DRAFT 2

Statistical methods for analysis of histopathological data with special reference to rank tests for graded observations

Authors: P.N. Lee & J.S. Fry

Date : 9.4.86

1. Introduction

Over the last few years, considerable advances have been made in statistical methods for analysis of histopathological data. An extensive up-to-date review of these methods is shortly to be published by the IARC (Gart et al, 1986) running to well over 300 pages. However, most of these methods are only applicable to presence/absence data, i.e. an animal has a particular cancer or it does not, and do not take account of the grade of the condition being studied, whereas our experience is that pathologists frequently record such a grade, not only scoring tumours as benign or malignant, but also commonly scoring non-neoplastic conditions as 0=absent, 1=minimal, 2=slight, 3=moderate, 4=severe, 5=very severe.

The purpose of this document is to point out that there are relatively simple methods based on rank tests for analysis of graded data analogous to all the standard methods used for analysis of presence/absence data, to describe these methods (some of which are new or not previously efficiently programmed, at least as far as we

are aware) and to show how the recommended methods for graded and presence/absence data interrelate.

The methods of analysis described for graded and presence/absence data relate to combinations of 3 factors:

(a) the type of comparison being made between groups,
(b) whether the test is asymptotic or exact,
(c) whether or not stratification is involved.

These three factors are explained briefly below:

Types of between-group comparisons

3 types of group comparisons are normally made:

(i) 2 group comparison. Typically the response in each treated group in turn is compared with that in the control group, although one may require comparison of response in specific treatment groups.

(ii) overall between group comparison. This test, often called a test of homogeneity (or heterogeneity), shows whether as a whole the groups vary significantly in response. It is not particularly informative on its own as it does not indicate where the significant variation arises from.

(iii) test for dose-related trend. Where, as is often the case, the experimental groups are different dose levels of the same substance, a test as to whether response varies linearly with dose can be particularly informative. While a significant trend

need not imply all dose levels have an effect, a test for trend can be much more powerful in detecting an effect of treatment than a 2-group comparison.

Asymptotic and exact tests

Given the total number of animals in each group and given the total number of animals responding at each level, it is in theory possible to calculate exactly the ("conditional") probability of having obtained a result at least as extreme as that observed. In practice, where numbers of animals are large, the sheer computational time rules out the possibility of carrying out such an exact calculation.

As an alternative approach to an exact test, ~~the alternative~~ is to calculate a test statistic from the observed data which is a close approximation to some standard statistic, the frequency distribution of which is known. Often this standard statistic is the chisquared statistic on 1 degree of freedom for which the 95th percentage point is the value of 3.84, as can be seen from standard tables. In this case one would calculate the approximate statistic and if it exceeded 3.84 deem the test to have shown significance at the 95% confidence level. The problem with such an approach is that the test is only asymptotic, i.e. the approximation improves with increasing numbers of animals but only becomes perfect for an infinite number. While the approximation may be quite poor for very small counts, the approximation rapidly improves with increasing numbers and in general is excellent well before the numbers have got too large to make exact tests impractical.

Stratification

In the simplest situation one is concerned with a situation where, if treatment has no effect, the probability of any animal responding is the same, i.e. all the animals in the analysis can be considered essentially homogeneous in respect of factors other than treatment. A typical situation might be when one is considering male survivors at 90 days in a subchronic toxicity study in which the males were randomly allocated to treatment group at the outset. Such data can be treated as a single stratum and be analysed by methods appropriate for unstratified analysis.

Commonly, however, the probability of response varies markedly by factors other than treatment, e.g. sex, age or body weight, but it is desired to obtain an overall view of the effect of treatment on response for the whole data. In these circumstances one divides the data into mutually exclusive sets by these other factors so that within any set ("stratum") the animals can be considered essentially homogeneous (apart from treatment). One then analyses the data by stratified methods, which involves making comparisons within strata and then combining the comparisons.

It is important to note that use of unstratified analysis when an important factor other than treatment exists is not advisable. Where the factor is correlated with the treatment, a biassed assessment of the relationship between treatment and response will be obtained. Even where it is not correlated, the unstratified analysis may be much less powerful.

Carrying out analyses stratified for factors unaffected by treatment is conceptually straightforward. Care should be taken however in the interpretation of analyses stratified for factors affected by treatment. If, for instance, treatment increases bodyweight and tumour incidence and if analysis stratified for bodyweight eliminates the treatment/tumour relationship, one should be wary about stating that treatment does not cause an increase in tumours. Here it would be better to present both stratified and unstratified analyses as evidence that treatment affects tumour incidence, but that this can be explained in terms of its effect on body weight.

ROELEE 84

Over the last few years a computer program for entry, validation, reporting and statistical analysis of histopathological data has been developed by us, in close conjunction with Dr. F.J.C. Roe and a number of other pathologists. All the statistical tests described in this report are available in this program except where clearly stated.

2. Structure of this report

In the sections that follow unstratified tests are considered first and then stratified tests, and within these, asymptotic tests are considered before exact tests. Where the test for ranked data is well known, this is described first and it it shown how the corresponding test for presence/absence data arises as a special case. Where it is less well known, the more ~~well known~~ familiar test for presence/absence data is considered first, with the test for ranked data arising as a generalisation. Asymptotic tests are considered for every circumstance, exact tests only in some cases. Details of computing algorithms for exact tests are relegated to appendices. Simple worked examples are given in many situations.

