STATISTICAL ANALYSIS METHODS FOR AVIAN REPRODUCTION EXPERIMENTS

D.A. MacLeod

Technical Report Series No. 211
Headquarters 1994
Canadian Wildlife Service

This publication may be cited as:

MacLeod, D.A. Statistical Analysis Methods for Avian Reproduction Experiments. Technical Report No. 211, Canadian Wildlife Service, Headquarters

Issued under the authority of the
Minister of the Environment
Canadian Wildlife Service

Catalogue No. CW69-5/211E
ISBN 0-662-22642-9

Copies may be obtained from:

Canadian Wildlife Service
National Wildlife Research Centre
100 Gamelin Blvd.
Hull, PQ
K1A 0H3

Statistical Analysis Methods for Avian Reproduction Experiments

Duncan A. MacLeod
Canadian Wildlife Service
Environment Canada

Prepared for:

Pierre Mineau
Brian Collins

Canadian Wildlife Service
100 Gamelin
Hull, Quebec

August 1994

SUMMARY

The Current Situation

The situation with respect to the statistical analysis of data for avian reproduction experiments is a fairly confused one. An examination of the submissions from approximately 100 experiments revealed that a wide variety of different methods are employed, and that many of these are questionable. Assumptions are made that are not statistically valid, and tests are carried out that are inefficient at detecting the kind of treatment effects that are expected. These problems reduce the ability of the experiments to assess whether the test substances have the potential to cause reproductive effects.

There appear to be a number of reasons for this situation:

1. Multiplicity of Variables and Methods

There are several types of variable to be analysed in an experiment, and for some variable types there are a wide variety of methods that could be applied.

2. Current Methods are Too General

The most efficient methods are those that test specifically for a negative effect on reproduction that increases as the dose level increases. But the methods actually employed are usually general methods as they are much better known. These general methods test for any pattern of differences among treatments.

3. Current Methods Ignore Data Structure

Methods should take into account the fact that treatments are applied on a pen basis, but this is difficult and time-consuming and in practice it is much easier to treat the data as if the treatments had been applied independently to each egg, chick or adult bird.

4. Data Complexities

The data sets frequently contain features that complicate the analysis, such as a multi-level structure or variation in the numbers of eggs or chicks per pen.

5. The Objectives are Not Clear

It is not clear whether the effect of the test substance should be tested at each dose level, or whether a general test over all levels is sufficient.

6. Lack of Information in Current Protocols

Existing protocols offer only general guidance on statistical methods and do not discuss the many complexities that can occur.

Main Objective: Identify the Best Methods

It was assumed that when a data set is analysed, the effect of the test substance should be tested at each dose level. The main objective of this report is to identify methods that are statistically valid and fully efficient at carrying out these tests. To achieve this, the following plan was adopted:

- Decide on a set of criteria that the methods should meet in order to ensure their validity and efficiency.

- Classify the variables to be analysed according to their statistical characteristics

- Consider what statistical assumptions are appropriate for each variable class.

- For each variable class and each set of assumptions, identify as many methods as possible that meet the criteria for validity and efficiency.

Results Achieved

The search for improved methods was successful. For each variable class a number of methods have been identified that appear to be a major improvement over current methods. Some of these are extensions of methods currently employed and some represent new approaches.

Recommended Methods

The methods most commonly recommended are:

- weighted t-tests
- weighted tests of linear trend

Weights are employed to accommodate complexities in the data.

Other methods are also recommended in certain circumstances. These include:

Least Squares Methods	Qualitative Methods	Non-Parametric Methods
unweighted t-tests	2x2 chi-square test	Mann-Whitney test
LSD test	Fisher's exact test	Rerandomization tests
unweighted trend tests	Cochran-Armitage test	
Williams test		

Recommended Test Type

For maximum efficiency, all statistical tests should be one-tailed tests that test specifically for the effect most likely to occur - a negative effect on reproduction that increases in magnitude as the dose level increases.

Second Objective: Identification of the NOEC

Once a result is obtained for each dose level in a data set, it is desirable to identify the NOEC (the highest dose level in the experiment at which there is no observed effect of the test substance). A second objective of this report is to examine procedures to determine the NOEC from the pattern of significant or non-significant results at the different dose levels.

Third Objective: Examination of Data Quality

A further objective is to consider the issue of data quality. Because of the nature of avian reproduction experiments, there is a danger that inadequate data quality could seriously reduce the chances of detecting harmful treatment effects. Factors that could affect data quality include mortality, disease, and inconsistency in the birds' reproductive capabilities. Some measures that can be taken to ensure a minimum acceptable quality level are suggested.

Conclusion

Despite the complexities involved in the statistical analysis of data from avian reproduction experiments, a number of promising methods were identified and the prospects for improving the methods of statistical analysis employed in submissions appear to be good.

Further Studies Planned

In a later study, some of the statistical methods recommended in this report will be evaluated further by testing them on actual and simulated data.

ACKNOWLEDGEMENT

The author wishes to thank Brian Collins of the Canadian Wildlife Service (CWS) for statistical advice and information, Pierre Mineau of CWS for advice on biological aspects, and Dora Boersma, formerly with CWS, for advice on biological aspects and for her summaries of the statistical methods currently employed.

TABLE OF CONTENTS

1. INTRODUCTION

1.1 Data to be Analysed for AR Experiments

In avian reproduction (AR) experiments, the treatments are a control and a series of dose levels of the test substance. Each treatment is applied to a specified number of pens, with each pen containing a specified number of male and female birds. Data are collected on a large number of variables (described in section 2). Some of these are measurements (e.g. egg weight), some are counts (e.g. number of eggs laid per pen), and some are proportions (e.g. the per cent of eggs set which hatch per pen). In general each data set has the structure of individual subjects (eggs, chicks or adult birds) within pens within treatments.

1.2 Scope of This Report

The planning and analysis of AR experiments touches on a number of subjects of a statistical nature:

- the design of the experiment
- the selection of the dose levels of the test substance
- the quality of the data produced
- statistical analysis of the data to test the effect of the test substance
- the conclusions to be drawn from the test results

Only some of these topics are within the scope of this report. The main focus is on statistical analysis, with some attention also given to the data quality and to the conclusions to be drawn. The experimental design is not discussed, as it appears to be reasonable to assume that a simple one-way design is employed. The selection of the dose levels is not considered as it is essentially a biological question (and is discussed in Mineau, Boersma and Collins, in press). The only assumption made in this report is that the levels can be considered to be equally spaced in some scale for analysis purposes.

1.3 Objectives

Main Objective: Identify Superior Methods for Testing Effects

The main objective of this report is to evaluate potential methods of data analysis for testing the effect of the test substance, and to identify the most promising ones. In a later study (Collins and Mineau, in prep.) these methods will be examined further by applying them to actual or simulated data.

Second Objective: Determine the NOEC

Another question is that of what conclusions to draw from the test results once they are obtained. The determination of the NOEC, or highest dose at which there is no observed effect, is considered to be an important conclusion in this respect. A second objective of this report is to examine procedures for determining the NOEC from the test results.

Third Objective: Examine Data Quality

There appears to be a danger that the ability of AR experiments to detect harmful effects could be compromised by inadequate data quality, caused by factors such as mortality, disease or reproductive failure unrelated to the treatments applied. A third objective of this report is to examine this aspect and recommend possible measures to ensure that the quality is acceptable.

1.4 The Complexity of the Data Analysis Situation

The statistical analysis of data from AR experiments is not a simple matter, and neither is the evaluation of potential methods. In a typical experiment there are several different types of variables to be analysed (measurements, counts and proportions), each of which has specific features that could affect the analysis.

Another factor is the complexity of the data sets. For many variables the data sets have a multi-level structure, with individual subjects (adult birds, eggs or chicks) grouped within pens and pens grouped within treatments. Another common feature is variation in the number of subjects (particularly eggs or chicks) from one pen to another. There can also be differences in the amount of random variation in a variable from one pen or treatment to another.

There is also the question of what statistical tests or test procedures are appropriate. Currently in some cases only an overall test of the test substance is run, in some cases each dose level is compared to the control, while in some each treatment is compared to each other treatment. Some of these tests are general tests that could be applied to any set of treatments, while some are tailored specifically to the ordered nature of the treatments in AR experiments (a control and a set of increasing dose levels of the test substance).

Because of this multiplicity of variable types, data complexities and test procedures, the situation facing those who analyse data from AR experiments is a difficult one. A number of different established methods could be employed, and the ongoing increase in computing power means that procedures that were once considered too calculation-intensive are now feasible.

No consensus has emerged as to what methods are most appropriate. Protocols on AR experiments generally do not help to resolve these questions as their focus is on the methodology of the experiment and not on the statistical analysis.

1.5 Organization of This Report

The first section of the report is this introduction. The other sections are:

Section 2. Description and Classification of the Variables

Four different classes of variable are defined, based on the variable type (measurement or count/proportion) and the subject that the variable is taken from (adult birds or eggs/chicks).

Class	Subject	Variable Type
1	Adult Birds	Measurement
2	Adult Birds	Count or Proportion
3	Eggs or Chicks	Measurement
4	Eggs or Chicks	Count or Proportion

Section 3. The Current Statistical Situation

In this section the statistical methods currently employed are evaluated, and areas are identified where improvement appears to be needed.

Section 4. Methods Recommended for Each Variable Class

The objectives of the analysis are defined, and criteria are set out concerning validity and efficiency. The characteristics of each class of variable are discussed, and methods are identified that appear to meet the criteria.

Section 5. Determination of the NOEC

The determination of the NOEC from the results of the tests of each dose level is discussed, and recommendations are made concerning the procedure to employ.

Section 6. Examination of Data Quality

The issue of data quality is discussed and some measures to ensure an acceptable level of quality are suggested.

Appendix A: This contains further information on the statistical methods referred to in sections 3 and 4.

Appendix B: This contains information on supplemental statistical procedures such as transformations, weighting schemes, and tests of homogeneity of variance.

2. DESCRIPTION AND CLASSIFICATION OF VARIABLES

2.1 Experimental Procedure

The following is a brief description of the experimental procedure for AR experiments. Further information is available from the U.S. EPA guidelines 'Avian Reproduction Test' (U.S. EPA Hazard Evaluation Division, 1986, EPA 540/9-86-139).

In AR experiments the effect of the test substance on avian reproduction is assessed under laboratory conditions by adding it to the birds' diet in various concentrations. The treatments are a control and a series of dose levels of the test substance. Each treatment is applied to a specified number of pens, with each pen containing a specified number of male and female birds (e.g. 1 male, 2 females).

The experiments are run in two phases - a pre-egg-laying phase and an egg-laying phase. In the first phase, which is typically 10 weeks in length, the birds are fed the test substance but no reproductive activity occurs as the diurnal light conditions are maintained at a normal winter cycle.

The egg-laying phase, which is typically from 8 to 14 weeks in length, begins when the lighting is changed to a spring cycle and the birds begin reproductive activities. Eggs are laid at a rate of up to one per day per female bird during this phase, and are incubated until hatching. The chicks are placed in an enclosure and raised for 14 days. The adult birds are sacrificed at the end of the experiment.

This section contains a description of the variables commonly analysed in AR experiments, and their classification according to statistical characteristics. These variables comprise a fairly comprehensive list and include all of the main variable types; however it was not feasible to include every possible variable.

2.2 Variables From Adult Birds

2.2.1 Observations Made on Adult Birds

The weight of each bird is typically measured at a number of points during the experiment. Any mortality during the course of the experiment is noted. At the end of the experiment the bird is sacrificed and a gross necropsy carried out. In addition the food consumption is measured for each pen.

2.2.2 Variables Derived

Two classes of variables were defined - those that are measurements or derived from measurements, and those that are counts or proportions.

Class 1. Measurement Variables for Adult Birds

- weight (at various times)
- weight change (over various periods)
- food consumption per bird

Class 2. Counts or Proportions for Adult Birds

- the number or proportion of hens that lay eggs
- mortality during the experiment
- necropsy observations (e.g. the number or proportion that develop a particular condition)

2.3 Variables From Eggs or Chicks

2.3.1 Observations Made on Eggs and Chicks

When the eggs are laid, they are sometimes weighed and are then inspected for cracks. Some eggs are removed from the experiment at this point for measurement of shell thickness. The remaining eggs are then set in an incubator, checked for fertility at 14 days, and checked for a viable embryo at 21 days.

At hatching the chicks may be weighed and are placed in an enclosure. Excess chicks are removed from the experiment if the enclosure can not hold them all. The remaining chicks then grow for 14 days, at which time they are sacrificed and possibly weighed again.

At each stage the number of eggs or chicks surviving that stage is recorded for each pen, and the eggs or chicks that have not survived or are not viable are removed. Figure 1 illustrates the progress of the eggs and chicks through the experiment.

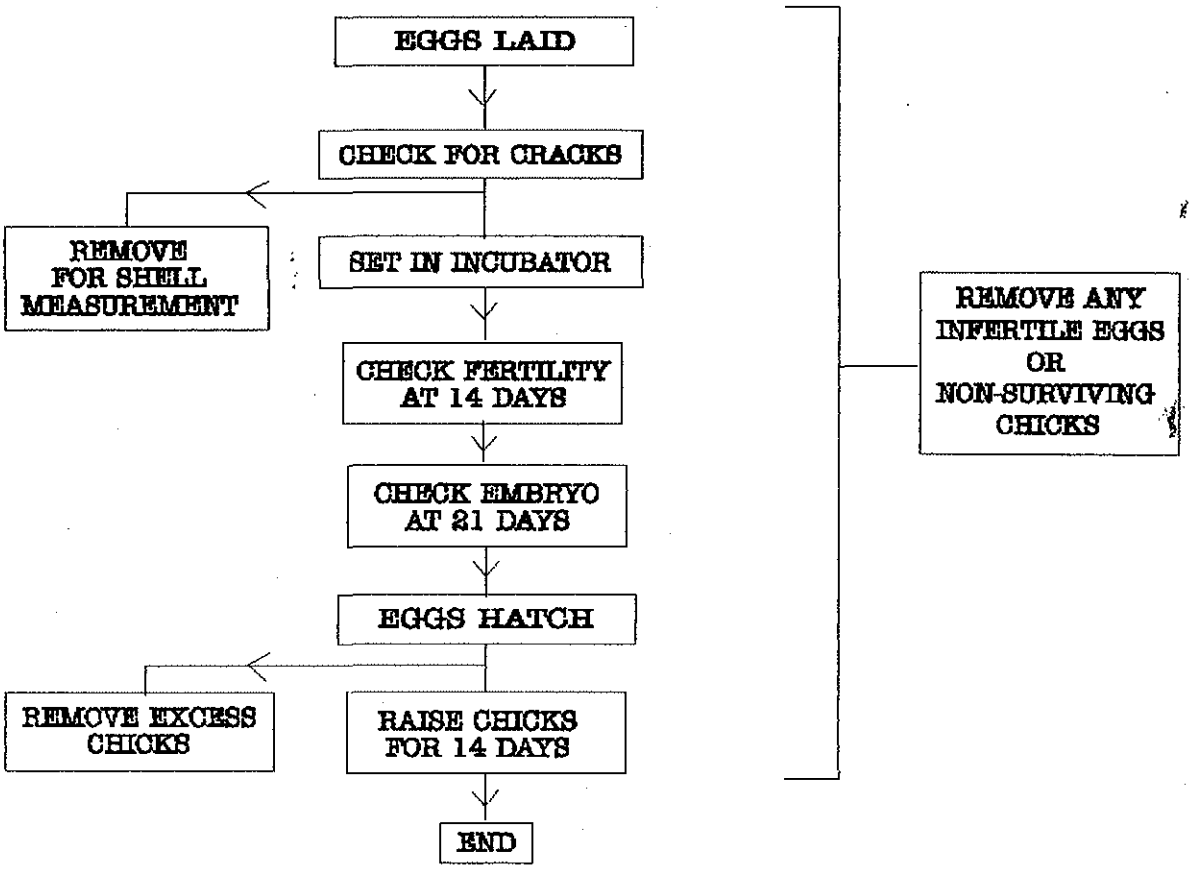

Figure 1. Stages of the Experiment

2.3.2 Variables Derived

Two different variable classes were defined - those that are measurements or derived from measurements, and those that are counts or proportions. Typical variables for these classes are:

Class 3. Measurement Variables for Eggs or Chicks

- egg weight
- egg shell thickness
- chick weight (at hatching and 14 days)
- chick weight gain

Class 4. Counts or Proportions for Eggs or Chicks

Counts

At each stage of the experiment, counts of the surviving eggs or chicks are tabulated for each pen. In addition, 'estimated counts' can be calculated that take into account the removal of eggs for shell measurement or the removal of excess chicks at hatching (which represent the numbers that would have survived if the removals had not occurred). Some examples of counts are:

Actual Counts

- number of eggs laid
- number of non-cracked eggs

Estimated Counts

- estimated number of eggs hatched
- estimated number of chicks alive at 14 days

Illustrations of the calculation of an estimated count are:

Estimated number of eggs hatched

$$= (\text{non-cracked eggs}) \times \frac{\text{eggs hatched}}{\text{eggs set}}$$

Estimated number of chicks alive at 14 days

$$= (\text{non-cracked eggs}) \times \frac{\text{eggs hatched}}{\text{eggs set}} \times \frac{\text{14-day chicks}}{\text{chicks retained}}$$

Proportions

The proportions of eggs or chicks that have survived a particular stage or set of stages are calculated from the counts. Some of these are actual proportions that are ratios of actual counts, and some are estimated proportions that take into account the removal of eggs for shell measurement or the removal of excess chicks at hatching. Some examples of proportions are:

Actual Proportions

- Proportion of eggs laid that are not cracked
- Proportion of eggs set that are fertile after 14 days
- Proportion of eggs set that hatch
- Proportion of chicks retained that survive to 14 days

Estimated Proportions

- Estimated proportion of eggs laid that hatch
- Estimated proportion of eggs laid that produce chicks that survive to 14 days

An illustration of the calculation of an estimated proportion is:

Estimated proportion of eggs laid that produce 14-day chicks

$$= \frac{\text{non-cracked eggs}}{\text{eggs laid}} \times \frac{\text{eggs hatched}}{\text{eggs set}} \times \frac{\text{14-day chicks}}{\text{chicks retained}}$$

Note: The calculations for estimated counts and proportions assume that the eggs removed for measurement of shell thickness were selected at random from the non-cracked eggs. However, it is possible that the eggs removed were designated in advance as suggested in the OECD protocol. In this case the counts and proportions are tabulated for the non-designated eggs only, and there is no need to allow for the removal of eggs for shell measurement since these eggs were not in the data set to start with.

