose head was cannibalized or autolysed so that the presence or absence of a pituitary tumour was not ascertainable but death was known not to be caused by a pituitary tumour. These animals were considered with their respective time to death as having no fatal pituitary tumour in the analysis of fatal tumours but were excluded from the analysis of incidental tumours.

For the prevalence analysis, the adaptive method of determining subintervals of the time axis, outlined in Section 5.5, was used and resulted in the following ten intervals:

$$[87, 591], [596, 680], [681, 796], [797, 891], [892, 905], [908, 964],$$
$$[965, 1028], [1029, 1030], [1033, 1071], [1073, 1234].$$

The 17 animals with codes -2 and -3 were, as explained above, excluded from the prevalence analysis. Thus, the number of animals considered in each group for the analysis of incidental tumours is not always equal to the number of animals considered in the analysis of fatal tumours less the number of fatal tumours observed.

The second part of Table 5.23 gives the calculated observed and expected numbers for each group in the prevalence analysis. Their difference is the vector $\mathbf{D}_P$, the corresponding covariance matrix $\mathbf{V}_P$ not being displayed.

In the third part of Table 5.23, the numbers of tumours observed and expected in either context are summed for each group. The difference is the vector $\mathbf{D}_C$, the corresponding covariance matrix being $\mathbf{V}_C = \mathbf{V}_L + \mathbf{V}_P$.

The chi-square statistic for heterogeneity calculated according to (5.3) for the combined situation is $X^2_{CH} = 12.11$. Because all animals in the highest dose group died before any pituitary tumour was observed in the experiment, the observed and expected numbers are zero. The expected number of tumours in group 15 is virtually zero ($E_{15} = 0.02$), and thus group 15 was also deleted prior to the computation of test statistics. The rank of the matrix $\mathbf{V}_C$ is therefore 13 rather than 15. Comparison of the computed X^2_{CH} to the percentiles of a chi-square distribution with 13 degrees of freedom gives a p-value of 0.52.

Using the scores $0, 1, \ldots, 12, 13$ as dose levels, corresponding roughly to a

Table 5.23 Pituitary tumours in male Colworth rats: observed and expected numbers of tumours by context of observation and experimental group

Group	Tumours observed in a fatal context			Tumours observed in an incidental context			Combined	
	No. of animals	Observed events	Expected events	No. of animals	Observed events	Expected events	Observed events	Expected events
1	192	26	32.24	159	24	24.12	50	56.36
2	48	10	9.44	38	6	6.44	16	15.88
3	48	8	8.11	40	6	5.96	14	14.07
4	48	5	8.45	41	9	6.63	14	15.08
5	48	8	7.77	39	3	5.46	11	13.23
6	48	10	9.48	38	8	6.82	18	16.30
7	48	9	7.14	39	3	5.14	12	12.28
8	48	11	5.92	36	6	4.76	17	10.68
9	48	5	7.09	43	7	6.22	12	13.31
10	48	6	3.43	40	5	3.41	11	6.83
11	48	5	2.70	43	1	2.71	6	5.41
12	48	1	1.85	47	1	2.33	2	4.18
13	48	0	0.23	48	2	0.59	2	0.82
14	48	0	0.14	48	0	0.41	0	0.55
15	48	0	0.02	47	0	0.00	0	0.02
16	48	0	0.00	45	0	0.00	0	0.00

logarithmic transformation of the actual dose levels (see Section 4.3), gives, according to (5.4), a trend statistic $X^2_{CT} = 1.66$ with a two-sided p-value of 0.20.

Thus, it appears that *N*-nitrosodimethylamine does not induce pituitary tumours in male Colworth rats.

This data set is of particular interest since, if one did not use the information on the context of observation but considered all the pituitary tumours to be found either in a fatal context or in an incidental context, different conclusions would be derived.

Considering all pituitary tumours as fatal would give, with the life-table methods of Section 5.6, a chi-square statistic for heterogeneity $X^2_{LH} = 28.34$ with 13 degrees of freedom ($p = 0.008$). The one-degree-of-freedom chi-square statistic for trend would be $X^2_{LT} = 7.38$ ($p = 0.007$), indicating a positive trend in the occurrence of pituitary tumours with increasing dose. Considering all tumours as incidental would give, using the prevalence methods of Section 5.5, a heterogeneity statistic $X^2_{PH} = 18.04$ with 13 degrees of freedom ($p = 0.156$). The one-degree-of-freedom chi-square statistic for trend would be $X^2_{PT} = 2.85$ ($p = 0.091$), suggesting a negative trend in occurrence of pituitary tumours with increasing dose.