Key formulae are numbered. Table 1 summarises the formulae and sections considering the various tests.

The notation used is summarized in the following section, in the order it appears.

References to the better known formulae are not given specifically in each case. The reader is referred to Breslow and Day (1980) for presence/absence methods and to Conover (1980) for rank methods.

TABLE 1

Tests considered in this report

Type of test	Type of comparison 2-group	k-group	Trend
Unstratified			
Asymptotic			
Rank	Section 5,F5	Section 4,F1-3	Section 6,F8
Presence/absence	Section 5,F6&7	Section 4,F4	Section 6,F9
Exact			
Rank	Section 9,F11;A2	Section 12	Section 12
Presence/absence	Section 7,F10;A1	Section 12	Section 8,F11;A1
Stratified			
Asymptotic			
Rank	Section 11,F15	Section 11,F17	Section 11,F16
Presence/absence	Section 10,F12	Section 10,F14	Section 10,F13
Exact			
Rank and Presence/absence	Section 12	Section 12	Section 12

F = Formula
A = Appendix
N.B. Tests considered under Section 12 are not available in ROELEE 84

3. Notation

k	number of groups being compared
i	subscript for group ($i=1,2...k$)
n_i	number of observations in group i
N	total number of observations in all groups ($n = \sum n_i$) (N.B. summation is over group unless stated)
j	subscript for member of group ($j=1,2...n_i$)
X_{ij}	observations of jth member of ith group
$R(X_{ij})$	rank assigned to X_{ij}
R_i	sum of ranks for ith group ($R_i = \sum_{j=1}^{n_i} R(X_{ij})$)
T	a test statistic calculated according to a formula, subscripted by the number of this formula
$\sum'$	summation over all N members
V^2	an estimate of variance
C	correction factor for Kruskal-Wallis test
q	number of runs of ties of length 1 or greater - equivalently number of rows of a contingency table layout of the data, groups being envisaged as columns
u	subscript for run/row
t_u	length of uth run of ties - equivalently the row total of a contingency table
O_{ui}	observed numbers in k x q contingency table

$\bar{R}_u$	mean rank for row u of a contingency table
O_i	observed numbers in k x 2 contingency table
E_i	expected numbers in k x 2 contingency table
Y_{ij}	generalized response of jth member of ith group
$\bar{Y}$	mean response for all N animals
$\bar{d}$	mean dose for all N animals
$\bar{D}$	dose-related trend statistic
$\bar{Y}_i$	mean response in group i
Z_i	deviation of group mean response from overall group mean
Q_i	$n_i - O_i$
W	$\sum d_i O_i$
S	number of strata
s	subscript for strata (s=1,2...S)

4. Asymptotic tests based on ranks for between group comparisons for a single stratum

Let us assign rank 1 to the smallest of the N observations, rank 2 to the second smallest and so on to the largest of all the observations which receives rank N. If several observations are equal to each other, assign the average rank to each of the tied observations.

The Kruskal-Wallis test for overall between group variation is then given by: FORMULA 1

$$T_1 = (\sum_i R_i^2/n_i - N(N+1)^2/4)/V^2$$

where

$$V^2 = (\sum_{ij} R(X_{ij})^2 - N(N+1)^2/4)/(N-1)$$

T_1 is asymptotically distributed as chi-squared with k-1 degrees of freedom.

If there are no ties V^2 simplifies to $N(N+1)/12$, and then the the test statistic reduces to: FORMULA 2

$$T_2 = (12 \sum_i R_i^2/n_i)/N(N+1) - 3(N+1)$$

An alternative form of the test with ties is to compute: FORMULA 3

$$T_3 = T_2/C$$

where C, the correction factor is given by

$$C = 1 - \left(\sum_{u=1}^{q} t_j^3 - \sum_{u=1}^{q} t_j \right) / (N^3 - N)$$

where there are q runs of ties of length $t_1, t_2 \ldots t_u \ldots t_q$. Note that it is convenient to consider a run of any length, even 1, as a run of ties so that the formula can be rewritten:

$$C = 1 - \left(\sum_{u=1}^{q} t_j^3 - N \right) / (N^3 - N)$$

If there are no ties at all, clearly C = 1. If there are only relatively short runs of ties, the correction factor will be close to 1, and T_2 will be a very good approximation to T_1 or T_3 (which are mathematically identical). Formula 3 is a more convenient form than formula 1 to use when the Kruskal-Wallis test is applied to a graded condition, where there are many ties (observations with the same grade). Thus, if the data are laid out as follows:

Condition	u	Group 1	2	k	Row totals	$\bar{R}_u$ = Average rank
absent	1	0_{11}	0_{12}	0_{1k}	t_1	$(t_1 + 1)/2$
minimal	2	0_{21}	0_{22}	0_{2k}	t_2	$t_1 + (t_2 + 1)/2$
			.....			
severe	q	0_{q1}	0_{q2}	0_{qk}	t_q	$\sum_{u=1}^{q-1} t_i + (t_q + 1)/2$
Column totals		n_1	n_2	n_k	N = Grand total	