3. CURRENT STATISTICAL TEST METHODS

3.1 Sources of Information on Current Methods

Information on the methods currently employed for testing the effect of the test substance was obtained from these sources:

1. Submissions from approximately 100 AR experiments

2. Existing protocols for AR experiments:

   - the U.S. EPA protocols
     'Avian Reproduction Experiments' (1986), EPA 540/9-86-139, and
     'Avian Reproduction Test' (1982), EPA 540/9-92-024

   - the ASTM protocol 'Standard Practice for Conducting Reproductive Studies With Avian Experiments' (1984)

   - the OECD protocol 'Avian Reproduction Test' (1984)

Although the main concern of these protocols was to set out the experimental procedure, they did contain some information on statistical methods. The ASTM protocol suggests a somewhat more comprehensive approach than the EPA or OECD protocols and also provides a fairly lengthy list of statistical references.

3. Documents on experiments in toxicology and teratology:

   - the OECD guidelines 'One-Generation Reproduction Toxicity Test' (1981) and 'Two-Generation Reproduction Toxicity Test' (1983)

   - the WHO publication 'Principles for Evaluating Health Risks to Progeny Associated with Exposure to Chemicals During Pregnancy' (1984)

These documents are relevant because from a statistical point of view experiments in toxicology and teratology have much in common with AR experiments. The treatments are a control and a set of concentrations of a test substance, and the structure of the data sets is usually similar. The experimental unit is generally a litter of mice or rats rather than a single subject, as the unit is the pen in AR experiments. Many of the variables have similar features, for example some of the key variables are counts or proportions (e.g. the proportion of subjects that develop a particular condition) as they are in AR experiments.

However, little information was obtained from these documents. The OECD guidelines simply state that appropriate statistical methods should be employed, while the WHO paper contains a general discussion of statistical analysis but does not specify any particular methods.

3.2 Background Information on Current Methods

Experimental Design

From the suggestions made with respect to data analysis, it appears that the EPA and ASTM protocols assume an experimental design with a simple random allocation of pens to treatments. Most of the submissions employed this design, although some employed a more complex design such as a blocked design in order to minimize the variation in temperature or humidity from one treatment to another. (The OECD protocol does not appear to make any specific assumptions concerning experimental design.)

Test Type and Confidence Level

The test type and confidence level to be employed are not specified in the protocols, although it is fairly safe to assume that statistical tests are to be two-tailed tests run at the 5% confidence level since this is generally the norm. In all of the submissions it appears that this was the type and level of test that was carried out.

General Approach to Statistical Analysis

The EPA and ASTM protocols distinguish between methods for measurement variables and methods for counts or proportions. They suggest an ANOVA-based approach for measurement variables, and either a chi-square-based or ANOVA-based approach for counts and proportions. The OECD protocol is less specific and makes only the general suggestion that an ANOVA or other acceptable method be employed.

In general the submissions followed the suggestions of the EPA and ASTM protocols, but there was a great deal of variation from one submission to another with respect to the complexity of the methods employed and the amount of information provided on these methods. Some provided a considerable amount of information and some provided almost no information.

Note: In order to preserve confidentiality, this report does not give precise information on the methods employed by any submission.

3.3 Current Methods for Measurement Variables

3.3.1 Methods for Measurement Variables Suggested in Protocols

The statistical approach suggested in the EPA protocols for measurement variables is to run an ANOVA, and to follow this with pairwise comparisons of treatment means if the ANOVA finds a significant effect. The test suggested for these comparisons is Duncan's test. The model for the ANOVA is not stated, although it appears that the one-way model is assumed.

The approach suggested in the ASTM protocol is basically the same, but wider in scope. For the first analysis it suggests either an ANOVA or a general linear models analysis, and for the pairwise comparisons Dunnett's test and the LSD test are suggested as well as Duncan's test. Again it appears that a one-way model is assumed. The ASTM protocol mentions the possibility of taking into account unequal sample sizes in the data, but does not suggest a procedure for this. It also suggests that if significant treatment effects are found in the pairwise comparisons, the trend in effect as the dose level increases should be tested using regression.

3.3.2 Methods for Measurement Variables Employed in Submissions

Measurement variables were analysed using an ANOVA in all of the submissions where the method was specified, as suggested by the protocols. The ANOVA model appeared to be the simple one-way model, except for those cases where a more complex design such as a blocked design had been employed.

If the ANOVA produced significant results, most submissions carried out pairwise comparisons of all treatment means. The most common tests were those suggested in the protocols, but some others were also employed. In some cases the submissions carried out tests that involved an ordering of the treatments, such as a test of the trend in the treatment means as the dose level increased or a comparison of control with the highest dose level.

Since a one-way ANOVA requires a 2-level data set while most AR data sets are 3-level (subjects within pens within treatments), a key question is that of how the data was reduced from 3 levels to 2 prior to the analysis. In most cases it appears that the data values from all pens within a treatment were pooled together. This in effect treats the data from a treatment as if it had come from a single pen, and reduces the data set to a 2-level structure of subjects within treatments. The validity of this practice is discussed in section 3.6.

Another way to reduce the data structure to 2 levels, that appears to have been employed in some cases, is to form a data set of the pen means and then to run an ANOVA or other analysis on these means. However there is then the complication that some pen means are more accurate than others, due to inherent differences in pen-to-pen variation or to differences in the number of subjects per pen.

Submissions differed in the extent to which the variation in accuracy of the pen means was tested for and taken into account. Some went to considerable lengths to deal with it, while others appeared to completely ignore it. A procedure followed in some cases was to employ a test for inhomogeneity of variance to compare the variance between pen means in one treatment to the variance in another. Bartlett's test was generally the test employed for this purpose.

If the differences among variances were significant, it appears that the most common approach was to attempt to equalize the variance by transforming the data (usually by a log transformation). The test for variance inhomogeneity was then repeated. If homogeneity was not achieved, in most cases the variance of the pen means was estimated separately within each treatment. Treatment effects were then tested by pairwise comparisons of treatment means using unequal-variance t-tests. Another method employed for problem cases was to compare treatments pairwise using a non-parametric test such as the Mann-Whitney test.

3.3.3 Evaluation of Current Methods for Measurement Variables

Clarity of Objectives

It is not clear exactly what results are to be obtained from the analysis. Since this is not stated, it must be inferred from the methods employed. The fact that an ANOVA is the recommended first step, and also the last step if the ANOVA test is not significant, implies that the only result that is needed is an overall result obtained over all dose levels. On the other hand the use of pairwise comparisons for all pairs of treatments suggests that the relative rank should be determined for each dose level.

It is not clear what objective is behind the suggestion in the ASTM protocol to run a regression of the treatment mean on the dose level if treatment effects are found. This may be part of the procedure for testing the treatment effects, or may imply that a secondary objective (which is optional and at the discretion of the experimenter) is to model the dose-response relationship.

Validity of Current Methods

The biggest problem appears to be the practice of pooling the data from all pens within a treatment, thus reducing the data structure to 2 levels from 3. This usually results in an analysis that is not statistically valid for reasons set out in section 3.6. It appears that many of the submissions employed this approach, although it is not possible to be sure of this because the information provided on the methodology was often very incomplete.

The alternative procedure for producing a 2-level data set, which is to calculate the mean for each pen and then run an analysis on the pen means, is statistically much more appropriate but requires that the variation in accuracy of the pen means be taken into account. Many submissions appeared to ignore this problem, since they used methods such as a one-way ANOVA or a multiple comparison test which assume that each data value has equal accuracy. For most AR data sets these methods would not be valid since this assumption is seriously violated.

Some of the submissions dealt with the variation in accuracy of the pen means by estimating the within-treatment variance separately for each treatment if a test for inhomogeneity of variance was significant. While this approach is an improvement over the practice of ignoring unequal variances, it also has problems. One is that the number of pens per treatment is often small, and the variances of the pen means would have only a few degrees of freedom.

But the main problem is that this approach is not flexible enough. Its basic assumption is that all the pen means within a treatment have the same accuracy, and the only question is whether this accuracy is constant overall or varies from one treatment to another. It dose not distinguish between variation in accuracy caused by differences in the number of subjects per pen, and variation in accuracy caused by differences in the underlying pen-to-pen variation. For this reason it is not flexible enough to deal separately with each pen mean according to how many subjects were in that pen and to the underlying pen-to-pen variation in that treatment.

Efficiency of Current Methods

Statistical efficiency is another area where improvement could be made. Most submissions started with an ANOVA, and concluded that there was no treatment effect if the result was not significant. But ANOVA is a general method that tests for any differences between treatments and is not efficient in cases where a specific pattern of effect is expected (as in AR experiments where we expect an increasing effect as the level of the test substance increases).

Similarly most of the tests employed for pairwise comparisons of treatment means are relatively inefficient for AR experiments. They are too conservative as they are intended for situations where the probability of error in any one of the tests is to be 5%. For Duncan's test the error rate is set at 5% over all pairs of treatments, while in Dunnett's test it is 5% over all comparisons of control with a non-control treatment. The LSD test, t-test or Mann-Whitney test employed in some submissions are better choices for pairwise comparisons as the error rate is 5% for each test.

Only a few submissions employed a method that involved the most efficient type of test for AR experiments - a test that tests specifically for an increasing effect as the dose level increases, such as a trend test or a test of control against the highest dose level. And for those cases where such a test was employed, it is not clear if it was run on its own (as it should be) or was run only if an ANOVA had been carried out first and had produced a significant result.

Finally, it appears that all submissions employed two-tailed statistical tests (they did not state whether their tests were one-tailed or two-tailed, but two-tailed tests are much more common in statistics generally). However one-tailed tests would be much more efficient for AR experiments since we are looking specifically for negative effects on reproduction.

3.4 Current Methods for Counts and Proportions

3.4.1 Methods for Counts and Proportions Suggested in Protocols

For counts and proportions, the EPA protocol suggests either a chi-square test or an ANOVA. If an ANOVA is employed, the arcsine transformation is recommended for proportions prior to analysis. If the ANOVA finds a significant effect, pairwise comparisons of the treatment means are recommended. Duncan's test is mentioned as a possible test for these comparisons. The model for the ANOVA is not stated, although it appears that the one-way model is assumed. No suggestion is made as to what to do if a chi-square test is run and finds a significant effect.

The methods suggested in the ASTM protocol are similar to those in the EPA protocols but somewhat wider in scope. For the first test either a chi-square analysis or a least squares analysis is suggested, with the least squares analysis being either an ANOVA or a general linear model analysis. If least squares analysis is employed, it mentions the possible use of weights to take into account unequal sample sizes in the data. Methods for deriving weights are not given. As with the EPA protocols, it appears that a one-way model is assumed for least squares analysis.

If a least squares analysis is run and a significant result obtained, the ASTM protocol suggests pairwise comparisons of treatment means using tests such as Duncan's test, Dunnett's test or the LSD test. If a chi-square test is run and a significant result obtained, it suggests that the treatment effect be examined in more detail by partitioning the chi-square statistic. No specific partitions are suggested.

The ASTM protocol also suggests that if significant treatment effects are found, the trend in effect as the dose level increases should be examined. If a least squares analysis was run, the analysis suggested for trend is a linear regression of treatment mean against dose level. If a chi-square test was run the analysis suggested for trend is Armitage's test for a linear trend in proportions (referred to in this report as the Cochran-Armitage test).

3.4.2 Methods for Counts and Proportions Employed in Submissions

Most submissions employed either an ANOVA or a chi-square approach, as suggested in the protocols. The ones that employed a chi-square approach generally stated only that fact and did not give any further details on the methods employed. Those that employed ANOVA usually provided additional information on their methods.

For the submissions that employed a chi-square approach, a key question is that of how the data was treated in order to allow a chi-square test to be run. A chi-square test for proportions requires that the data be in the form of a contingency table. This means that the data must be reduced from its original structure (with a proportion for each pen) to a simplified structure with one proportion per treatment. It appears that this simplification was usually carried out by pooling the data from all pens within a treatment (this practice is discussed in section 3.6).

The submissions that employed the ANOVA approach usually applied a transformation to proportional data prior to the analysis to stabilize the variance. This was generally an angular transformation although some others were also used.

If the ANOVA produced significant results, most submissions then carried out pairwise comparisons of treatment means. The most common tests were those suggested in the protocols, but some others were also employed. In some cases tests were run that involved an ordering of the treatments, such as a trend test or a comparison of control with the highest dose level.

For proportional variables, some of the submissions that followed the ANOVA approach also employed measures to allow for variation in the accuracy of pen counts or proportions due to variation in the number of eggs or chicks per pen. Often the measure was to employ weighted data analysis, assigning larger weights to pens with a larger number of subjects.

The most common procedure to derive a weighting scheme was Cochran's method (described in section B.2.2.1 in Appendix B), which involves establishing the relationship between the number of subjects per pen and the accuracy of the pen proportions. In other cases a much simpler plan was adopted and the weights were simply set equal to the number of subjects per pen. It is not clear if weights were employed in the pairwise comparisons in addition to the ANOVA.

Some submissions appeared to follow a different approach to the problem of unequal variances among the pen proportions, and ran a test of homogeneity of variance to compare the variance within each treatment prior to carrying out the pairwise comparisons. (It is not clear if they did this for measurement variables only, or for proportions also.)

If the inhomogeneity was significant, the procedure usually followed was to estimate the variance of the pen counts or proportions separately for each treatment and then to carry out pairwise comparisons of treatment means using unequal-variance t-tests. An alternative procedure followed in some cases was to carry out the pairwise treatment comparisons using a non-parametric test (the Mann-Whitney test).

3.4.3 Evaluation of Current Methods for Counts and Proportions

Clarity of Objectives

As was the case with measurement variables, it is not clear exactly what results are required. Since this is not stated, it must be inferred from the methods employed. The fact that a general test such as a chi-square test or an ANOVA is the recommended first step, and also the last step if the test is not significant, implies that the only result needed is an overall result for all dose levels. On the other hand the use of pairwise comparisons for all pairs of treatments suggests that the relative rank should be determined for each level.

In addition, the suggestion in the ASTM protocol to examine the trend in treatment effect as the dose level increases (if a treatment effect was found) may imply that a secondary objective is to model the dose-response relationship.

Validity of Current Methods

The biggest problem appears to be that an appreciable number of submissions pooled the proportions from all pens within a treatment, in order to simplify the data structure so that a chi-square test could be run. This procedure is not statistically valid, for reasons set out in section 3.6.