As noted in Chapter 2, the combination of two analyses, one based on tumour death rates and the other based ostensibly on tumour prevalence rates, may seem somewhat contrived. In fact, it is difficult to justify this analysis rigorously. The tests based on X^2_{CH} and X^2_{CT} can be shown to test the null hypothesis of interest, that is, that the underlying tumour onset rates are equal in all exposure groups, only under rigid assumptions (Lagakos, 1982; McKnight & Crowley, 1984). Nevertheless, this analysis

is a useful attempt to solve the difficult problem of using data on age at death in evaluating data on occult tumours. Because the underlying tumour onset rates are not identifiable (Lagakos, 1982; McKnight & Crowley, 1984), any test for equality of onset rates will be only approximate, and efforts to improve upon the analyses in this section are likely to require changes in experimental design. For example, McKnight and Crowley (1984) have shown that tumour onset rates are approximately identifiable in experiments with frequent planned sacrifices. Thus, it is likely that better tests can be developed, but only at the cost of additional animals. Methods of testing for differences in tumour incidence rates using data from planned sacrifices are now available (McKnight & Crowley, 1984; Dewanji & Kalbfleisch, 1986), and research in this area is progressing rapidly.

LIST OF ESSENTIAL SYMBOLS – CHAPTER 5 (in order of appearance)

$I+1$ number of experimental groups ($i = 0, 1, \ldots, I$)
d_i dose level ($d_0 = 0$, $d_1 < \cdots < d_I$)
t_k time of observation of an event (death, occurrence of tumour) ($k = 1, \ldots, K$)
x_{ik} number of events in group i at time t_k
n_{ik} number of animals at risk (to experience event) in group i at time t_k
A_{ik} proportion of animals at risk in group i compared to total of all groups at time t_k
E_{ik} expected number of events in group i at time t_k
O_i total number of events in group i
E_i total number of expected events in group i
D_i difference between O_i and E_i
$\mathbf{D}$ vector of D_i's
$\mathbf{V}$ covariance matrix of vector $\mathbf{D}$
V_{hi} element of $\mathbf{V}$ ($h, i = 0, 1, \ldots, I$)
X_H^2 test statistic for heterogeneity (chi-squared distribution with I degrees of freedom)
X_T^2 two-sided test statistic for linear trend (chi-squared distribution with 1 degree of freedom)
X_Q^2 test statistic for departure from linear trend (chi-squared distribution with $I-1$ degrees of freedom)

(N.B.: A subscript W to the above quantities D_i, $\mathbf{D}$, $\mathbf{V}$, V_{hi}, X_H^2, X_T^2 and X_Q^2 indicates their derivation using non-negative weights w_k ($k = 1, \ldots, K$))

$\hat{S}_i(t)$ Kaplan–Meier estimate of survival function in group i
$V\{\hat{S}_i(t)\}$ variance of $\hat{S}_i(t)$
y_i number of animals with tumour in group i ($i = 0, 1, \ldots, I$)
m_i number of animals at risk in group i ($i = 0, 1, \ldots, I$)
$\hat{R}_i$ ratio of odds of a tumour in group i to corresponding odds in control group
Z one-tailed test statistic for difference of two proportions (normal distribution); subscript c when used with continuity correction

Z_T one-tailed test statistic for monotone trend (normal distribution); subscript c when used with continuity correction
$\tilde{R}_i$ odds ratio estimate from regression model
T duration of experiment
$\mathscr{I}_j$ subinterval of experimental time spanning from 0 to T ($j = 1, \ldots, J$)
Y_{ij} number of animals with tumour among those that died or were killed in group i in interval j
N_{ij} number of animals that died or were killed in group i in interval j

(N.B.: Subscript P, L and C given to quantities X_H^2, X_T^2 and X_Q^2 when derived in prevalence analysis (Section 5.5), life-table analysis (Section 5.6) or analysis using context of observation (Section 5.7))

$\lambda_i(t)$ hazard rate function in group i
ρ_i ratio of hazard rates between group i and control group: $\lambda_i(t)/\lambda_0(t)$
$\hat{r}_i$ estimate of ρ_i by ratio of observed and expected values ($i = 1, \ldots, I$)
$\tilde{r}_i$ estimate of ρ_i by weighted odds ratios ($i = 1, \ldots, I$)