C is readily calculated from the previous formula by summing the cubes of the row totals, while T_2 is readily calculated from Formula 2 noting that

$$R_i = \sum_{u=1}^{q} O_{ui} \bar{R}_u$$

An extreme situation of ties is when the condition can only be observed at 2 levels, i.e. a presence/absence condition, normally analysed by statistical methods appropriate for 2 x k contingency tables. Here the data are normally laid out in the form:

	Group 1	Group 2	.....	Group k	Total
Present	O_1	O_2		O_k	t_1
Absent	$(n_1 - O_1)$	$(n_2 - O_2)$		$(n_k - O_k)$	t_2
Total	n_1	n_2		n_k	N

and the classical asymptotic chisquared test statistic for testing overall between group variation is: FORMULA 4

$$T_4 = (N - 1)(1/t_1 + 1/t_2) \sum_i (O_i - E_i)^2/n_i$$

where E_i, the expected number with the condition in group i is given by

$$E_i = t_1 n_i / N$$

It is important to note that if T_1 (or its equivalent T_3) is applied to such data, exactly the same answers are reached as if T_4 is used.

Formula 1 is thus of general application for testing between-group variation. At one end of the spectrum, it is well known that if the data derive from a continuous normal distribution, the asymptotic relative efficiency (ARE) of the Kruskal-Wallis test relative to the classical one-way analysis of variance F-test is 0.955 and may be much higher than 1 for certain types of non-normality, particularly data containing outliers or occasional true extreme values.

At the other end of the spectrum, it is identical to the classical 2 x k chisquared formula for the most heavily tied situation of all, when only two values can occur. Furthermore Conover (1980) has stated that, as a result of recent advances in the theory of rank tests, there should no longer be any hesitation to apply Formula 1 to situations that have many ties.

Given biological data are rarely normally distributed and often heavily skewed, there seems to be no real reason from a statistical point of view why one should not use the Kruskal-Wallis test routinely when dealing with essentially continuous data, with graded

data or with presence-absence data. From a presentational point of view, since scientists are more familiar with their means and standard deviations and with their chisquared tests than with rank tests, one may continue in practice to use one-way analyses of variance for data that are fairly normal and to use classical chisquared tests for contingency tables. A problem with the Kruskal-Wallis test is how to present the findings in terms of a mean response in each group. For essentially continuous data, the appropriate statistic to use is the median rather than the mean. For presence-absence data, one can continue to give the proportion responding. For graded data, the median is not helpful as it will often be the same in each group even where there is significant between-group variation. More meaningful here is the mean rank as it indicates which groups are above and which below the overall mean rank $(N + 1)/2$. If the mean grade is presented it should be made clear that significance tests were not based on them - mean grades assume a scale to the grading system (e.g. 2 animals with grade 2 are equivalent to 1 animal with grade 1 and 1 animal with grade 3) but rank tests do not.

EXAMPLE

Grade	Group 1	2	3	Raw Totals	Mean Ranks
1	4	10	6	20	10.5
2	14	6	7	27	34
3	17	9	8	34	64.5
4	6	7	6	19	91
5	2	6	1	9	105
Total number of animals	43	38	28	109	
Rank sums R_i	2370.5	2156.5	1468	5995	

(N.B. Check overall sum of ranks 5995 = N(N + 1)/2)

Formula 1

$$\sum_{ij} R(X_{ij})^2 = 20 \times 10.5^2 + 27 \times 34^2 + 34 \times 64.5^2 + 19 \times 91^2 + 9 \times 105^2 = 431429.5$$

$$V^2 = 941.708333$$

$$\sum_i R_i^2/n_i = 330027.2214$$

$$T_1 = 0.320929$$

Formula 3

$$T_2 = 0.302474$$

$$\sum_{u=1}^{q} t_j^3 = 74575$$

$$C = 1 - \frac{74575 - 109}{1295029 - 109} = 0.942494$$

$$T_3 = 0.320929$$

If we now collapse the table containing grades 3, 4, 5 as condition present and grades 1, 2 as condition absent we have:

	Group 1	2	3	Raw Totals	Mean Ranks
Present	25	22	15	62	31.5
Absent	18	16	13	47	86
Total	43	38	28	109	
Rank sums R_i	2335.5	2069.0	1590.5	5995	

Formula 1

$$\sum R(X_{ij})^2 = 62 \times 31.5^2 + 47 \times 86^2$$

$$= 409131.5$$

$$V^2 = 735.245370$$

$$\sum R_i^2/n_i = 329847.9241$$

$$T_1 = 0.167188$$

Formula 4

$O_1 = 25$ $O_2 = 22$ $O_3 = 15$

$E_1 = 24.458716$ $E_2 = 21.614679$ $E_3 = 15.926606$

(N.B. Check sum of E = sum of O = 62)

$$\sum_i (O_i - E_i)^2/n = 0.041385$$

$$T_4 = 0.167188$$

It might be interesting to know what amount an exact test would give for comparison?

5. Asymptotic tests based on ranks for unstratified 2 group comparison

Formula 1 is a perfectly valid asymptotic test for 2 group comparisons as well as for k group comparisons. It should be noted, however, that in pairwise group comparison for a k group study each comparison must be based on a ranking of the observations specifically in the two groups being compared.