For the submissions that ran an ANOVA, the most difficult aspect was to deal with the unequal numbers of subjects per pen and the consequent variation in accuracy of the pen proportions. Some submissions ignored this problem, which is not advisable. Some used weighted analysis, with Cochran's approach used to determine the weights. This is statistically the most valid but also the most difficult. Some submissions used weighted analysis with weights equal to the numbers per pen; however it is not clear if this is any improvement on an unweighted analysis.

Unequal numbers of subjects per pen also caused complications for the pairwise comparisons of treatments. Some submissions again followed the convenient practice of ignoring this problem. Others allowed for it by estimating the variance between pen proportions within treatments separately for each treatment, if a test for inhomogeneity of variance was significant.

While the use of a homogeneity of variance test is an improvement over the practice of ignoring unequal variances, it is not completely satisfactory. It allows the accuracy of pen proportions to vary from one treatment to another, but still makes the questionable assumption that the accuracy is constant within a treatment.

Consequently this approach is not flexible enough to take into account the accuracy of each individual pen proportion, which it should be with AR experiments. Another consideration is that a test for inhomogeneity would not be that powerful in any case, since the numbers of pens per treatment is generally small in AR data sets and the variances of the pen proportions would have only a few degrees of freedom.

Efficiency of Current Methods

Another area of concern is that of the efficiency of the statistical tests. Most submissions started with a general test - a chi-square or an ANOVA - and concluded that no treatment effect was present if the result was not significant. But for situations where a specific pattern of effect is expected, as in AR tests where we expect an increasing effect as the level of the test substance increases, a test for that specific pattern would be much more efficient.

Similarly most of the tests employed for pairwise comparisons of treatment means are too general in nature and thus are not that efficient for AR experiments. Duncan's test is intended for cases where all pairs of treatments are to be tested, while Dunnett's test is intended for cases where the control is to be compared with each non-control treatment. The LSD test, t-test or Mann-Whitney test employed in some submissions are better choices for pairwise comparisons.

The most efficient methods for AR experiments were carried out in only a few submissions. These involved tests designed specifically to detect an increasing effect of the test substance as the dose level increases, such as a trend test or a test of control against the highest dose level. It is not known whether the method for these submissions consisted of this test by itself, or involved running an ANOVA first and then running the test only if the ANOVA was significant. If it was the latter, the advantage of using an efficient test would be lost.

Current methods for counts and proportions can also be made more efficient by employing one-tailed tests instead of the more usual two-tailed tests, in order to test specifically for negative effects on reproduction. The submissions did not state whether they employed one-tailed or two-tailed tests, but they presumably used two-tailed tests since these are much more common in statistics generally.

3.5 Summary of Assessment of Current Methods

There is considerable variation with respect to validity and efficiency among the statistical methods currently employed, and a definite need for a more effective and consistent set of methods. While methods were reasonably well chosen and well described in some submissions, in others they were deficient in many respects. And in some the methods were not described beyond a very brief reference. Improvement is needed in the areas of

- defining objectives
- choosing statistical methods that are efficient at meeting these objectives
- choosing methods that are statistically valid, and that can deal with the complexities of AR data sets
- providing an informative description of the methods employed

In fairness to the submissions it should be recognized that the analysis of AR data sets in a statistically valid and efficient manner is not a simple task, because of the difficulties and complexities that are often present. The analyst is frequently faced with the need to choose between statistical validity on one hand and feasibility and computational convenience on the other. Another factor is the lack of clearly stated objectives and guidelines in the protocols. In addition, there do not appear to be any statistical papers or texts on the subject of appropriate methods for AR experiments.

3.6 Discussion of Pooling of Data

3.6.1 Introduction

A key question both for measurement variables and for counts or proportions is that of whether it is permissible to pool data values from all pens within a treatment, and then ignore the pen structure in the analysis and treat the data as if the treatments had been applied to individual subjects. This pooling reduces the data structure from its actual 3-level structure of subjects within pens within treatments to a 2-level structure of subjects within treatments.

To illustrate this, consider the following simplified example for a measurement variable. Pooling the data for the following 3-level data set

Treatment 1	Treatment 2
Pen 1: 1.5, 1.7, 1.4	Pen 1: 0.6, 0.9
Pen 2: 1.3, 1.2	Pen 2: 0.9, 1.1, 1.4
Pen 3: 2.0, 2.2, 2.5	Pen 3: 0.5, 0.8, 0.7

would produce the 2-level set

Treatment 1	Treatment 2
1.5, 1.7, 1.4, 1.3, 1.2, 2.0, 2.2, 2.5	0.6, 0.9, 0.9, 1.1, 1.4, 0.5, 0.8, 0.7

Similarly the following data set for a proportion

Treatment 1	Treatment 2
Pen 1: 25/30	Pen 1: 10/20
Pen 2: 15/25	Pen 2: 5/15
Pen 3: 15/20	Pen 3: 10/15

would be reduced by pooling to

Treatment 1	Treatment 2
55/75	25/50

The data sets produced by pooling are obviously much easier to analyse than the original sets. The pooled data set for the measurement variables could be analysed by simple least squares methods, and the pooled set for proportions could be analysed by contingency-table methods. Unfortunately, for reasons set out in 3.6.3, this practice is not statistically valid for either variable type.

3.6.2 Position of Protocols on Pooling of Data

The OECD protocol does not consider this question, while the EPA and ASTM protocols take a position that is somewhat inconsistent. They do not state directly that it is permissible to pool data from all pens within a treatment and then ignore the pen structure in the subsequent data analysis. But the fact that they suggest an ANOVA approach with multiple comparison of treatment means for measurement data suggests that they assume that the data structure has first been reduced to 2 levels, and then analysed as a simple 2-level data set. The obvious procedure to achieve this reduction is by pooling.

For proportional variables, the EPA and ASTM protocols accept the chi-square test as a valid method of analysis. Since this test requires the data to be in the form of a simple contingency table, this implies that the data have been reduced to a single proportion per treatment. Presumably this was achieved by pooling the pen proportions within each treatment. Also, presumably the pen structure was ignored in the contingency table analysis.

The WHO document on teratological experiments also deals with this issue. In these experiments, treatments are applied to entire litters of mice or rats. The question discussed is whether the analysis should reflect the litter-based structure of the data, or whether the data for all litters in a treatment should be pooled and the litter-based structure ignored in the analysis. The paper attempts to compromise by suggesting that data be analysed twice - once taking the litter structure into account and once with it ignored. Presumably a treatment effect would be considered to be present if either analysis produced a significant result, although this was not stated.

3.6.3 Validity of Pooling

This question of pooling has been discussed in a number of papers in scientific journals, mainly with respect to toxicological experiments where the experimental unit is the litter of animals (generally mice or rats). The strong consensus is that the practice of pooling the data for all litters within a treatment, and then analysing the data as if the treatments had been applied to individual animals rather than on a litter basis, is not valid (e.g. Weil (1970), and Haseman and Soares (1976)). The reason is that subjects from the same litter will tend to be have similar responses, resulting in a cluster of similar data values.

When data values within a treatment are averaged to obtain the treatment mean, the accuracy of this mean depends on the extent to which the random errors of the individual values cancel each other out. If the data values occur in clusters, there will be less cancellation of error than if the values were independent since there is more chance that many errors will occur in the same direction.

As a result of this reduced error cancellation, the variance of the treatment means will be underestimated if the data from all pens or litters in a treatment are pooled and then analysed as if they were independent observations. The consequence of this underestimation of the variance is to increase the probability of finding the treatment effect to be significant.

Paradoxically this can be used as an argument in favour of pooling, since it increases the chance of detecting a harmful effect of the test substance. But any benefits from this are offset by the fact that the decrease in the variance is very inconsistent from one case to another since it depends on the number of subjects per pen, the size of the pen-to-pen variation and other factors. Also, in principle it should not be necessary to employ invalid statistical methods in order to detect treatment effects.

4. IDENTIFICATION OF PROMISING STATISTICAL TEST METHODS

4.1 Introduction

A large number of statistical methods that are currently being applied or could potentially be applied to AR data sets were evaluated, and of these a number were recommended for each variable class. The criteria employed to evaluate the methods are set out in section 4.2, and the methods themselves are described in sections 4.3 to 4.6. A summary is presented in Table 1.

Note: The statistical methods presented in this section assume that the experiment followed a 'one-way' design in which pens were assigned to treatments by simple random allocation. The methods also assume that the same number of pens was assigned to each treatment, except for the possibility that extra pens could have been assigned to the control.

Further Information

The statistical methods presented in this section are not described in detail; however, further information on them and on other aspects of the statistical analysis is given in the Appendices.

Appendix A

This contains information on the statistical methods, including reviews of any recent developments.

Appendix B

This contains information on supplemental statistical procedures:

- data transformations
- weighting procedures
- testing homogeneity of variance
- combining pen proportions

Table 1. Recommended Methods by Variable Class and Set of Assumptions

Possible Assumptions for Each Variable Class

Note: X_{ij} denotes a pen mean, P_{ij} a pen proportion, N_i the number of pens per treatment, σ_{Xij}^2 the variance of X_{ij}, and σ_{Pij}^2 the variance of P_{ij}

Statistical Assumptions	Measurement Adult Birds	Prop/Count Adult Birds	Measurement Eggs/Chicks	Prop/Count Eggs/Chicks
1. σ_{Xij}^2 or σ_{Pij}^2 are const., N_i are all equal	Yes	---	Yes	Yes
2. σ_{Xij}^2 or σ_{Pij}^2 are const., N_i not all equal	Yes	---	Yes	Yes
3. σ_{Xij}^2 or σ_{Pij}^2 are const. within treatments but vary between treatments	Yes	---	Yes	Yes
4. σ_{Xij}^2 or σ_{Pij}^2 vary within treatments	Yes	---	Yes	Yes
5. P_{ij} can be reduced to a single proportion within each treatment	---	Yes	---	Yes
6. X_{ij} or P_{ij} have an irregular distribution	---	Yes	---	Yes

Methods Recommended for Each Set of Assumptions

Set 1	Set 2	Set 3
t-test	t-test	t-test
LSD test	LSD test	trend test
Williams test	Williams test*	
trend test	trend test	
Abelson-Tukey test		

Set 4	Set 5	Set 6
weighted t-test	chi-square test	Mann-Whitney test
weighted trend test	Fisher exact test	rerandomization test
	Cochran-Armitage test	jackknife method

* Williams test is applicable here only to the case of extra control pens.

4.2 Criteria for Evaluating Statistical Methods

4.2.1 Introduction

Each method was evaluated by examining the following aspects:

- statistical validity, in particular
  - whether it reflects the inherent structure of the data
  - whether it deals with the data complexities that commonly occur
- efficiency at detecting treatment effects
- whether the effect is tested at each dose level

These topics are discussed in sections 4.2.2 to 4.2.4, and a set of criteria for evaluating methods are set out in section 4.2.5.

4.2.2 Validity of the Methods

Since treatments are applied to pens rather than to individual subjects in AR experiments, the statistical methods employed should use the pen as the basic experimental unit. The methods should reflect the fact that the basic source of random error in the experiment is the variation between pens within a treatment.

Since the pen is the basic unit in the experiment, the starting point for any method should be the calculation of pen means, counts or proportions. Treatment means or proportions should then be calculated from these pen quantities, and their accuracy should be derived from the variation between the pen quantities within each treatment. To pool the data from all pens within a treatment, and then ignore the pen structure in the subsequent analysis, is not acceptable as the pen-to-pen variation is lost.

The method should also be flexible enough to able to deal with the data complexities that can be expected to occur. These include

- variation in the number of subjects per pen
- differences in the inherent pen-to-pen variation
- irregularities in the data

One potential source of irregularity is that of reproductive failure in particular pens for reasons unrelated to the treatments applied. This is mentioned in the literature as a feature of AR data sets for certain variables (Picirillo and Quesenberry, 1980).

4.2.3 Efficiency of the Methods

Different statistical methods test for different kinds or patterns of treatment effects. The efficiency of a method for a given experiment depends on how well matched it is to the patterns of treatment effects that occur with that type of experiment. The most efficient methods for AR experiments are those that test for one specific pattern, which has two important features:

- It is negative in direction with respect to reproductive capability
- It increases in magnitude as the dose level increases

To test for an effect that increases as the dose level increases, the tests should make use of the ordered nature of the treatments. To test specifically for negative effects, the tests should be one-tailed tests rather than the two-tailed type that are more commonly employed in statistical analysis.

4.2.4 Testing at Each Dose Level

A decision was made that the statistical method should include a test of the effect of the test substance at each dose level. The reason is that it is considered important to be able to identify the NOEC (the highest dose at which there is no observed effect), and this requires that a significant or non-significant result be obtained for each dose. The process for identifying the NOEC is described in section 5.

For the methods recommended in this section, the test of the effect at a given dose level is carried out in one of two ways:

- by testing the difference between that dose level and control, or
- by testing the trend over the set of treatments from control up to that dose level

Note: The test of trend is carried out solely for the purpose of determining whether the effect is significant at that particular dose level. It should not be confused with procedures for modelling the dose-response relationship. (Modelling the dose-response curve is outside the scope of this report.)

4.2.5 Summary of Criteria Employed

From the discussions in the preceding sections, the following criteria have been drawn up for evaluating statistical methods:

1. Methods should be reasonably well established in statistics.

2. Methods should employ the pen as the basis of the analysis. The quantities analysed should be pen means, pen counts or pen proportions. It is not acceptable to pool the data from all pens within a treatment, and then ignore the pen-based structure of the data set in the subsequent analysis.

3. Methods should take into account those complexities that commonly occur in data from AR experiments, particularly variation in the numbers of subjects per pen and variation in the accuracy of pen means, counts or proportions.

4. Methods should test the effect of the test substance at each dose level.

5. Each of the tests should be a one-tailed test at the 5% confidence level, and should test for a negative effect on reproduction that increases as the dose level increases.

4.3 Methods for Measurement Variables - Adult Birds

4.3.1 Requirements to be Met By These Methods

The criteria are those set out in section 4.2.5, with the following condition added:

Employ least-squares methods only. Non-parametric methods should not be necessary for this variable class and are not usually applied to it.

4.3.2 Basic Variable Characteristics and Model

For most of the variables in this class the data values are measurements made on individual adult birds, and the data sets have a three-level structure of birds within pens within treatments. (An exception is food consumption, where there is one measurement per pen per time period.)

In general the number of birds, and thus of measurements, will be the same for all pens. The initial number per pen is the same, and any variation would be the result of mortality. The initial number is either 1, 2, 3, 5 or 7, depending on the species and caging parameters and whether the variable is measured on males only, females only or on both sexes.

For most variables the data should follow approximately the standard linear model:

$$X_{ijk} = \mu + T_i + E_{ij} + e_{ijk}$$

where X_{ijk} is the data value for the k'th bird in the j'th pen in treatment i, μ is the overall mean, T_i is the effect for treatment i, E_{ij} is the random pen effect and e_{ijk} is the random error for an individual data value. In the standard model the E_{ij} and e_{ijk} have approximately normal distributions and their variances σ_{Eij}^2 and σ_{eijk}^2 are roughly constant over the data set.

Pen Means: The methods considered are all based on least squares analysis of pen means. The pen means $\bar{X}_{ij}$ have the form

$$\bar{X}_{ij} = \mu + T_i + E_{ij} + \bar{e}_{ij}$$

where $\bar{e}_{ij}$ is the mean of the e_{ijk} for that pen. The variance σ_{Xij}^2 of $\bar{X}_{ij}$ is

$$\sigma_{Xij}^2 = \sigma_{Eij}^2 + \sigma_{eij}^2/n_{ij}$$

where n_{ij} is the number of birds in the pen and σ_{eij}^2 is the mean of the σ_{eijk}^2.

Treatment Means: If σ_{Xij}^2 is constant (or approximately so) for all pens within a treatment, the treatment mean is calculated as the unweighted mean $\bar{X}_i$ of the pen means and its variance σ_{Xi}^2 has the form

$$\sigma_{Xi}^2 = \sigma_{Xij}^2/N_i = (\sigma_{Eij}^2 + \sigma_{eij}^2/n_{ij})/N_i$$

where N_i is the number of pens in the treatment.

4.3.3 Common Data Complexities

1. Variation among the n_{ij}. The initial number of birds per pen is the same for all pens. But if deaths have occurred in some pens during the experiment, the n_{ij} will vary from pen to pen for those variables measured after the deaths.