An alternative form for Formula 1 specifically for 2-group comparison, commonly known as the Mann-Whitney test or the Wilcoxon test is : FORMULA 5

$$T_5 = \frac{R_1 - n_1(N+1)/2}{\sqrt{n_1 n_2 \sum R^2/(N(N-1) - n_1 n_2 (N+1)^2/4(N-1)}}$$

$\sum R(x_{ij})^2$

which is asymptotically distributed as a normal variate with mean 0 and variance 1. Positive values of T_5 indicate a greater response in group 1 than in group 2. When squared it becomes an asymptotic chi-squared statistic with 1 degree of freedom, yielding identical answers to that from Formula 1.

In the special case of a 2 x 2 contingency table an alternative to Formula 1 or 5 is the uncorrected chisquared statistic : FORMULA 6

$$T_6 = \frac{(O_1(n_2 - O_2) - O_2(n_1 - O_1))^2(N-1)}{n_1 n_2 t_1 t_2}$$

In this particular case it is well known that the corrected chisquared statistic : FORMULA 7

$$T_6 = \frac{(|O_1(n_2 - O_2) - O_2(n_1 - O_1)| - N/2)^2(N - 1)}{n_1 n_2 t_1 t_2}$$

yields a better approximation to the p-values obtained from the exact tests discussed below. It appears that this correction (the N/2 in the formula) is only an improvement for 2 x 2 tables and not for 2 x k tables (k groups, presence/absence variable) or for k x 2 tables (2 groups, graded variable with k levels).

Formulae 6 and 7 can also be rewritten in the form

$$T_6 = (O - E)^2/varO$$

or $$T_7 = (|O - E| - 1/2)^2/varO$$

where O is the observed number in any of the 4 cells of the 2 x 2 table, E is the expected number under the null hypothesis, and var(O) is the variance of the observed number. If we take the "presence" cell in group 1, we have

$$O = O_1$$

$$E = n_1 t_1 /N$$

$$varO = n_1 n_2 t_1 t_2 /N^2(N - 1)$$

EXAMPLE

Using the same data as before for groups 1 and 2, we have:

Grade	Group 1	Group 2	Raw Tables	Mean Ranks
1	4	10	14	7.5
2	14	6	20	24.5
3	17	9	26	47.5
4	6	7	13	67
5	2	6	8	77.5
Total	43	38	81	
Rank sums R_i	1737.5	1583.5	3321	(= 81 x 82/2)

Formula 1

$$\sum' R(X_{ij})^2 = 14 \times 7.5^2 + .. 8 \times 77.5^2 = 177862$$

$$V^2 = 521.2625$$

$$\sum_i R_i^2/n_i = 136193.2339$$

$$T_1 = 0.061838$$

Formula 5

$$T_1 - n_1 (N + 1)/2 = - 25.5$$

$$n_1 n_2 \sum' R^2/N(N - 1) = 44849.76975$$

$$\sum' R(X_{ij})^2/$$

$n_1 n_2 (N + 1)^2/4(N - 1) = 34334.42500$

$T_5 = 0.248673$

$T_5^2 = 0.061838$

To illustrate the case of a 2 x 2 contingency table we have:

	Group 1	Group 2	Raw Tables	Mean Ranks
Present	7	20	27	14.0
Absent	23	8	31	43.0
Total	30	28	58	
Rank R_i	1087.0	624.0	1711	(= 58 x 59/2)

Formula 1

$\sum_{ij} R(X_{ij}')^2 = 27 \times 14^2 + 31 \times 43^2 = 62611$

$V^2 = 212.921053$

$\sum_i R_i^2/n_i = 53291.9190$

$T_1 = 13.232224$

Formula 6

$$(O_1(n_2 - O_2) - O_2(n_1 - O_1)) = 404$$

$$T_6 = 13.232224$$

Formula 7

$$(|O_1(n_2 - O_2) - O_2(n_1 - O_1)| - N/2) = 375$$

$$T_7 = 11.400730$$

6. Asymptotic tests based on ranks for dose-related trend for a single stratum

Often in a k group study the groups represent different doses, d_i, of a particular treatment. In these circumstances it is usually of more interest to know whether there is a relationship between dose and response than to know the results of a general test of between-group homogeneity. If the mean response for all the animals is $\bar{Y}$ and the mean dose applied is $\bar{d}$, it is natural to consider the statistic

$$D = \sum_{ij}' (Y_{ij} - \bar{Y})(d_{ij} - \bar{d})$$

where the summation is over all animals, the dose for an individual animal being written for generality as d_{ij} rather than d_i. The steeper the slope of the dose-response relationship the larger the value of this statistic, which is termed a test for linear trend or often simply a test for dose-related trend. This statistic can be written

$$D = \sum_{ij}' Y_{ij} d_{ij} - \sum_{ij}' Y_{ij} \sum_{ij}' d_{ij} / N$$

where the first term only depends on the individual values, the second term being constant for given marginal totals.