2. Skewness in the distribution of the X_{ijk}. This may produce non-linearity in the model and cause σ_{Xij}^2 to vary with the size of $\bar{X}_{ij}$. In general these problems can be handled by applying a logarithmic transformation. For some variables the use of a log transformation is a fairly standard practice (e.g. pesticide concentrations).

3. Variation among the number N_i of pens per treatment. Differences will occur if extra pens were assigned to the control, or if certain pens had to be left out of the analysis because of problems such as sickness or mortality.

4. Variation in σ_{Xij}^2 (that is not removable by transformation). Possible causes for this include variation among the n_{ij}, and non-removable variation in σ_{Eij}^2 or σ_{eij}^2. The variation among σ_{Eij}^2 or σ_{eij}^2 could be such that they are different for each pen, or they could be constant for all pens within a treatment but vary from one treatment to another.

4.3.4 Cases to be Considered

The methods that are appropriate for a given data set depend on σ_{Xij}^2 and N_i. Four different cases have been identified. The choice of case affects not only the testing of treatment effects but the actual calculation of the treatment means.

Note: It is assumed that any variation in σ_{Xij}^2 that is due to skewness alone has been removed by transformation.

Case 1: σ_{Xij}^2 Constant Over All Treatments, N_i Equal

Since σ_{Xij}^2 is equal to $(\sigma_{Eij}^2 + \sigma_{eij}^2/n_{ij})$, for σ_{Xij}^2 to be constant the parameters n_{ij}, σ_{Eij}^2 and σ_{eij}^2 must all be constant. For n_{ij} to be constant over all treatments, there must not have been any deaths among the adult birds up to this point. Each treatment mean $\bar{X}_i$ is calculated as an unweighted average of the pen means $\bar{X}_{ij}$, and the variance σ_{Xi}^2 of the treatment means is constant and equal to σ_{Xij}^2/N where N is the common value of the N_i.

Case 2: σ_{Xij}^2 Constant Over All Treatments, N_i Not Equal

The pen parameters n_{ij}, σ_{Eij}^2 and σ_{eij}^2 must be constant, but N_i can vary. The treatment means $\bar{X}_i$ are calculated as unweighted averages of the pen means $\bar{X}_{ij}$. The variance σ_{Xi}^2 of each treatment mean is equal to σ_{Xij}^2/N_i, and varies inversely with N_i.

Case 3: σ_{Xij}^2 Constant Within Each Treatment, But Varies Between Treatments

Here the parameters n_{ij}, σ_{Eij}^2 and σ_{eij}^2 must be constant within a treatment but can vary from one treatment to another. Treatment means $\bar{X}_i$ are calculated as unweighted averages of the pen means $\bar{X}_{ij}$. The variance σ_{Xi}^2 of each treatment mean is equal to σ_{Xij}^2/N_i, and varies with both σ_{Xij}^2 and N_i.

Case 4: σ_{Xij}^2 Varies Between Pens Within a Treatment

If σ_{Xij}^2 varies from pen to pen within a treatment (due to variation in n_{ij}, σ_{Eij}^2 or σ_{eij}^2), the situation is more complicated. To accommodate this variation, the statistical method must involve weighted least squares. In order to set up a suitable weighting scheme, σ_{Xij}^2 must be modelled as a function of n_{ij} or other variables and then estimated separately for each pen. The weight assigned to each pen mean $\bar{X}_{ij}$ is the inverse of this estimate of σ_{Xij}^2.

The treatment means are calculated as weighted means of the pen means, and treatment effects are tested by comparisons among these weighted treatment means. Methods for deriving an appropriate weighting scheme are not described here, but are discussed in Appendix B. It is assumed that weighting is necessary only at the pen level, and not at the level of individual measurements.

4.3.5 Recommended Methods for the Different Cases

No guidelines have been set out as how to determine which case to select for a given data set, as this is largely a matter of subjective judgement.

Note - Each method involves a set of one-tailed tests at the 5% level, that test the effect of the test substance at each dose level. For more information on the methods see Appendix A.

Case 1. σ_{Xij}^2 Constant, N_i Equal

These methods involve standard least-squares analysis.

- t-tests, each test compares the control mean with the mean for a particular dose level
- LSD (least significant difference) tests, each test compares the control mean with the mean for a particular dose level
- Williams tests, each test compares the control mean with the mean for a particular dose level
- Linear trend tests, each test looks at the trend in treatment means from control up to a particular dose level
- Abelson-Tukey tests, each test looks at the trend in treatment means from control up to a particular dose level

Case 2. σ_{Xij}^2 Constant, N_i Not All Equal

These methods involve least-squares analysis, adjusted for the variation in N_i.

- t-tests, each test compares the control mean with the mean for a particular dose level

- LSD tests, each test compares the control mean with the mean for a particular dose level

- Linear trend tests, each test looks at the trend in treatment means from control up to a particular dose level

- Williams tests, adjusted for extra control pens, each test compares the control mean with the mean for a particular dose level. Note that this test can not be applied if N_i varies from one dose level to another. It can be applied to the case where extra pens were assigned to the control, but the number of pens in the other treatments is constant.

Case 3: σ_{Xij}^2 Varies Between Treatments

These methods involve least-squares analysis, adjusted for the variation in σ_{Xij}^2.

- t-tests, adjusted for variation in σ_{Xij}^2 between treatments, each test compares the control mean with the mean for a particular dose level

- Linear trend tests, adjusted for variation in σ_{Xij}^2 between treatments, each test looks at the trend in treatment means from control up to a particular dose level

Case 4: σ_{Xij}^2 Varies Between Pens Within a Treatment

These are weighted least-squares methods, with weights applied to the pen means.

- t-tests, adjusted for weighted analysis, each test compares the weighted control mean with the weighted mean for a particular dose level

- Linear trend tests, adjusted for weighted analysis, each test looks at the trend in weighted treatment means from control up to a particular dose level

4.4 Methods for Counts or Proportions - Adult Birds

4.4.1 Requirements to be Met By These Methods

The criteria are those set out in section 4.2.5.

4.4.2 Conversion of Counts to Proportions

It is assumed that counts will be converted into proportions before analysis, by dividing by the initial number n_0 of adult birds in the pen. This will take into account the fact that the counts are restricted to the range of 0 to n_0. After conversion they will be restricted to the range of 0 to 1.

4.4.3 Basic Variable Characteristics and Model

For these variables the data set has a two-level structure of pens within treatments, and each data value is the proportion within a particular pen. Let P_{ij} be the proportion for pen j in treatment i. Then P_{ij} is calculated from

$$P_{ij} = y_{ij}/n_{ij}$$

where n_{ij} is the number of adult birds in the pen for that variable and y_{ij} is the number of the n_{ij} for which a particular characteristic was recorded (e.g. the number that died in a particular phase of the experiment).

4.4.4 Common Data Complexities

1. Variation among the n_{ij}. Although the initial number of adult birds per pen is the same for all pens, n_{ij} will vary between pens if deaths have occurred prior to the tabulation of that variable.

2. The discrete nature of the proportions. In general the P_{ij} will have a distribution that is approximately binomial. Since the denominators n_{ij} are small (either 1, 2, 3, 5 or 7 depending on caging parameters and on whether the proportion is calculated for males, females, or both sexes), this distribution will be discrete to the point where it is not reasonable to treat the P_{ij} as if they were continuous variables.

3. Skewed distribution for the proportions. The distribution of the P_{ij} may be quite skewed in some cases. For example consider the proportion of deaths per pen for an experiment where all the n_{ij} are equal to 7. A possible result would be 0 deaths in most pens, 1 death in a few pens, and a very few pens with 2 or more deaths. The corresponding pen proportions will be skewed with most values concentrated at 0, a few values at .14 and a very few at .28 or more.

4. Variation among the N_i. Differences in the number N_i of pens per treatment will occur if extra pens were assigned to the control, or if pens had to be left out of the analysis because of problems such as sickness or mortality.

4.4.5 Cases to be Considered

The simplest methods for these variables are those methods that are only applicable after the data structure has been reduced to a single proportion in each treatment. But as discussed earlier, to be statistically valid these methods must still take into account the pen-based structure of the experiment. There are procedures that appear to achieve a reduction in data structure while maintaining statistical validity. Two of them are discussed in section 4.4.7. However they do not appear to be suitable for all data sets. In particular they are probably not suitable for data sets with irregularities in the data.

Note: One of these methods is Rao and Scott's method. The acceptance of this method is tentative and depends on the results of a forthcoming evaluation.

Thus there are two cases to consider:

Case 1: Methods for Treatment Proportions

These methods analyse the data after the pen proportions P_{ij} within each treatment have been combined into a single proportion.

Case 2: Methods for Pen Proportions

These methods are employed for situations where the P_{ij} have not been combined.

4.4.6 Recommended Methods for These Cases

No guidelines have been drawn up on how to select the appropriate case for a given data set as this is largely a matter of subjective judgement.

Note - Each method involves a set of one-tailed tests at the 5% level, that test the effect of the test substance at each dose level. For more information on the methods see Appendix A.

Case 1: Methods for Treatment Proportions

The methods recommended are contingency-table methods, since the data sets have the structure of a contingency table for this case.

- Chi-square tests, each test compares the proportion for control with the proportion for a particular dose level

- Fisher's exact tests, each test compares the proportion for control with the proportion for a particular dose level

- Cochran-Armitage tests, each test looks at the trend in treatment proportions from control up to a particular dose level

Case 2: Methods for Pen Proportions

The methods set out are all non-parametric methods. Least-squares methods are considered to be inappropriate for this case because of the discrete nature and skewed distribution of the P_{ij}.

- Mann-Whitney tests, each test compares control with a particular dose level

- Rerandomization tests, each test either compares control with a particular dose level or looks at the trend in treatments from control up to a particular dose level

- Jackknife tests, each test either compares control with a particular dose level or looks at the trend in treatments from control up to a particular dose level

4.4.7 Statistically Valid Methods for Reducing the Data Structure

Method 1. Tabulating at the Pen Level

In some situations the data structure can be reduced to that of a single proportion per treatment without appreciable loss of information, by tabulating entire pens rather than subjects within pens. For example if the pen proportions for a variable consisted mainly of zero values, there would probably be little loss of information in tabulating the proportion of pens that are zero or non-zero for each treatment and analysing those treatment proportions. Similarly if the pen proportions consist mainly of values of 1, the proportion of pens in a treatment that are equal to 1 could be tabulated.

Method 2. Rao and Scott's Method for Combining Pen Proportions

The procedure of pooling the data from all pens in a treatment into a single proportion, and then ignoring the pen information in the data analysis, is not statistically valid as the error in the analysis would not take pen-to-pen variation (as discussed in section 3.6). However Rao and Scott have set out a procedure for combining pen proportions that appears to overcome this problem (Rao and Scott, 1992). Their solution is to pool the data from all pens in a treatment into a single proportion in the usual manner, but in the statistical analysis to employ a variance formula that takes the variation between pens into account.

Let P_{ij} be the pen proportion for pen j in treatment i, with each P_{ij} being equal to y_{ij}/n_{ij}, and let PP_i be the pooled proportion for treatment i. Then

$$PP_i = \Sigma_j y_{ij}/\Sigma n_{ij} = Y_i/N_i$$

Statistical analysis is then run on the PP_i, but they are not treated as if they were the simple proportions Y_i/N_i as this would underestimate the variance of the PP_i. Instead the variance of PP_i is derived from the pen-to-pen variation in P_{ij} within treatments. An effective sample size $(N_i)_{eff}$ is then obtained for each PP_i that corresponds to its variance. In general $(N_i)_{eff}$ is less than N_i.

The PP_i are then put into the form of proportions:

$$PP_i = (Y_i)_{eff}/(N_i)_{eff}$$

where $(Y_i)_{eff}$ is the effective numerator and is defined as $(PP_i)(N_i)_{eff}$. According to Rao and Scott, PP_i can be entered into statistical formulas as if it were the simple proportion $(Y_i)_{eff}/(N_i)_{eff}$ since its denominator corresponds to its accuracy.

Rao and Scott's method is currently being studied for its applicability to AR experiments. A current limitation to its use is that it takes the pooled pen proportions as the optimal estimate of the overall proportion for each treatment, thus assuming that each pen proportion P_{ij} should be weighted by its denominator n_{ij}. This is a drawback since such a weighting scheme may be inappropriate for many AR data sets. However it should be possible to make the method more flexible by extending it to other weighting schemes. Rao and Scott's method is described further in Appendix B, section B.1.

4.5 Methods for Measurement Variables - Eggs or Chicks

4.5.1 Requirements to be Met By These Methods

The criteria are those set out in section 4.2.5, with the following condition added:

Employ least-squares methods only. Non-parametric methods should not be necessary for this variable class and are not usually applied to it.

4.5.2 Basic Variable Characteristics and Model

The data sets for these variables have a three-level structure of eggs or chicks within pens within treatments. The number of eggs or chicks will vary from one pen to another. The data should follow approximately the standard linear model:

$$X_{ijk} = \mu + T_i + E_{ij} + e_{ijk}$$

where X_{ijk} is the data value for the k'th bird in the j'th pen in treatment i, μ is the overall mean, T_i is the effect for treatment i, E_{ij} is the random effect for pen j in that treatment and e_{ijk} is the random error for the k'th data value from that pen. In the standard model the E_{ij} and e_{ijk} have approximately normal distributions and their variances σ_{Eij}^2 and σ_{eijk}^2 are constant or approximately so over all treatments.

Pen Means: The methods considered are all based on least squares analysis of the pen means $\bar{X}_{ij}$ which have the form

$$\bar{X}_{ij} = \mu + T_i + E_{ij} + \bar{e}_{ij}$$

where $\bar{e}_{ij}$ is the mean of the e_{ijk} for that pen. The variance σ_{Xij}^2 of $\bar{X}_{ij}$ is

$$\sigma_{Xij}^2 = \sigma_{Eij}^2 + \sigma_{eij}^2/n_{ij}$$

where n_{ij} is the number of eggs or chicks in the pen for that variable and σ_{eij}^2 is the mean of σ_{eijk}^2.

4.5.3 Common Data Complexities

1. Variation among the n_{ij}. The numbers n_{ij} of eggs or chicks in a pen at any given stage in the experiment will in general vary from pen to pen. Often this variation is quite large.

2. Skewness in the distribution of the X_{ijk}. This may produce non-linearity in the model and cause σ_{Xij}^2 to vary with the size of $\bar{X}_{ij}$. In general this situation can be handled by applying a logarithmic transformation.

3. Variation among the number N_i of pens per treatment. Differences will occur if extra pens were assigned to the control, if pens had to be left out of the analysis because of problems such as sickness, mortality or extraneous reproductive failure, or if there were no surviving eggs or chicks in one or more pens at the stage of the experiment when the variable was measured.

4. Variation among the σ_{Xij}^2. It is expected that there will be inhomogeneity in σ_{Xij}^2 within treatments (in addition to any inhomogeneity that is removable by transformation), because of pen-to-pen variation in the n_{ij}. In addition there could be non-removable variation in σ_{Eij}^2 or σ_{eij}^2. This could be such that σ_{Eij}^2 or σ_{eij}^2 are different for each pen, or they could be constant for all pens within a treatment but vary from one treatment to another.

4.5.4 Cases to be Considered

Only least squares methods were considered for this variables class, the complexity of the method depending on the complexity of σ_{Xij}^2 and also on the N_i. The cases to consider are set out in section 4.3.4 in the discussion of measurement variables for adult birds.

The first three cases are:

Case 1: σ_{Xij}^2 constant over all treatments, N_i all equal

Case 2: σ_{Xij}^2 constant over all treatments, N_i not all equal

Case 3: σ_{Xij}^2 constant over all pens in a treatment, but varies between treatments

For these cases unweighted least squares methods can be applied. However they are unlikely to be suitable, since σ_{Xij}^2 depends on n_{ij} and in general the n_{ij} have a large variation from pen to pen. Case 4, which requires weighted least squares analysis, is the case that is expected to be applicable for most data sets.

Case 4: σ_{Xij}^2 varies from pen to pen within a treatment

Note: It is assumed that any variation in σ_{Xij}^2 that is due to skewness alone has been removed by transformation.

Note: Methods for determining an appropriate weighting scheme and for testing for homogeneity of variance are described in Appendix B.

4.5.5 Recommended Methods

No guidelines have been given concerning how to determine which case to select for a given data set, as this is largely a matter of subjective judgement.

Note - Each method involves a set of one-tailed tests at the 5% level, that test the effect of the test substance at each dose level. For more information on the methods see Appendix A.

Cases 1 to 3. Unweighted Least Squares Methods

The methods are the same as those listed for measurement variables for adult birds in section 4.3.5, cases 1 to 3. Since these cases are unlikely to occur, the methods are not repeated here.

Case 4. Weighted Least Squares Methods

- t-tests, adjusted for weighted analysis, each test compares the weighted control mean with the weighted mean for a particular dose level

- Linear trend tests, adjusted for weighted analysis, each test looks at the trend in weighted treatment means from control up to a particular dose level

4.6 Methods for Proportions or Counts - Eggs or Chicks

4.6.1 Requirements to be Met By These Methods

The criteria are those set out in section 4.2.5.

4.6.2 Conversion of Counts to Proportions

It is assumed that counts will be converted into proportions before analysis, by dividing by a suitable denominator. This will take into account the fact that the counts have a restricted range, since there is a limit to the number of eggs that can be laid in a pen. The denominator n_U represents an upper limit to the number of eggs or chicks in a pen, so that after conversion the proportions are restricted to the range of 0 to 1.

For most data sets, n_U can be taken as the theoretical maximum n_{TH} which is the number of eggs produced if each female bird were to lay one egg per day during the egg laying period. However pen counts may occasionally be greater than n_{TH}, as some eggs may be laid just before the designated egg-laying period starts. To cover all possibilities it is suggested that n_U be set either at n_{TH} or at a value that is 20% larger than the largest number of eggs laid in any pen, whichever is greater.

Note: If the counts for a variable are small compared to n_U, the upper limit will have little effect and the counts will tend to follow a Poisson distribution. Normally a square root transformation would be applied to such variables. However this is not necessary in this case. The reason is that an angular transformation is applied to proportions prior to analysis, and this transformation is in fact equivalent to the square root transformation for these variables. This is discussed in Appendix B, section B.3.2.

4.6.3 Basic Variable Characteristics and Model

In the terminology of this report, proportional variables are either 'actual proportions' or 'estimated proportions'. Actual proportions are standard proportions, and estimated proportions are the product of two or more actual proportions (see section 2.3.2 for more information). For actual proportions, each data value is the proportion for a particular pen and the data set has a structure of pens within treatments. Let P_{ij} be the proportion for pen j in treatment i. Then

$$P_{ij} = y_{ij}/n_{ij}$$

where n_{ij} is the number of eggs or chicks in the pen for that variable and y_{ij} is the number of the n_{ij} for which some characteristic was recorded (e.g. the number that survived a particular phase of the experiment). For proportions that have been converted from counts, n_{ij} is equal to the assigned value n_U.

The standard linear model for P_{ij} is

$$P_{ij} = \mu + T_i + E_{ij}$$

where μ is the overall average, T_i is the effect of treatment i and E_{ij} is the random error associated with that pen.

A complicating factor is that in general E_{ij} is the sum of two components, one that depends on the denominator n_{ij} and one that is independent of n_{ij}. Models for E_{ij} and its variance are discussed in some detail in section 4.6.5.1. For estimated proportions the situation is still more complicated (see section 4.6.5.2).

4.6.4 Common Data Complexities

1. Variation among the n_{ij}. n_{ij} will generally vary widely from pen to pen. (An exception occurs if the variable is a count that was converted to a proportion, in which case n_{ij} is constant and equal to the assigned value n_U.)

2. The distribution of the P_{ij}. The distribution of P_{ij} is complicated by the fact that there are two error components.

3. Extreme data values. Extraneous low values of P_{ij} could occur due to reproductive failure in one or more pens for reasons unrelated to the treatments applied.

4. Variation in the N_i. The number N_i of pens in a treatment can vary from one treatment to another if extra pens were assigned to the control, if pens have been removed from the experiment due to problems such as sickness, mortality or extraneous reproductive failure, or if there were no surviving eggs or chicks in one or more pens at the stage of the experiment when the variable was measured.

4.6.5 The Variance of the Pen Proportions

Modelling the Variance of Actual Proportions

Consider an actual pen proportion P_{ij} with denominator n_{ij}. Its variance is that of its error E_{ij}, which can be considered to consist of two components:

$$E_{ij} = b_{ij} + (eb)_{ij}$$

Here b_{ij} is the 'binomial' component that represents the deviation of P_{ij} from the expected value for that pen (i.e. sampling error), and $(eb)_{ij}$ is the 'extra-binomial' component which represents the variation among these expected values from pen to pen within a treatment (i.e. real differences between pens).

The variance σ_{Pij}^2 of P_{ij} is given by the sum of the variances of b_{ij} and $(eb)_{ij}$:

$$\sigma_{Pij}^2 = \sigma_{bij}^2 + \sigma_{ebij}^2$$

Since b_{ij} has a binomial distribution, σ_{bij}^2 is equal to $P_{ij}(1-P_{ij})/n_{ij}$. The form of σ_{ebij}^2 will generally not be known exactly, but it is independent of n_{ij}.

It appears to be reasonable to express σ_{ebij}^2 as $\lambda\ P_{ij}(1-P_{ij})$ for some constant λ. The reason is that it should have the same tendency as σ_{bij}^2 to be a maximum when P_{ij} is near .5 and to decrease to 0 as P_{ij} increases to 1 or decreases to 0. With this assumption, we can write

$$\sigma_{Pij}^2 = P_{ij}(1-P_{ij})/n_{ij} + \lambda P_{ij}(1-P_{ij})$$

$$= P_{ij}(1-P_{ij})(1/n_{ij} + \lambda)$$

The factor $P_{ij}(1-P_{ij})$ can be removed from the variance by applying an angular transformation:

$$A_{ij} = \arcsin(\sqrt{P_{ij}})$$

Note: An angular transformation is probably not necessary if the P_{ij} are within the range of 0.2 to 0.8, as the factor $P_{ij}(1-P_{ij})$ is relatively constant within that range. It is also possible for some cases that a different transformation would be more suitable for equalizing the variance. A discussion of transformations for proportions is presented in Appendix B.

The transformed proportions A_{ij} can be expressed either in degrees or in radians. If they are expressed in degrees, their variance σ_{Aij}^2 is:

$$\sigma_{Aij}^2 = 821[(1/n_{ij}) + \lambda] \qquad (1)$$

The simplest situation for analysis purposes is one where σ_{Aij}^2 is constant or approximately constant. This occurs if the factor $[(1/n_{ij}) + \lambda]$ is approximately constant, which requires that λ be large enough that the variation in $1/n_{ij}$ does not have much effect. σ_{Aij}^2 also will be constant for counts converted to proportions, since n_{ij} is set to n_U in these cases and there is no variation in $(1/n_{ij})$. If σ_{Aij}^2 is constant, it is not necessary to fit a model to it.

However in general it is expected that σ_{Aij}^2 will not be constant, and must be modelled by estimating λ and substituting into equation (1). Some possible procedures for modelling σ_{Aij}^2 are discussed in Appendix B, section B.2.2.

The Variance of Estimated Proportions

Consider an estimated proportion P_{ij} that is the product of two actual proportions:

$$P_{ij} = Q_{ij} R_{ij}$$

where Q_{ij} and R_{ij} are actual proportions. The variance of P_{ij} is too complicated to model precisely. But a first order approximation can be derived from the formula for the variance of a product of two variables:

$$\sigma_{Pij}^2/P_{ij}^2 = \sigma_{Qij}^2/Q_{ij}^2 + \sigma_{Rij}^2/R_{ij}^2 + 2\,\rho\,\sigma_{Qij}\sigma_{Rij}/Q_{ij}R_{ij} \qquad (2)$$

where ρ is the correlation coefficient between Q_{ij} and R_{ij}. Using this formula it is possible to estimate σ_{Pij}^2 for each pen, provided that estimates have been developed for ρ, σ_{Qij}^2 and σ_{Rij}^2.

The procedure suggested is to derive models for σ_{Qij}^2 and σ_{Rij}^2, and to obtain an estimate of ρ from the data for Q_{ij} and R_{ij} using the standard formula for a correlation coefficient. Estimates of ρ, σ_{Qij}^2 and σ_{Rij}^2 can then be entered into equation (2) to obtain a value for σ_{Pij}^2. A corresponding procedure could be developed for the case where P_{ij} is the product of three or more actual proportions.

Formula (2) can also be expressed in terms of angular-transformed variances by substituting $\sigma_{Aij}^2 P_{ij}(1-P_{ij})$ for σ_{Pij}^2, and similarly for σ_{Qij}^2 and σ_{Rij}^2.

4.6.6 Cases to be Considered

The situation is quite complex for this variable class, as there are a wide variety of approaches that could be appropriate for a given data set depending on the circumstances.

Least-Squares Methods

Least-squares methods are expected to be the most efficient for analysing the pen proportions, provided that they can deal with the complexities of the variance of the P_{ij}. It is assumed that the transformation

$$A_{ij} = \arcsin(\sqrt{P_{ij}})$$

has been applied (with A_{ij} measured in degrees) and that the variance of the A_{ij} has the form

$$\sigma_{Aij}^2 = 821[(1/n_{ij}) + \lambda]$$

The optimum situation is one where σ_{Aij}^2 is constant or approximately so, and unweighted least squares methods can be employed to analyse the A_{ij}. But this situation is unlikely as discussed in section 4.6.5, since it would require that λ be large enough to smooth out any variation in $(1/n_{ij})$.

If σ_{Aij}^2 is not constant, then any least squares analysis will have to involve a weighting scheme in which weights are applied to the A_{ij}. It will be necessary to fit a model to σ_{Aij}^2, estimate it for each A_{ij}, and set the weights to the inverse of these variance estimates. Some procedures for this are discussed in Appendix B, section B.2.2.

Methods for Qualitative Data

Methods for qualitative data may also be applicable, depending on the circumstances. For this to be the case it is necessary that the structure of the data first be reduced to that of a single proportion in each treatment. Methods for qualitative data such as contingency-table methods can then be applied, provided that they take into account the variance between pens in a treatment.

As discussed in section 3.6, it is not valid to simply pool the pen proportions in a treatment and then analyse the pooled proportions as if they were simple proportions, as this does not take pen-to-pen variation into account. However a recently developed procedure appears to overcome this problem (Rao and Scott, 1992). This procedure is described briefly in section 4.4.7, and in more detail in Appendix B, section B.1.1.

The suitability of Rao and Scott's procedure for analysing variables in this class depends on its general validity, and also on whether it is flexible enough to accommodate data sets with moderate or large extrabinomial error terms. In its current form, the procedure assumes that the pen proportions P_{ij} should be weighted by their denominators n_{ij} in calculating treatment proportions. But this weighting scheme is probably not suitable for many AR data sets, since it assumes that the extra-binomial variance component is small (as discussed in section 4.6.5).

To be applicable to AR data sets in general, it will probably be necessary to make the method more flexible by extending it to include other weighting schemes. Thus the applicability of this method depends on whether a suitable weighting scheme can be identified. It is currently being studied for its applicability to avian reproduction experiments and is tentatively recommended.

Note: Rao and Scott's procedure is probably not suitable for data sets where there are irregularities in the distribution of the pen proportions.

Non-Parametric Methods

There may be data sets for which neither least-squares methods nor qualitative methods are suitable, for reasons such as irregularities in the data. For these cases the pen proportions can be analysed using non-parametric methods that are more robust and less affected by such problems.

Summary: In all there are six cases to consider. The first three cases are those for which the transformed variance σ_{Aij}^2 is constant, and thus unweighted least squares methods area applicable. They correspond to the cases set out in section 4.3.4 in the discussion of measurement variables for adult birds.

Case 1: σ_{Aij}^2 is constant over all treatments, and the number N_i of pens per treatment are all equal

Case 2: σ_{Aij}^2 is constant over all treatments, but the N_i are not all equal

Case 3: σ_{Aij}^2 is constant over all pens in a treatment, but varies from one treatment to another

However for most data sets σ_{Aij}^2 will probably not be constant within treatments, so that cases 1 to 3 would not be appropriate. The choice is then among cases 4, 5 and 6:

Case 4: σ_{Aij}^2 is not constant within treatments, but can be modelled in terms of n_{ij} or other parameters. From this model a weighting scheme can be derived. Thus weighted least squares methods can be employed.

Case 5: The P_{ij} can be combined into a single proportion in each treatment, and these treatment proportions can be analysed by methods for qualitative data (taking pen-to-pen variation into account).

Case 6: σ_{Aij}^2 can not be modelled, and qualitative methods can not be applied in a statistically valid manner. This requires that non-parametric methods be employed.

4.6.7 Recommended Methods for These Cases

No guidelines have been set out concerning how to determine which case to select for a given analysis, as this is largely a matter of subjective judgement.

Note - Each method involves a set of one-tailed tests at the 5% level, that test the effect of the test substance at each dose level. For more information on the methods see Appendix A.

Cases 1 to 3. Unweighted Least Squares Methods

The methods are the same as those listed for measurement variables for adult birds in section 4.3.5, cases 1 to 3. Since these cases are unlikely to occur, the methods are not repeated here.

Case 4. Weighted Least Squares Methods

- t-tests, each test compares the control mean with the mean for a particular dose level

- Linear trend tests, each test looks at the trend in treatment means from control up to a particular dose level

Case 5. Methods for Qualitative Data

- Chi-square tests, each test compares the proportion for control with the proportion for a particular dose level

- Fisher's exact tests, each test compares the proportion for control with the proportion for a particular dose level

- Cochran-Armitage tests, each test looks at the trend in treatment proportions from control up to a particular dose level

Case 6. Non-Parametric Methods

- Mann-Whitney tests, each test compares control with a particular dose level

- Rerandomization tests, each test either compares control with a particular dose level or looks at the trend in treatments from control up to a particular dose level

- Jackknife tests, each test either compares control with a particular dose level or looks at the trend in treatments from control up to a particular dose level

5. IDENTIFICATION OF THE NOEC

Position of CWS on the NOEC

The position of the Canadian Wildlife Service is that for each variable analysed in an AR experiment, the dose level at which the treatment effect (if any) begins should be determined by identifying the NOEC (the highest dose level at which there is no observed effect). Reasons for this are discussed in Mineau, Boersma and Collins (in press). The NOEC is also specified in the OECD protocol as one of the results to be produced.

Identifying the NOEC if Results are Consistent

The procedure to identify the NOEC involves testing the effect at each dose level and examining the pattern of significant and non-significant results. Normally this pattern is consistent from one dose level to another in that the effect will be non-significant for all dose levels up to a certain value and significant for all levels above this value. However, exceptions can occur. Examples of consistent and inconsistent patterns are:

Dose	Consistent	Inconsistent
Low	Not Sig.	Not Sig.
Medium	Not Sig.	Sig.
High	Sig.	Not Sig.

If a pattern is consistent the NOEC is determined by finding the lowest level at which there is a significant effect. The NOEC is the level immediately below this. For the consistent pattern above, the NOEC is the medium dose. If the effect is not significant at any of the dose levels the NOEC is at or above the highest dose level in the experiment.

Identifying the NOEC if Results are Inconsistent

If a pattern is not consistent, the situation is more difficult. It is up to the experimenter to decide how to identify the NOEC and to support his or her decision. One possible approach is to derive two different NOEC values, one by working down from the highest dose and one by working up from the lowest. Denote these by NOEC1 and NOEC2 respectively.

Working down from the highest dose, NOEC1 is the first dose level at which there is a non-significant effect. For the inconsistent pattern given on the previous page, NOEC1 is the high dose. Working up from the lowest dose, NOEC2 is the dose level just below the first level at which there is a significant effect. For the inconsistent pattern given on the previous page, the first level with a significant effect is the medium dose and therefore NOEC2 is the low dose.

NOEC1 and NOEC2 bracket the NOEC. In the event of an inconsistent pattern of results, both values could be presented with the statement that the NOEC is somewhere within their range. An explanation could also be required as to why this inconsistency occurred.

6. DATA QUALITY CONSIDERATIONS

6.1 The Need for Data Quality Control Measures

Statistical methods are in general conservative in detecting significant treatment effects. For AR experiments this presents the danger that if the data are of poor quality, then harmful effects of a test substance may go undetected. There is a particular concern for data from the birds in the control, since poor data quality for the control would cause the control mean to be an artificially low standard against which to compare the means for the different dose levels. It appears to be advisable to have measures in place that will help to ensure that the data quality is acceptable.