Rewriting in terms of the group means $\bar{Y}_i$ and d_i we have

$$D = \sum_i n_i d_i \bar{Y}_i - \sum_i (n_i \bar{Y}_i) \sum_i (n_i d_i) / N$$

If we define the deviation of the group mean from the expected value under the null hypothesis as

$$Z_i = \bar{Y}_i - \bar{Y}$$

D can be rewritten as

$$D = \sum_i n_i d_i Z_i$$

Alternatively, D may be rewritten as

$$D = \sum_i n_i (d_i - \bar{d})\bar{Y}_i$$

which is of the form

$$D = \sum_i a_i \bar{Y}_i \quad \text{with} \quad \sum_i a_i = 0$$

Although the latter form is as discussed by Marascuilo and McSweeney (1967) in their paper proposing a rank trend test, the form in terms of Z_i is more convenient for accumulation over strata and the one we will use.

Where the data are ranks, a test for trend is given by the statistic : FORMULA 8

$$T_8 = D^2 / \text{CvarD}$$

where

$$D = \sum_i n_i d_i (R_i/n_i - (N + 1)/2)$$

and

$$\text{var } D = (N + 1)(N\sum_i d_i^2 n_i - (\sum_i d_i n_i)^2)/12$$

C being the same correction for ties as in formula 3.

T_8 is asymptotically distributed as chisquared on one degree of freedom.

If there are only 2 possible data values, i.e. presence/absence data, this formula collapses to Armitage's test for trend: FORMULA 9

$$T_9 = D_2^2/\text{var}D_2$$

where

$$D_2 = \sum_i d_i (O_i - E_i)$$

and

$$\text{var}D_2 = t_1 (\sum_i d_i^2 E_i - (\sum_i d_i E_i)^2/t_2)(N - 1)$$

Although Armitage's trend test is widely used for presence/absence data, Formula 8 is not commonly used to test for trend with graded data. One possible reason for this is that if one is using non-parametric methods it is seen by some to be unsatisfactory to use the actual dose levels in the formula, which implicitly implies some sort of scale. While the author feels from considerable experience of the test that it is a very useful and sensitive method of picking up departures from the null hypothesis,

two points should be made. Firstly, it would be perfectly possible to use the rank of the dose applied rather than the dose itself to make this a fully non-parametric method. Secondly, a commonly used alternative test, that of Jonckheere (1954), (which essentially is based on comparing the number of pairwise comparisons of individual data items which show a difference in the direction suggested by the difference in dose with the corresponding number which show a difference in the opposite direction) is clearly less sensitive in the common situation where a high dose has an effect but lower doses do not. The same objection can be levelled at a test based on a sum of two-sample rank tests (comparing group 1 with 2 to k, 1 + 2 with 3 to k and so on) as proposed by Wahrendorf (in Gart et al 1986).

EXAMPLE

If we have data as follows

Group	i	1	2	3	4	
Dose	d_i	0	1	3	6	
Response					Total	Mean Rank
Grade 1		3	2	2	1	8	4.5
Grade 2		5	5	5	1	16	16.5
Grade 3		2	4	4	3	13	31
Grade 4		0	0	1	3	4	39.5
Total	n_i	10	11	12	8	41	
Rank sums	R_i	158.0	215.5	255.0	232.5	861 (= 41 x 42/2)	
Rank means R_i/n_i		15.80	19.5909	21.25	29.0625	21.0	
Deviations from expected		-5.20	-1.4091	+0.25	+8.0625		
$n_i d_i$		0	11	36	48	95	

$D = 380.50$

$$C = 1 - \frac{(8^3 + 16^3 + 13^3 + 4^3) - 41}{41^3 - 41}$$

$$= 0.900871$$

$d_i^2 n_i$	0	11	108	288	407

$VarD = 26817$

$T_8 = 5.992892$

Note that if we use formulae 2 and 3 to compute an asymptotic chisquared statistic for overall between group variation we have

$$T_3 = 6.289090$$

which only exceeds T_8 by 0.296. As 0.296 does not approach significance on 2 d.f. as a chisquared statistic and 5.993 is quite highly significant ($p<0.02$) on 1 d.f., the data are well described by the linear trend.

If we now collapse the data into a 2 x 4 table taking grades 1 and 2 as absent and grades 3 and 4 as present we have:

i	1	2	3	4		
d_i	0	1	3	6	Total	Rank
Absent	8	7	7	2	24	12.5
Prsent	2	4	5	6	17	33.0
Total	10	11	12	8	41	45.5
R_i	166.0	219.5	252.5	223.0	861	
R_i/n_i	16.60	19.9545	21.0417	27.875	21	
Deviations	-4.40	-1.0455	+0.0417	+6.875		
n_i/d_i	0	11	36	48	95	

$$D = 320.00$$

$$C = 1 - \frac{(24^3 + 17^3) - 41}{41^3 - 41}$$

$$= 0.728571$$

$T_8 = 5.241042$

E_i	4.1463	4.5610	4.9756	3.3171	17
O_i	2	4	5	6	17

$D_2 = 15.609756$

$\sum d_i {}^2E_1 = 168.7560975$

$\sum d_i E_i = 39.3902439$

$VarD_2 = 46.491612$

$T_9 = 5.241042$

(N.B. In the above example, we have reversed the order of absent and present in the table but this of course makes no difference to the answers.)