6.2 Measures Based on Variable Means

The issue of data quality is given prominence in the OECD protocol, in which a number of criteria are specified that the data for the control birds should meet. These criteria set lower limits for the mean value for control for certain key variables:

- mortality among the adult birds
- number of 14-day surviving chicks
- shell thickness

For certain other variables a normal range is set out, and the mean value for control is expected to be within or close to this range. These variables are

- number of eggs laid
- proportion of cracked eggs
- proportion of viable embryos
- proportion of eggs set that hatch
- proportion of hatchlings that survive to 14 days

The other protocols do not discuss data quality to the same extent. The 1986 EPA protocol states that sickness, injuries or excessive mortality among the chicks may indicate that the quality of the adult birds in the experiment was not adequate, but does not set out specific quality control measures. The ASTM protocol does not consider the question of data quality.

Data quality is also mentioned in some of the submissions, mainly with respect to the possible effects of sickness or injury. Some of the submissions state that if there is disease or mortality in more than a certain proportion of the pens, the experiment may be rejected.

6.3 Measures to Detect Reproductive Failure

A particular concern for AR experiments is the possibility that extraneous low data values can occur as a result of reproductive failure in particular pens, for reasons unrelated to the treatments applied (Picirillo and Quesenberry, 1980). The presence of such a value in the control pens would be of particular concern as it would artificially lower the control mean. The presence of extraneous low data values would also reduce the power of statistical tests by artificially raising the pen-to-pen variation.

Measures could be set out to detect reproductive failure in the control pens, by specifying minimum acceptable pen values for certain variables just as minimum overall control values are specified in the OECD protocol.

Another possible approach for identifying cases of extraneous reproductive failure in particular pens is to employ a statistical procedure for detecting 'outliers' or extreme data values that do not belong in a data set. This is a difficult area as many different procedures have been developed - some for general use and some for specific contexts. Care must be taken not to identify genuine data values by mistake. The choice of method is essentially a subjective one. Some possible approaches to the detection of outliers are discussed in Appendix B.

6.4 Measures Based on Statistical Power

There may also be a need for measures to ensure that the statistical tests being applied have sufficient power to detect treatment effects if those effects are large enough to be potentially harmful. The reason for additional measures is that even if treatment means are within an acceptable range and there are no obvious cases of reproductive failure, the power of the tests could be inadequate due to excessive pen-to-pen variation within treatments.

A possible approach to deriving a measure to ensure sufficient statistical power could involve specifying a minimum size of treatment effect that the experiment should be able to detect. A possible minimum size is in fact suggested by the 1982 EPA protocol which states that the objective of an AR experiment is to detect reproductive impairment at or above 20%.

Suppose that a figure of 25% is selected. For certain key variables the per cent effect of the test substance could be calculated at each dose level:

$$\text{Per Cent Effect} = \frac{\text{Control Mean} - \text{Dose Level Mean}}{\text{Control Mean}} \times 100\%$$

If for any dose level the per cent effect is 25% or greater but the test of that dose level is not statistically significant, the power of the experiment could be considered to be insufficient.

APPENDIX A - INFORMATION ON STATISTICAL METHODS

Table of Contents

A.1 Introduction

In this appendix the methods are divided into three classes:

Recommended Methods

These are considered to be the most promising for AR experiments. They meet the objectives and the criteria for validity and efficiency set out in section 4.2.1 and are recommended in section 4.

Potential Methods

This section contains information on some additional methods that are relevant to AR experiments, but appear to be too complex to be applied on a routine basis.

Methods Not Recommended

These are methods that are currently employed, but are not recommended because they do not appear to meet the criteria set out in section 4.2.1.

A.2 Recommended Methods

A.2.1 Least Squares Methods

A.2.1.1 t-Tests

Let X_{ij} be the j'th data value within treatment i. (Here X_{ij} corresponds to a pen mean $\bar{X}_{ij}$ in the notation of section 5.) For application to AR data sets, let the control be treatment 1 and the dose levels be treatments 2, 3, 4, etc. Let σ_{Xij}^2 be the variance of X_{ij}, let N_i be the number of data values (i.e. number of pens) in treatment i, let $\bar{X}_i$ be the mean for treatment i, and let S_i^2 be the variance between pen means within treatment i. Then

$$\bar{X}_i = \Sigma_j X_{ij}/N_i \qquad \text{and} \qquad S_i^2 = [\Sigma_j (X_{ij}-\bar{X}_i)^2]/(N_i-1)$$

There are a number of different versions of the t-test, with the choice in a given case depending on σ_{Xij}^2 and N_i.

Standard t-Test (σ_{Xij}^2 constant over all treatments, N_i all equal)

This test is described in all statistics texts. Let N be the number of data values in each treatment. To test the difference between the control mean $\bar{X}_1$ and the mean $\bar{X}_k$ for treatment k, first calculate the combined variance S_{1k}^2 among pen means within treatments 1 and k:

$$S_{1k}^2 = (S_1^2 + S_k^2)/2$$

The variance S_{X1-Xk}^2 of the difference $\bar{X}_1-\bar{X}_k$ is given by

$$S_{X1-Xk}^2 = 2S_{1k}^2/N$$

The test statistic to compare control and dose level k is

$$t = (\bar{X}_1-\bar{X}_k)/S_{X1-Xk}$$

which has a t distribution with 2(N-1) degrees of freedom.

t-Test Adapted to Unequal N_i (σ_{Xij}^2 constant over all treat., N_i not all equal)

This test is similar to the standard t-test, with some adjustments. The combined variance S_{1k}^2 within treatments 1 and k is

$$S_{1k}^2 = [(N_1-1)S_1^2 + (N_k-1)S_k^2]/[(N_1-1) + (N_k-1)]$$

The variance of the difference $\bar{X}_1-\bar{X}_k$ is given by

$$S_{X1-Xk}^2 = S_{1k}^2/N_1 + S_{1k}^2/N_k$$

The statistic is $$t = (\bar{X}_1-\bar{X}_k)/S_{X1-Xk}$$

which has a t-distribution with $[(N_1-1)+(N_k-1)]$ degrees of freedom.

t-Test Adapted to Unequal Variances (σ_{Xij}^2 constant within treatments but varies from one treatment to another)

Here further adjustments are necessary. The S_i^2 are employed directly and are not combined into a common estimate.

The variance of $\bar{X}_1-\bar{X}_k$ is $$S_{X1-Xk}^2 = S_1^2/N_1 + S_k^2/N_k$$

and the statistic t_{UV} is $$t_{UV} = (\bar{X}_1-\bar{X}_k)/S_{X1-Xk}$$

t_{UV} does not follow a simple t-distribution, and this causes the process of testing t_{UV} to be somewhat involved. There are several different test procedures, which are described in most statistical texts (e.g. Snedecor and Cochran, 1967, p.115). The most powerful is the Welch-Aspin test (Welch, 1947) which requires special tables.

An alternative procedure developed by Cochran is slightly less powerful but more commonly used as it employs the standard t-table (Cochran, 1964). If N_1 is equal to N_k the significance of t_{UV} can be determined directly from the t-table using N_1-1 degrees of freedom. However if N_1 is not equal to N_k, the t-table can still be used but a rather complicated calculation is required to get the significance level of t_{UV}. A good discussion on the effects of unequal variances is given in Miller (1986).

t-test for Weighted Analysis (σ_{Xij}^2 varies within treatments)

Let w_{ij} be the weight applied to X_{ij}. Assume the w_{ij} are the inverses of the variance of X_{ij}. The mean for treatment i is the weighted average $\bar{X}_{Wi}$ of the X_{ij}:

$$\bar{X}_{Wi} = \Sigma_j w_{ij} X_{ij} / W_i$$

where $W_i = \Sigma_j w_{ij}$.

The variance of $\bar{X}_{W1} - \bar{X}_{Wk}$ is $S_{XW1-XWk}^2 = 1/W_1 + 1/W_k$

and the test statistic is $t_W = (\bar{X}_{W1} - \bar{X}_{Wk}) / S_{XW1-XWk}$

This formula is actually a special case of the formula for the mean of a stratified sample, as discussed in some texts on sampling (e.g. Cochran, 1977, p. 91-96). The distribution of t_W can be approximated by a t-distribution with a reduced number of degrees of freedom. The test uses the t-table, but the effective number n_e of degrees of freedom must be worked out. A formula for this number was derived by Satterthwaite (1946). For our case the formula is

$$n_e = \frac{(1/W_1 + 1/W_k)^2}{\Sigma_j \frac{[w_{1j}^2/(n_{1j}-1)]}{W_1^4} + \Sigma_j \frac{[w_{kj}^2/(n_{kj}-1)]}{W_k^4}}$$

A.2.1.2 LSD Tests

Least Significant Difference (LSD) tests are described in most standard texts (e.g. Snedecor and Cochran, 1967, p.271). They are the same as t-tests, except that the variance S_{X1-Xk}^2 of each mean difference $\bar{X}_1 - \bar{X}_k$ is calculated using the data from all of the treatments rather than the data from the two treatments in the test. This increases the accuracy of the variance compared to that for the t-test.

Two different versions of the LSD test are described - one with equal numbers N_i of data values per treatment and one with unequal N_i. The notation used is the same as in section A.2.1.1. For both versions the variance σ_{Xij}^2 of the data values must be approximately constant over all treatments. The LSD test is not applicable to data sets with unequal variances.

Standard LSD Test (σ_{Xij}^2 constant over all treatments, N_i all equal)

Let N be the number of pens per treatment and let R be the number of treatments. The combined variance S^2 within all treatments is:

$$S^2 = [\Sigma_i\Sigma_j(X_{ij}-\bar{X}_i)^2]/[R(N-1)]$$

The variance of $\bar{X}_1-\bar{X}_k$ is $S_{X1-Xk}^2 = 2S^2/N$

and the test statistic is $t_{LSD} = (\bar{X}_1-\bar{X}_k)/S_{X1-Xk}$

t_{LSD} has a t-distribution with R(N-1) degrees of freedom.

LSD Test Adapted to Unequal N_i (σ_{Xij}^2 constant over all treat., N_i not all equal)

The combined variance S^2 within all treatments is:

$$S^2 = [\Sigma_i\Sigma_j(X_{ij}-\bar{X}_i)^2]/[\Sigma(N_i-1)]$$

The variance of $(\bar{X}_1-\bar{X}_k)$ is $S_{X1-Xk}^2 = S^2/N_1 + S^2/N_k$

and the test statistic is $t_{LSD} = (\bar{X}_1-\bar{X}_k)/S_{X1-Xk}$

t_{LSD} has a t-distribution with degrees of freedom equal to $\Sigma(N_i-1)$.

A.2.1.3 The Williams Test

This test, described in Williams (1971) and Williams (1972), was designed specifically for experiments that consist of a control and a series of dose levels of a test substance. It is only applicable to data sets where the variance σ_{Xij}^2 of the data values is approximately constant. The number N_i of data values per treatment must be the same for all dose levels, and the number of data values for control must be equal to or greater than for the dose levels.

The test is similar to an LSD test, the only difference being that the treatment means may have been adjusted prior to the test. This feature is designed to overcome certain problems of interpretation that can occur with t-tests or LSD tests.

With t-tests or LSD tests it can happen that the results of the tests at different dose levels are not consistent with each other. It is possible to have a significant effect at a lower level but a non-significant effect at a higher level (this problem is discussed in Appendix B, sections B.2 and B.3. For example suppose that the treatments are a control C and dose levels D1, D2 and D3, and that the effect at D3 is less than at D2 (see Figure A1). It is possible that the effect at D2 is significant while that at D3 is not significant.

Williams' innovation is to remove the possibility of inconsistent results in these cases by adjusting the dose level means. In this process, the problem means are averaged and the average is then substituted for these means. This produces a set of treatment means that form a monotonic series. In the example the original D2 and D3 means would both be replaced by the D2-D3 average.

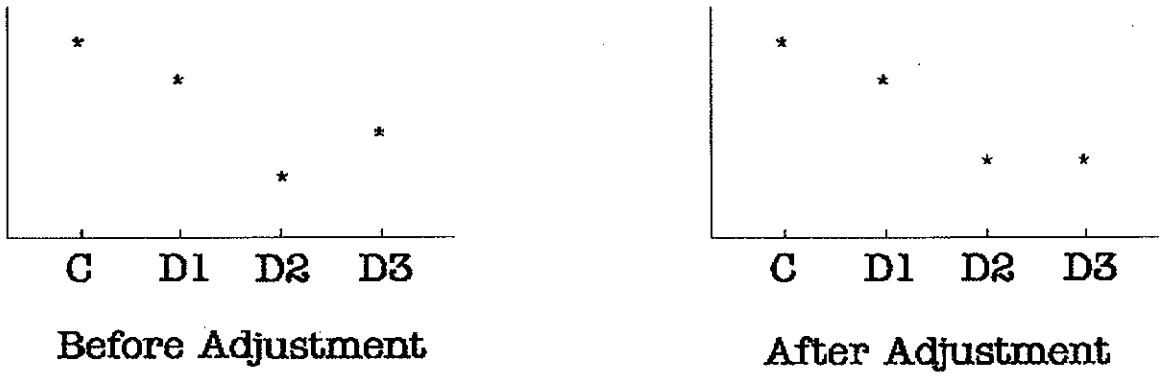

Figure A1. Averaging of consecutive means to produce a monotonic series.

When the adjusted dose level means are compared to the control mean, the results are always consistent which simplifies the interpretation of the results.

Standard Williams Test (σ_{Xij}^2 constant over all treatments, N_i all equal)

This has the same form as the standard LSD test, in that each dose level mean is compared to the control mean using a within-treatments variance that is based on the data from all treatments (see section A.2.1.2). The variance S_{X1-Xk}^2 of the difference between the control and dose level means is the same as in an LSD test, and the test statistic t_{WIL} is given by:

$$t_{WIL} = (\bar{X}_1 - \bar{X}'_k)/S_{X1-Xk}$$

If the means have been adjusted prior to analysis, $\bar{X}'_k$ is the adjusted mean. If they have not been adjusted $\bar{X}'_k$ is the original mean. The control mean X_1 is always the original mean. The test requires special tables of critical values, since the distribution of the test statistic deviates from the t-distribution because of the provision for averaging of dose level means. Williams' 1971 paper gives the table of critical values for the one-tailed version of the test and the table for the two-tailed version is in his 1972 paper.

Williams Test Adjusted for Extra Control Values (σ_{Xij}^2 constant over all treatments, N_i all equal except for extra control pens)

This has the same form as the LSD test for unequal N_i (see section A.2.1.2). The variance S_{X1-Xk}^2 of the difference between the control and dose level means is the same as in an LSD test for unequal N_i, and the test statistic t_{WILL} is given by:

$$t_{WILL} = (\bar{X}_1 - \bar{X}'_k)/S_{X1-Xk}$$

If the means have been adjusted prior to analysis, $\bar{X}'_k$ is the adjusted mean. If they have not been adjusted $\bar{X}'_k$ is the original mean. The control mean X_1 is always the original mean. Critical values for this test are obtained by adjusting the critical values for the standard Williams test. Williams' 1972 paper gives a formula for this adjustment.

The Williams test has advantages over the t-test and the LSD test concerning ease of interpretation of the results, but would probably have only limited application for AR data sets. The reason is that it can not be employed in cases where the number of pens per treatment varies among dose levels or where the accuracy of the pen means is not constant.

A.2.1.4 Trend Tests

Trend tests are carried out over a set of consecutive treatments, and so are only applicable to data sets where the treatments are ordered. The test for trend over a set of treatments involves the linear contrast for that set. These contrasts are given in most statistical texts. The coefficients for contrasts of 2 to 5 treatments are:

2 Treatments:	1, -1
3 Treatments:	1, 0, -1
4 treatments:	3, 1, -1, -3
5 treatments:	2, 1, 0, -1, -2

For example if an experiment consists of a control C and increasing dose levels D_1, D_2 and D_3, and the treatment means are M_C, M_{D1}, M_{D2} and M_{D3}, the linear contrasts from C to D1, C to D2 and C to D3 (for testing for effects at D1, D2 and D3 respectively) are:

C to D1:	$M_{D1} - M_C$
C to D2:	$M_{D2} - M_C$
C to D3:	$3M_{D3} + M_{D2} - M_{D1} - 3_{MC}$

Note: The trend tests for effects at D1 and D2 are the same as the LSD tests.

Note: It is assumed that the control and the dose levels can be treated as if they are all equally spaced along some axis. If this is not reasonable, an alternate spacing will have to be devised and different linear contrasts worked out.