7. Exact tests for presence/absence data - 2 group comparisons

Formula 7, the corrected chisquared statistic, will generally suffice to give accurate p values for a 2 x 2 contingency table provided that the expected frequencies for all 4 cells are at least 5 under the null hypothesis. For smaller samples, or when in doubt, recourse should be made to Fisher's exact test. In this test one noted that for a given 2 x 2 table

	Group 1	Group 2	
Absent	Q_1	Q_2	t_1
Present	0_1	0_2	t_2
	n_1	n_2	N

the probability of observing these frequencies, given the null hypothesis with fixed marginal totals, is: FORMULA 10

$$P = \frac{t_1! \, t_2! \, n_1! \, n_2!}{0_1! \, 0_2! \, Q_1! \, Q_2! \, N!}$$

The one-tailed probability of observing a result as or more extreme than that actually observed is given by summing the probabilities from all such tables consistent with the marginal totals. Thus if the alternative to the hypothesis is that group 2 will increase frequency of the condition, one sums the probabilities from the tables:

$$\begin{vmatrix} Q & Q \\ 1 & 2 \\ & \\ 0 & 0 \\ 1 & 2 \end{vmatrix} \qquad \begin{vmatrix} Q+1 & Q-1 \\ 1 & 2 \\ & \\ 0-1 & 0+1 \\ 1 & 2 \end{vmatrix} \qquad \cdots\cdots \qquad \begin{vmatrix} Q+u & Q-u \\ 1 & 2 \\ & \\ 0-u & 0+u \\ 1 & 2 \end{vmatrix}$$

were $u = \min (n_2 - 0_2 , t_2 - 0_2) = \min (0_1 , Q_2)$

If the alternative to the hypothesis is that group 2 will decrease frequency of the condition, one sums the probabilities from the tables:

$$\begin{vmatrix} Q & Q \\ 1 & 2 \\ & \\ 0 & 0 \\ 1 & 2 \end{vmatrix} \qquad \begin{vmatrix} Q-1 & Q+1 \\ 1 & 2 \\ & \\ 0+1 & 0-1 \\ 1 & 2 \end{vmatrix} \qquad \cdots\cdots \qquad \begin{vmatrix} Q-v & Q+v \\ 1 & 2 \\ & \\ 0+v & 0-v \\ 1 & 2 \end{vmatrix}$$

where $v = \min (Q_1 , 0_2)$

If a two-tailed probability is required it is normal to double the lower of these two probabilities. In practice it is obvious from the data which probability to calculate when the hypothesis is in doubt. ~~What about J. Tox. Environ. Health 15, 745-767 (1985)?~~

An algorithm for rapid calculation of this exact test has been programmed by Dr. J.S. Fry and is described in Appendix 1. It is incorporated in ROELEE 84 and is sufficiently rapid to make its use practical for any 2 x 2 table of the size normally met in animal studies.

I don't agree with this – the p value cannot be calculated by accumulating on each side of the hypergeometric distribution e.g. in the case of 4/45 versus 32/92

$p_0 \rightarrow p_4 = 7.07955 \times 10^{-4}$

and then it again falls below p_4 at p_{21}

$p_{21} \rightarrow p_{32} = 2.16732 \times 10^{-4}$

$\therefore p$ 2 tailed $= 9.24689 \times 10^{-4}$ and not 1.41591×10^{-3}

EXAMPLE

Assuming the data

	Control	Treated	
Absent	49	37	86
Present	1	8	9
Total	50	45	95

it is obvious that the likely alternative to the null hypothesis is that treatment increases response. The probability associated with the table as exists (i.e. with 8 treated animals having the condition) is:

$$P_8 = \frac{50!\ 45!\ 86!\ 9!}{49!\ 37!\ 8!\ 1!\ 95!}$$

$$= \frac{50 \times 9 \times 45 \times 44 \times 43 \times 42 \times 41 \times 40 \times 39 \times 38}{95 \times 94 \times 93 \times 92 \times 91 \times 90 \times 89 \times 88 \times 87}$$

The probability associated with the only more extreme table, where 9 treated animals have the condition is:

$$P_9 = \frac{50!\ 45!\ 86!\ 9!}{\cancel{48}50!\ 36!\ 9!\ 0!\ 95!}$$

$$= \frac{45 \times 44 \times 43 \times 42 \times 41 \times 40 \times 39 \times 38 \times 37}{95 \times 94 \times 93 \times 92 \times 91 \times 90 \times 89 \times 88 \times 87}$$

P_9 can be calculated as 0.000754186

$$P_8 = \frac{50 \times 9}{37} \quad P_9 = 0.009172536$$

so the total 1-tailed probability = 0.009927,

i.e. a 2-tailed probability of 0.0198534

8. Exact tests for presence/absence data - unstratified test for trend

As noted in the section on asymptotic tests for trend, the trend statistic contains a term which is constant for given marginal totals. It is clear therefore that an exact test for trend can be calculated by considering only the remaining term and summing the probabilities for those tables which give a value as or more extreme than that observed. For presence/absence data this reduces to summing the probabilities of those tables which give a value of $W = \sum_i d_i O_i$ equal to or greater than that observed, where d_i is the dose in group i and O_i the number within the condition of interest.