As with the t-test and LSD test, there are a number of different versions of the trend test that could be applicable to AR data sets. The appropriate choice for a given case depends on the variance σ_{Xij}^2 of the pen means (or proportions) and on the number N_i of pens per treatment. The notation used is described in section A.2.1.1.

Standard Trend Test (σ_{Xij}^2 constant over all treatments, N_i all equal)

The trend test for 2 or 3 treatments is the same as the LSD test. To run a trend test for 4 or more treatments, let the linear contrast be

$$F = \Sigma \lambda_i \bar{X}_i$$

Thus for a 4-treatment contrast the values of λ_1 to λ_4 are -3, -1, 1 and 3.

Then calculate the combined variance S^2 within all treatments:

$$S^2 = [\Sigma_i \Sigma_j (X_{ij} - \bar{X}_i)^2] / [R(N-1)]$$

where N is the number of data values per treatment and R is the number of treatments. The variance S_F^2 of F is

$$S_F^2 = \Sigma\lambda_i^2 S_{Xi}^2$$

where S_{Xi}^2 is the variance of $\bar{X}_i$ and is equal to S^2/N. Thus

$$S_F^2 = (\Sigma\lambda_i^2)S^2/N$$

The test statistic is $t_{TR} = F/S_F$

which has a t-distribution with R(N-1) degrees of freedom.

Trend Test Adapted to Unequal N_i (σ_{Xij}^2 constant over all tr., N_i not all equal)
The test for 2 or 3 treatments is the same as the LSD test. To test 4 or more treatments, let the linear contrast be

$$F = \Sigma\lambda_i\bar{X}_i$$

The combined variance S^2 within all treatments is

$$S^2 = [\Sigma_i\Sigma_j(X_{ij}-\bar{X}_i)^2]/[\Sigma(N_i-1)]$$

The variance of F is $S_F^2 = \Sigma\lambda_i^2 S_{Xi}^2 = S^2\Sigma(\lambda_i^2/N_i)$

since S_{Xi}^2, the variance of $\bar{X}_i$, is equal to S^2/N_i. The test statistic t_{TR} is

$$t_{TR} = F/S_F$$

which has a t-distribution with degrees of freedom equal to $\Sigma(N_i-1)$.

Trend Test Adapted to Unequal Variances (σ_{Xij}^2 constant within treatments but varies from one treatment to another)
For this case the within-treatment variances S_i^2 are employed directly and are not combined into a common estimate. The test for 2 or 3 treatments is the same as the t-test for unequal variances. To test 4 or more treatments, let the linear contrast be

$$F = \Sigma\lambda_i\bar{X}_i$$

The variance of F is $S_F^2 = \Sigma\lambda_i^2 S_{Xi}^2 = \Sigma(\lambda_i^2 S_i^2/N_i)$

since S_{Xi}^2, the variance of $\bar{X}_i$, is equal to S_i^2/N_i.

The test statistic $t_{TR} = F/S_F$

has a distribution that approximates a t-distribution. Critical values can be obtained from the t-table, but a special calculation is required for the effective number n_e of degrees of freedom. An approximate formula is given by Satterthwaite (1946):

$$n_e = [\Sigma(\lambda_i^2 S_i^2/N_i)]^2 \; / \; \Sigma[(\lambda_i^2 S_i^2/N_i)^2/(N_i-1)] \qquad (A1)$$

Trend Test for Weighted Analysis (σ_{Xij}^2 varies within treatments)

Let w_{ij} be the weight applied to X_{ij}. Assume that w_{ij} is the inverse of the variance of $\bar{X}_{ij}$. The mean for treatment i is the weighted average $\bar{X}_{Wi}$ of the pen means:

$$\bar{X}_{Wi} = \Sigma_j w_{ij} X_{ij}/W_i \qquad \text{where} \qquad W_i = \Sigma_j w_{ij}.$$

The test for 2 or 3 treatments is the same as the weighted t-test. To test 4 or more treatments, let the linear contrast be

$$F = \Sigma\lambda_i \bar{X}_{Wi}$$

Its variance is $S_F^2 = \Sigma\lambda_i^2 S_{XWi}^2 = \Sigma\lambda_i^2/W_i$

since S_{XWi}^2, the variance of $\bar{X}_{Wi}$, is equal to $1/W_i$.

The test statistic is $t_{TR} = F/S_F$

which has a distribution that approximates a t-distribution. The t-table can be used for critical values, but a special calculation is required to get the effective number n_e of degrees of freedom. Equation (A1), which gives n_e for the case of unequal variances (Satterthwaite, 1946), can be modified by substituting $(1/W_i)$ for (S_i^2/N_i):

$$n_e = [\Sigma(\lambda_i^2/W_i)]^2 \; / \; \Sigma[(\lambda_i^2/W_i)^2/(N_i-1)]$$

A.2.1.5 The Abelson-Tukey Test

This test, described in Abelson and Tukey (1963), is similar to a trend test but employs a contrast which gives greater weight to the more extreme treatments and is intermediate between a trend test and a t-test or LSD test. It assumes equal numbers of data values per treatment and a constant variance for the data values. As with the trend test, it is only applicable to sets of 4 or more ordered treatments and it assumes that the treatments are equally spaced in some scale. The coefficients for 4 and 5 treatments are:

4 treatments: -.866, -.134, .134, .866
5 treatments: -.894, -.201, 0, .201, .894

For the case of a control C and dose levels D1, D2 and D3, the Abelson-Tukey contrast for C to D3 would be $(.866M_{D3} + .134M_{D2} - .134M_{D1} - .866M_C)$ where M_C, M_{D1}, M_{D2} and M_{D3} are the treatment means. The variance for this contrast can be obtained from the formula for a linear contrast given in section A.2.1.4 for the standard trend test, by substituting -.866, -.134, .134 and .866 for the λ_i.

According to Abelson and Tukey, their contrast is the optimal one for situations where the pattern of treatment effect is monotonic but unknown. The contrasts are optimal in the sense that they maximize the minimum power of the test over all possible monotonic treatment effect patterns. (The minimum power of a test occurs when the real pattern is as far as possible from that of the contrast.)

Abelson and Tukey also suggest the use of a slightly modified version of their contrast, as an alternative that is simpler than the original and almost as efficient. In this contrast the coefficients are the same as those for a linear trend, except that the coefficients for the highest and lowest treatment are doubled. For example for a set of 4 ordered treatments the contrast for the linear trend test is (-3, -1, 1, 3), and that for the modified version of the Abelson-Tukey test is (-6, -1, 1, 6). They refer to this as the 'linear-2' contrast.

The standard Abelson-Tukey test would only be applicable to AR data sets with equal numbers of pens per treatment and constant variance of the pen means. However the 'linear-2' variation of it appears to be applicable to any data set for which the standard trend test is applicable. To derive the 'linear-2' contrast for a given situation, identify the appropriate trend test and double the coefficients for the control and the highest dose level.

A.2.2 Methods for Qualitative Data

A.2.2.1 The Chi-Square Test (to compare 2 treatments)

The chi-square test is the standard test to compare treatments for data in contingency tables, and is described in any statistics text. It is applicable to data sets for proportional variables in which the data have been reduced to a single proportion per treatment. The data is then in the form of a 2 X N contingency table, e.g.

	Control	Dose 1	Dose 2	Dose 3
No. of Eggs That Hatched	x	x	x	x
No. of Eggs That Did Not Hatch	x	x	x	x
Total (No. of Eggs Set)	x	x	x	x

To compare two treatments, such as control and a particular dose level, first select the appropriate 2 X 2 subset of the table. The subset to compare control and dose level 3 for the above example is

	Control	Dose 3	Total
No. of Eggs That Hatched	x	x	R_1
No. of Eggs That Did Not Hatch	x	x	R_2
Total (No. of Eggs Set)	C_1	C_2	T

A table of expected values is obtained from the row and column totals. The expected value E_{ij} for the i'th row and j'th column is

$$E_{ij} = R_i C_j / T$$

The test statistic χ^2 is then calculated from the squares of the deviations of the E_{ij} from the original data values O_{ij}:

$$\chi^2 = \Sigma_i \Sigma_j (O_{ij} - E_{ij})^2 / E_{ij}$$

χ^2 has a chi-square distribution with 1 degree of freedom if certain conditions are satisfied. Its critical values are in standard tables. The critical value for a one-tailed test at the 5% level is 2.71.

One condition to be met is that the O_{ij} for any row must represent the results of independent trials. (This condition would not hold if the table were produced by simply pooling data from all pens within a treatment.) The other is that the E_{ij} must not be too small. The usual restriction is that the E_{ij} should all be 5 or greater, although some texts suggest that it is permissible to have some values of E_{ij} that are under 5.

Alternatively the test statistic z_{2T} (the square root of χ_{2T}^2) may be used, which follows a standard normal distribution. A positive or negative sign is assigned to z_{2T} depending on whether the per cent for the dose level being tested is greater than or less than the per cent for control. The critical value for a one-tailed test at the 5% level is 1.645.

One point at issue is the question of whether or not to make a 'correction for continuity' as suggested in some texts. To make this correction, replace each $(O_{ij}-E_{ij})$ in χ^2 by $(|O_{ij}-E_{ij}|-0.5)$. If $|O_{ij}-E_{ij}|$ is less than 0.5, then replace $(O_{ij}-E_{ij})$ by 0. This reduces the value of the test statistic considerably if the data values are small. Some authors (e.g. (Miller, 1986) or (Conover, 1974)) advise against this correction and claim that the test is then too conservative.

A.2.2.2 Fisher's Exact Test

This test, discussed in most statistics texts, is applicable to 2 x 2 tables and is the test generally recommended when the expected values are too small for the chi-square test to be applied. The first step in the test is to identify all possible 2 x 2 tables that have the same row and column totals as the original data. For the data set

	Control	Dose 3	Total
No. of Eggs That Hatched	8	2	10
No. of Eggs That Did Not Hatch	3	4	7
Total (No. of Eggs Set)	11	6	17

the set of all possible tables with the same totals is

10	0	9	1	8	2	7	3	6	4	5	5	4	6
1	6	2	5	3	4	4	3	5	2	6	1	7	0

These tables have been arranged in the order of their treatment effect, the leftmost table having the maximum negative effect of the dose level and the rightmost table the maximum positive effect.

The probability is calculated of obtaining each table in the series, under the assumption that there is no difference between the treatments (assuming fixed row and column totals). Let a, b, c and d be the data values in a table, R_i and C_j be the row and column totals, and T be the overall total:

a	b	R_1
c	d	R_2
C_1	C_2	T

The probability P of obtaining a given table by chance is

$$P = \frac{R_1!\ R_2!\ C_1!\ C_2!}{a!\ b!\ c!\ d!\ T!}$$

For our example, these probabilities are

	10	0	9	1	8	2	7	3	6	4	5	5	4	6
	1	6	2	5	3	4	4	3	5	2	6	1	7	0
Prob:	.0006		.0170		.1273		.3394		.3563		.1425		.0170	

To test whether there is a significant negative effect for this dose level using a one-tailed test, we sum the probabilities of the actual data set plus the other sets in the same tail:

Sum of probabilities = .0006 + .0170 + .1273 = .1449

Since this is greater than .05, the test is not significant at the 5% level.

In general Fisher's exact test gives results that are very similar to those obtained for the chi-square test with the continuity correction (described in section A.2.2.1). Like this latter test, it has also been criticized as being too conservative (e.g. (Kempthorne, 1979), (Upton, 1982) and (Rice, 1988)).

A.2.2.3 The Cochran-Armitage Test

This test, described in Cochran (1954) and Armitage (1955), tests for linear trend in a set of ordered treatments within a contingency table. Consider the following table:

	Control	Dose 1	Dose 2	Dose 3
No. of Eggs That Hatched	x	x	x	x
No. of Eggs That Did Not Hatch	x	x	x	x
Total (No. of Eggs Set)	x	x	x	x

Let P_i be the proportion and N_i be the total for each treatment. For our example, P_i = (Eggs That Hatched)/(Eggs That Did Not Hatch) and N_i = (Eggs Set).

In order to carry out this test it is necessary to define a treatment scale and to assign each treatment a value U_i on this scale. For example the control could be assigned a value of $U_1 = 1$ and the dose levels assigned values of $U_i = i$ for $i = 2, 3$ and 4. The test involves the calculation and testing of the trend in P_i as U_i increases. The formula for the trend coefficient b_{AC} is

$$b_{AC} = [\Sigma N_i(P_i-\bar{P}_W)(U_i-\bar{U}_W)]/[\Sigma N_i(U_i-\bar{U}_W)^2]$$

where $\bar{P}_W$ and $\bar{U}_W$ are weighted means of the P_i and U_i respectively, weighted by N_i:

$$\bar{P}_W = \Sigma N_i P_i/\Sigma N_i \qquad \text{and} \qquad \bar{U}_W = \Sigma N_i U_i/\Sigma N_i$$

The test statistic is $\chi_{AC}^2 = b_{AC}^2[\Sigma N_i(U_i-\bar{U}_W)^2]/[\bar{P}_W(1-\bar{P}_W)]$

which has a chi-square distribution with 1 degree of freedom. The critical value for a one-tailed test at the 5% level is 2.71.

Alternatively, the test statistic z_{AC} (the square root of χ_{AC}^2) may be used, which has the standard normal distribution. A positive sign is assigned to z_{AC} if b_{AC} is positive and a negative sign if b_{AC} is negative. The critical value for a one-tailed test at the 5% level is 1.645.

The formula for b_{AC} is actually the same as that for a trend coefficient in a weighted linear regression of P_i on U_i with weights of N_i, but its variance differs somewhat from the corresponding regression variance because it is calculated from the assumption that the P_i are binomially distributed rather than from the residual mean square of the regression. This method is also mentioned in Snedecor and Cochran (1967), p. 246.

A.2.3 Non-Parametric Methods

A.2.3.1 The Mann-Whitney Test

This test (and the Wilcoxon rank test which is equivalent to it) is the rank-based counterpart to the t-test. It is widely employed and is described in most statistics texts. To compare two treatments using this test, assign each data value its rank within the combined data set.

For example consider the data set

	Treatment 1					Treatment 2				
	.85	.45	.60	.25	.78	.27	.10	.58	.49	.07
with ranks	10	5	8	3	9	4	2	7	6	1

The sum of the ranks is then calculated for each treatment. The test statistic is the smaller of the two sums. The significance of the treatment effect is obtained from a special table of critical values. In situations with well-behaved data this method is less powerful than the t-test which is its least-squares counterpart. However it is more robust and less affected by data irregularities.

A factor to consider in using the Mann-Whitney test is that it compares the mean ranks of treatments rather than treatment means. In doing so it makes the assumption that the distribution of the data values is the same within each treatment, so that if the mean ranks are different the means will be different also. This may be a problem for AR data sets where the distributions are not the same for all treatments, such as those where the within-treatment variance is larger in one treatment than another.

Another potential problem is that of tied ranks. When a group of data values are equal, they are each assigned the average rank for the group. In data sets where the data are categorical or discrete in nature, it is possible to have a large number of tied ranks and this can result in a test statistic that is erratic and non-normally distributed ((Lehman, 1961) or (Klotz, 1966)). Recent developments in statistical computing have made it feasible to overcome this problem by generating the exact distribution of the test statistic (Mehta et al, 1984), although this adds considerably to the amount of work required.

A number of papers have compared the performance of the Mann-Whitney test with that of the standard and the weighted t-tests. The data sets employed in these comparisons were similar to those that occur in AR experiments in that the data values were proportions with varying denominators. Among these papers are Haseman and Soares (1976), Gladen (1979), and Shirley and Hickling (1981)). They found that there was some loss of efficiency for the Mann-Whitney test relative to both the weighted and unweighted t-test, although the loss was relatively small. This loss was attributed to the fact that the Mann-Whitney test did not take into account the differences in variance among the proportions. The Mann-Whitney test is also described in general references on rank-based methods such as Conover (1971), and Van der Laan and Verdooren (1987).

A.2.3.2 The Jackknife Method

This is another method for dealing with data sets that have complications such as irregular distributions or data values of varying but unknown accuracies. It is relatively simple to apply as it does not require the use of complex models, but is generally not described in standard texts. A good reference for this method is Miller (1974).