For a given 2 x k table

	Group				
	1	2	.....	k	Total
Absent	Q_1	Q_2		Q_k	t_1
Present	O_1	O_2		O_k	t_2
Total	n_1	n_2		n_k	N

the probability, given the null hypothesis and the marginal totals is: FORMULA 11

$$P = \frac{n_1! \, n_2! \, \ldots\ldots \, n_k! \, t_1! \, t_2!}{O_1! \, Q_1! \, \ldots\ldots \, O_k! \, Q_k! \, N!}$$

Dr. Fry has also programmed an algorithm for calculation of this exact test for trend, which is described in Appendix 2. It is also incorporated into ROELEE 84. It is slower than the Fisher exact test algorithm and as a result the algorithm will not be attempted for

larger tables.

For small samples it is practical to carry out the test by hand and it is often easier than one would think to determine all the tables for which W is at least as great as in the original table. In some tables, it is worth noting that the exact probability does not depend on the actual dose values at all, but only on their ordering. If, for example, all the positive values occur in the top dose level, then the test collapses to a 2 x 2 comparison of top dose versus other groups combined, which is independent of the actual dose values.

EXAMPLE

	Group 1	2	3	4	Total
Q_i	5	4	4	2	15
O_i	0	1	1	3	5
Total	5	5	5	5	20
d_i	0	1	3	6	

Observed value of W = 1 + 3 + 18 = 22

A W value of 22 or more is obtained if

(i) $O_4 = 5$ Table

0	0	0	5	5
5	5	5	0	15
5	5	5	5	20

W = 30 P = 15! x 5!/20!

(ii) $O_4 = 4$ with the other positive value anywhere in groups 1-3. (W = 24, 25 or 28). These can be combined to give

Table

1	4	5	
14	1	15	
15	5	20	

P = 15 x 5 x 15! x 5!/20!

(iii) $O_4 = 3$, $O_3 = 2$, W = 24

Table

0	0	2	3	5
5	5	3	2	15
5	5	5	5	20

P = 10 x 10 x 15! x 5!/20!

(iv) $O_4 = 3$, $O_3 = 1$, $O_2 = 1$, W = 22

Table

0	1	1	3	5
5	4	4	2	15
5	5	5	5	20

P = 5 x 5 x 10 x 15! x 5!/20!

(v) No other table

$$\text{Total one-tailed probability} = \frac{15! \times 5!}{20!}(1 + 75 + 100 + 250)$$

$$= 0.027477$$

$$\therefore \text{ Two-tailed trend p value } = 0.054954$$

9. Exact tests for graded data - unstratified 2 group comparison

Here the table is as follows

	Group 1	Group 2
Grade 1	O_{11}	O_{12}
Grade 2	O_{21}	O_{22}
.		
.		
.		
Grade q	O_{q1}	O_{q2}
Total	n_1	n_2

Here one has to sum the probabilities of all tables which give a value of W equal to or greater than the observed value $W = \sum_{u=1}^{q} \bar{R}_u O_{ui}$, $\bar{R}_u$ being the mean rank associated with grade level u. This is precisely the same mathematical exercise as the test for trend for presence/absence data (described in Appendix 2) with R_u replacing d_i, the O's being considered vertically rather than horizontally.

EXAMPLE

	Group 1	2	Total	Rank
Grade 1	0	4	4	2.5
Grade 2	1	1	2	5.5
Grade 3	3	1	4	8.5
Grade 4	2	0	2	11.5
	6	6	12	

$W = 2 \times 11.5 + 3 \times 8.5 + 1 \times 5.5 = 54.0$

Relationship of W to values of O_{11}, O_{21}, O_{31}, O_{41}.

	O_{11}	O_{21}	O_{31}	O_{41}	W	
1)	0	0	4	2	57	largest possible W
2)	0	1	3	2	54	
3)	1	0	3	2	51	
4)	0	2	2	2	51	
5)	0	1	4	1	51	

It is clear only tables 1) and 2) are as or more extreme.

Table 1

0	4	4
0	2	2
4	0	4
2	0	2
6	6	12

$P = 6! \times 6!/12!$

Table 2

0	4	4
1	1	2
3	1	4
2	0	2
6	6	12

$P = 2 \times 4 \times 6! \times 6!/12!$

Total exact one-tailed probability $= 9 \times 6! \times 6!/12! = 0.00974026$

$\therefore$ Two-tailed probability $= 0.0194805$

10. Presence/absence data - asymptotic stratified tests for 2 group comparison, k group comparison and dose-related trend

In the section on asymptotic tests for unstratified 2 group comparison of presence/absence data, the final formula

$$T_7 = (|O - E| - 1/2)^2/varO$$

gives the test statistic in terms of a deviation of observed from expected response and the variance of the observed response. Within-strata deviations, $O_s - E_s$, and variances, $varO_s$, may be accumulated to form an overall stratified test: FORMULA 12

$$T_{12} = \left(\left|\sum_{s=1}^{S} O_s - \sum_{s=1}^{S} E_s\right| - 1/2\right)^2 \Big/ \sum_{s=1}^{S} varO$$

Note that the continuity correction of 1/2 is not accumulated over strata.

A similar method may be used to carry out a stratified test for trend based on the unstratified formula 9. This yields FORMULA 13

$$T_{13} = \left(\sum_{s=1}^{S} D_{2s}\right)^2 \Big/ \sum_{s=1}^{S} varD_{2s}$$

where D_{2s} and $varD_{2s}$ refer to the values of D_2 and $varD_2$ of formula 9 calculated for stratum s.

The stratified test for k group homogeneity is rather less straightforward. For a single stratum the data are as follows.

	Group 1	2	...	k	Total
Present	O_1	O_2		O_k	t_1
Absent	Q_1	Q_2		Q_k	t_2
	n_1	n_2		n_k	N

Corresponding to each of groups 1, 2 ... k-1 we accumulate over strata the following

$$O_i$$

$$E_i = t_1 n_i / N$$

$$\text{var}O_i = n_i (N - n_i) t_1 t_2 / N^2 (N - 1)$$

$$\text{cov}(O_i O_h) = - n_i n_h t_1 t_2 / N^2 (N - 1)$$

If we denote by O., E. and V. the accumulated vector of observed and expected values and variance-covariance matrix the stratified test for k group homogeneity is given by: FORMULA 14

$$T_{14} = (O. - E.)^T V.^{-1} (O. - E.)$$

T_{14} is asymptotically distributed as chi-squared on k-1 degrees of freedom. This reduces to formula 4 if there is only a single stratum.

11. Ranked data - asymptotic stratified tests for 2 group comparison, k group comparison and dose-related trend

Tests analogous to those in section 10 for presence/absence data can be computed for ranked data, by defining observed and expected mean responses and trend statistics as follows for a stratum:

$$O_i = R_i/(N/2)$$

$$E_i = (n_i(N+1)/2)/(N/2)$$

$$D = \sum d_i(O_i - E_i)$$

noting that

$$\mathrm{Var}R_i = C(N+1)(N-n_i)n_i/12$$

$$\mathrm{cov}(R_i, R_h) = -C(N+1)n_i n_h/12$$

where as before C is the correction factor defined in section 4.

Thus, for a stratified 2-group comparison we consider a particular group and calculate: FORMULA 15

$$T_{15} = \left(\sum_{s=1}^{S} O_s - \sum_{s=1}^{S} E_s\right)^2 / \sum_{s=1}^{S} \mathrm{var}O_s$$

Note that here the continuity correction is not included, and O_s and E_s refer to observed and expected values for the group in question for stratum s.

For a stratified test for trend based on ranks we note that for a stratum

$$\mathrm{var}D_s = C(N+1)(N\sum d_i^2 n_i - (\sum d_i n_i)^2)/3N^2$$

so that we calculate: FORMULA 16

$$T_{16} = \left(\sum_{s=1}^{S} D_s \right)^2 / \sum_{s=1}^{S} \mathrm{var}D_s$$

where D_s is the value of D for stratum s.

For a stratified test of k group homogeneity the method involves matrix manipulation in exactly the same way as for formula 14, except that in FORMULA 17

$$T_{17} = (O. - E.)^T V.^{-1} (O. - E.)$$

the vectors O. and E. and the matrix V. are derived from the formulae for O_i, E_i, $\mathrm{var}R_i$ and $\mathrm{cov}(R_i, R_h)$ noting that

$$\mathrm{var}O_i = \frac{4}{N} \mathrm{var}R_i$$

$$\mathrm{cov}(O_i, O_h) = \frac{4}{N} \mathrm{cov}(R_i, R_h)$$

12. Other exact tests

Other exact tests have not been implemented in ROELEE 84. The exact test for trend in a 2 x k table already takes considerable computing time and more complex exact tests would clearly involve impractically large amounts of computing except for very small data sets.

It should be noted that in the case of the Kruskal-Wallis test (Formulae 1-3) extensive exact tables are given in Iman, Quade and Alexander (1975).

In essence exact tests involve enumerating the probabilities associated with all combinations of the data, given the marginal totals, that produce a value of the asymptotic test statistic as or more extreme than that encountered in the actual data. The problem is the sheer number of combinations.

13. References

Breslow, N.E. & Day, N.E. (1980) Statistical Methods in Cancer Research, Vol. 1, The Analysis of Case-Control Studies, IARC Scientific Publications No. 32, Lyon, International Agency for Research on Cancer.

Conover, W.J. (1980) Practical nonparametric statistics, 2nd Edition. John Wiley & Sons, New York.

Gart, J.J., Krewski, D., Lee, P.N., Tarane, R.E. & Wahrendorf, J. (1986) Statistical methods in cancer research, Volume III: The design and analysis of long-term animal experiments, (IARC Scientific Publications), Lyon, International Agency for Research on Cancer, in the press.

Iman, R.L., Quade, D. & Alexander, D.A. (1975) Exact probability levels for the Kruskal-Wallis test. Selected Tables in Mathematical Statistics, 3, 329-384 (5.2, Appendix).

Jonckheere, A.R. (1954) A distribution-free k-sample test against ordered alternatives. Biometrika, 41, 133-145.

Marascuilo, L.A. & McSweeney, M. (1967) Nonparametric post hoc comparisons for trend. Psychol. Bull., 67, 401-412.

Appendix 1

Programming Fisher's exact test

To be completed.

Appendix 2

Programming the exact test for trend

To be completed.