The following is the procedure to estimate a treatment proportion from a set of pen proportions using the jackknife method (the procedure would be the same to estimate a treatment mean from a set of pen means). Let (P_j) be a set of pen proportions for a particular treatment, with each P_j being derived from

$$P_j = x_j/n_j$$

For example n_j could be the number of eggs set in a pen and x_j could be the number of these that hatch. The first step is to obtain a preliminary estimate P_{PL} of the treatment proportion by simply pooling the proportions:

$$P_{PL} = \Sigma x_j/\Sigma n_j = \Sigma n_j P_j/\Sigma n_j$$

However, for reasons discussed earlier in this report, P_{PL} may not be a satisfactory estimate as it weights each P_j by n_j. The jackknife method derives an improved estimate P_{JK} of the treatment proportion that is more efficient and less biased than P_{PL}, and also provides a variance estimate for P_{JK} that takes pen-to-pen variation into account.

The next step in the jackknife test is to calculate for each pen the pooled proportion that would be obtained if the proportion for that pen were removed from the data set. Denote by P_{-j} the pooled proportion with P_j removed:

$$P_{-j} = \Sigma x_k / \Sigma n_k$$

where k takes on all values except j. Then for each pen calculate the pseudo-value R_j:

$$R_j = P_{PL} + (N-1)(P_{PL}-P_{-j})$$

where N is the number of pens. The jackknife estimate P_{JK} of the treatment proportion is derived from the R_j, as is its variance:

$$P_{JK} = \bar{R} = \Sigma R_j / N$$

and $$\text{Var}(P_{JK}) = \Sigma (R_j - \bar{R})^2 / [N(N-1)]$$

P_{JK} is in fact a weighted mean of the P_j, with the weights being a complex function of N and the n_j. The result is to give more importance to the data values with larger n_j.

The jackknife method has been employed on a test basis for analysing proportions by Gladen (1979) and by Crump and Howe (1988). They found it to perform about as efficiently as the standard t-test or the Mann-Whitney test. However, one problem is that in some cases the weights are not that stable and the value of P_{JK} can be somewhat erratic. For example under extreme circumstances the value obtained for P_{JK} could be outside the range of the pen means P_j.

If applied to AR data sets, the jackknife method would only be the first step in the analysis and would provide estimates of treatment means and their associated variances. It would then be necessary to carry out statistical tests on these means, such as a sequence of t-tests or trend tests.

A.2.3.3 The Rerandomization Method

Another non-parametric method is the rerandomization method, also referred to as the randomization method or the permutation method. In this method the significance of treatment effects is tested by means of new data sets that are generated by carrying out permutations of the original data values. In these new data sets the data values are the same as in the original set but are assigned to treatments by random allocation.

To illustrate this permutation process, consider a simplified data set with treatments C, D1 and D2, and two data values per treatment (for example these values could be pen proportions):

C	D1	D2
.30	.25	.15
.50	.35	.20

The following 5 sets can be generated by reassigning the data values for D1 and D2 to different treatments while retaining the original values for treatment C:

C	D1	D2	C	D1	D2	C	D1	D2	C	D1	D2	C	D1	D2
.30	.20	.15	.30	.15	.20	.30	.20	.15	.30	.15	.20	.30	.15	.25
.50	.35	.25	.50	.35	.25	.50	.25	.35	.50	.25	.35	.50	.20	.35

If the values for treatment C were included in the reassignment process, a total of 89 permuted data sets could be produced. The number of possible permutations increases exponentially with the size and complexity of the data set, so that for a typical AR data set the number would be astronomically large.

To carry out a statistical test, the appropriate test statistic is calculated for the original data set and for each of the generated sets (e.g. the statistic might be the difference between the D2 and C means). If the value of the statistic for the original set is within the most extreme 5% of the set of statistic values (from the original plus the permuted sets), the test result is considered to be significant at the 5% level. No tables of critical values are required.

The rationale for this conclusion is that original data set plus the permuted sets constitute a workable approximation to the set of all possible outcomes of the experiment under the assumption that there is no treatment effect. Thus the distribution of the values of the test statistic obtained from these sets approximates the actual distribution of the test statistic.

Although it is very calculation-intensive, the rerandomization method should be more robust that least squares methods in dealing with irregular distributions or in handling situations where the data values are of varying but unknown accuracy (Edgington, 1987). The reasoning is that with the rerandomization approach it is not the actual value of a test statistic that is important, but its value relative to the values for other permuted sets. If an outlier tends to affect the value for all permutations in the same direction, the effect on the relative value of the test statistic would not be that large.

It may be necessary to select a simpler test statistic than would be the case for least squares analysis. For example if the treatment means are of varying accuracy, the optimum test statistic might be a difference of weighted means. But since no linear model is developed, it would probably be necessary to employ a simple statistic such as a difference of unweighted means.

The concept of rerandomization is well established in specific situations. Fisher's exact test involves a permutation of contingency table data, for example. Rank-based methods are also based on the principle of permutation of data values, as is the jackknife method. However only recently has rerandomization become feasible for moderate or large data sets because of the computing power needed to generate the permuted sets. For large data sets, where the generation of all permuted sets is still an intractable problem, a rerandomization test can still be carried out by generating a large random sample of permuted sets and treating it as a workable approximation to the complete set ((Edgington, 1980), (Miller, 1986) or (Crump and Howe, 1988)). Some decision is required as to the number of permuted sets to generate.

The rerandomization approach appears to be receiving more attention because its use is facilitated by the increase in computing power generally available. For example Petrondas and Gabriel (1983) have employed it for a multiple comparison test using a multi-stage procedure. It has recently been applied to proportional data from a teratogenicity experiment, with encouraging results (Crump and Howe, 1988). For dealing with difficult data situations the power of the rerandomization method appears to be as good as or better than least-squares methods and rank-based methods.

However, the suitability of the rerandomization method for regulatory purposes has been questioned in some papers. One perceived drawback is that it is not possible to calculate exact confidence levels (Haseman and Kupper, 1979). Another is that the calculations are too complex for routine use (Shirley and Hickling, 1981).

A.3 Potential Methods

A.3.1 Least-Squares Methods

A.3.1.1 Bartholomew's Test

This test, described in Bartholomew (1959), is a least-squares test applicable to experiments with ordered treatments. It is similar to a trend test, except that it tests for any monotonic pattern of treatment effects rather than for one specific pattern.

Consider an experiment that consists of a control C and increasing dose levels D1, D2 and D3, in which a test is to be made for a monotonic decrease in the underlying treatment means μ_C, μ_{D1}, μ_{D2} and μ_{D3} as the dose level increases. Bartholomew's test tests hypothesis H_1 against the null hypothesis H_0.

H_1: $\mu_C \geq \mu_{D1} \geq \mu_{D2} \geq \mu_{D3}$ with $\mu_C > \mu_{D3}$

H_0: $\mu_C = \mu_{D1} = \mu_{D2} = \mu_{D3}$

The test involves the application of an averaging process to remove any inconsistency in the series of mean values. This process is the same as that employed for Williams test in which pairs of consecutive treatment means are replaced by their average (as illustrated in section A.2.1.3).

In this example, let the treatment means obtained from the data be M_C, M_{D1}, M_{D2} and M_{D3}. If the means are in the expected pattern of $M_C \geq M_{D1} \geq M_{D2} \geq M_{D3}$ no adjustment is necessary. But if there is a deviation from this pattern, for example if $M_{D2} < M_{D3}$, then both M_{D2} and M_{D3} would be replaced by their average. This averaging is repeated if necessary, until a series of adjusted means M_i' is obtained such that $M_C' \geq M_{D1}' \geq M_{D2}' \geq M_{D3}'$.

Assuming that the treatment means are of equal accuracy with common variance S_M^2, the test statistic is

$$\chi^2 = \Sigma[(M_i' - \bar{M})/S_M]^2$$

where $\bar{M}$ is the average of the treatment means M_i (and also of the M_i').

If the treatment means are of varying accuracy, the test can still be applied using weighted analysis. The averaging process to produce a monotonic series of treatment means M_i' involves the weighted averaging of consecutive treatment means. The weights are $1/S_{Mi}^2$, where S_{Mi}^2 is the variance of M_i. The test statistic is

$$\chi^2 = \Sigma[(M_i'-\bar{M}_w)/S_{Mi}]^2$$

where $\bar{M}_w$ is the weighted mean of the M_i using weights of $1/S_{Mi}^2$.

Unfortunately, χ^2 has a complex distribution that is a mixture of several different chi-square distributions, and special calculations are required to obtain critical values. This complication is a result of the possibility of averaging to obtain a monotonic series of means. For this reason it is probably not suitable for application to AR data sets.

A.3.1.2 Pattern-Specific Tests

These tests, described in Ruberg (1989), are interesting because they are designed specifically for experiments with ordered treatments, and involve linear contrasts that are tailored to specific patterns of treatment effects. The treatments are assumed to be equally spaced in some scale. Unfortunately these tests require the treatment means to have equal accuracy, and so are probably not applicable to AR data sets in their present form.

A.3.1.2.1 Step Contrasts

Step contrasts test for a sudden change in response at one of the dose levels. Each contrast compares the mean of all dose levels that are at or above a particular value with the mean of all levels below that value. Consider an experiment with a control C and dose levels D1, D2 and D3, in which the treatment means are expected to decrease as dose level increases. Let the treatment means be M_C, M_{D1}, M_{D2} and M_{D3}. The step contrast $(3M_C - M_{D1} - M_{D2} - M_{D3})$ tests for a sudden drop in the mean between control and D1, the contrast $(M_C + M_{D1} - M_{D2} - M_{D3})$ tests for a drop between D1 and D2, and $(M_C + M_{D1} + M_{D2} - 3M_{D3})$ tests for a drop between D2 and D3.

In the test procedure, all step contrasts are tested together. For each contrast F, a variance S_F^2 is derived from the within-treatment variances and a t-value is calculated:

$$t = F/S_F$$

The highest t-value is identified and tested using a special table of critical values. If it is significant, it is concluded that a change in treatment effect occurs at that point in the series of treatments. The NOEC is also identified. For example if the contrast that has the highest t-value for our experiment is ($M_C + M_{D1} - M_{D2} - M_{D3}$) and the test is significant, it is concluded that the treatment effect changes between D1 and D2 and that the NOEC is at level D1.

The method is most useful for cases where a sudden response occurs at some dose level but it is not known where this threshold occurs. It has the advantage that the NOEC is identified by a single test rather than by a series of tests. The disadvantage is that the test is inefficient at identifying treatment effects that increase linearly as the dose level increases.

A.3.1.2.2 Basin Contrasts

These test for a response pattern in which there is no response to the treatments up to a certain level and a linear response thereafter as the level increases. For example, consider an experiment where the treatments are a control C and a set of increasing dose levels D1, D2 and D3, with treatment means expected to decrease as the dose level increases. Let the treatment means be M_C, M_{D1}, M_{D2} and M_{D3}. The contrast ($3M_C + M_{D1} - M_{D2} - 3M_{D3}$) tests for a linear decrease starting at C, ($3M_C + 3M_{D1} - M_{D2} - 5M_{D3}$) tests for a linear decrease starting at D1, and ($M_C + M_{D1} + M_{D2} - 3M_{D3}$) tests for a decrease starting at D2.

In the test procedure, all basin contrasts are tested together. For each contrast F, a variance S_F^2 is derived from the within-treatment variances and a t-value is calculated:

$$t = F/S_F$$

The highest t-value is identified and tested using a special table of critical values. If it is significant, it is concluded that a change in treatment effect occurs at that point in the series of treatments. The NOEC is also identified. For example if the contrast that has the highest t-value in our experiment is ($3M_C + 3M_{D1} - M_{D2} - 5M_{D3}$) and the test is significant, it is concluded that there is a linear treatment response starting at D1 and that D1 is the NOEC.

The method is designed for cases where there is no response up to a some threshold level and a linear response thereafter. It has the advantage that the NOEC is identified by a single test rather than series of tests. Some disadvantages are that it is difficult to identify the correct level at which the linear response starts because the three contrasts are very similar, and that it is inefficient at identifying treatment effects that involve sudden changes at certain dose levels.

A.3.1.2.3 Helmert Contrasts

This third set of contrasts are similar to step contrasts in that they also test for a sudden jump at a particular dose level. They compare a particular treatment mean with the average of the means for all treatments that are lower in the order. For example, consider an experiment where the treatments are a control C and a set of increasing dose levels D1, D2 and D3, with treatment means expected to decrease as dose level increases, and let the treatment means be M_C, M_{D1}, M_{D2} and M_{D3}. The first Helmert contrast is $(M_C - M_{D1})$ and tests for drop in the mean between control and D1, the second $(M_C + M_{D1} - 2M_{D2})$ tests for a drop between D1 and D2, and the third $(M_C + M_{D1} + M_{D2} - 3M_{D3})$ tests for a drop between D2 and D3.

Unlike step and basin contrasts, Helmert contrasts are tested sequentially starting with the lowest dose. For each contrast F, a variance S_F^2 is derived from the within-treatment variances and a t-value is calculated:

$$t = F/S_F$$

For our experiment, the first test would be of D1 against control. If the result is significant, it is concluded that there is a jump in response between C and D1 and that C is the NOEC. If it is not significant, the second contrast is then tested to compare D2 to the average of C and D1 and so on. Because the contrasts are orthogonal, the tests of the different contrasts are independent of each other. Critical values for the tests can be calculated from the maximum modulus distribution. These values are set out in Hochberg and Tamhane (1987).

The method is efficient for cases where a sudden response occurs at some threshold level but there is minimal change below that level. The disadvantage is that the test is inefficient at identifying treatment effects that increase linearly as the dose level increases.

A.3.2 Methods for Qualitative Data

A.3.2.1 Complex-Model Methods for Proportions

Methods of analysis have been developed specifically for proportional variables, based on models that are more complex than the standard models. These models have the same basic form as those employed in section 5.5.5. In the notation of that section, the variance σ_{Pij}^2 of a proportion P_{ij} is the sum of a binomial and an extra-binomial component:

$$\sigma_{Pij}^2 = P_{ij}(1-P_{ij})/n_{ij} + \sigma_{ebij}^2$$

However a more complex model is employed for the extra-binomial component σ_{ebij}^2 with these methods. The methods are theoretically interesting but are probably not robust enough to deal with the data irregularities that can occur with data from AR experiments.

The Beta Binomial Model

This is a generalization of the binomial model for proportional variables, developed by Williams (Williams, 1975), which models σ_{ebij}^2 by assuming that the extra-binomial term follows a beta distribution. A number of papers have compared the beta binomial method with other methods using Monte Carlo techniques, with mixed results. Although it fits some data sets quite well (Crowder, 1977), it is sensitive to departure from its assumptions ((Paul, 1982), (Haseman and Soares, 1976), and (Shirley and Hickling, 1981)). A study by Pack (1981) found that it did not provide much improvement over simpler techniques such as t-tests.

Other Complex Models

A number of other models have been suggested for proportional variables, each making different assumptions about the extra-binomial term. One of these is the 'correlated binomial' model (Kupper and Haseman, 1978), and another is the 'multiplicative binomial' model (Altham, 1978). Other proposed models are mentioned in Haseman and Kupper (1979). These methods have not received as much attention in the literature as the beta binomial model, but indications are that they have essentially the same problems ((Paul,1982) and (Crump and Howe, 1988)).

A.3.3 Non-Parametric Methods

A.3.3.1 Jonckheere's Test

Described in Jonckheere (1954) and also in Van Der Laan and Verdooren (1987), this test is applicable to experiments with ordered treatments. It is similar to Bartholomew's test in that it tests for the presence of any monotonic pattern of treatment effects rather than for a specific pattern such as a linear trend.

Let (T_i) be a set of ordered treatments. Suppose that it is expected that the data values means will decrease as the treatment number i increases. For any pair (i,j) of treatments, let i be less than j. Thus we expect the data values (X_{jl}) in T_j to be less than the values (X_{ik}) in T_i.

To carry out Jonckheere's test, for each treatment pair (i,j) tabulate the number of data value pairs (X_{ik},X_{jl}) such that $X_{jl} > X_{ik}$ (the opposite of what is expected), and denote this number by N_{ij}. The test statistic is $\sum N_{ij}$, with the summation taken over all pairs of treatments. Tables of critical values for $\sum N_{ij}$ are given in Hollander and Wolfe (1973).

Jonckeere's test is employed from time to time in biological experiments (e.g. Hewett and Bair, 1986), but was not considered well enough established to be included among the recommended methods.

A.3.3.2 Shirley's Test

This rank-based counterpart of the Williams test has been developed by Shirley (1977), and represents an extension of the Mann-Whitney test in the same way that the Williams test represents an extension of the t-test. The objective is to provide a rank-based test that removes the possibility of inconsistent results (this possibility is discussed in Appendix A section A.2.1.3, and in Appendix B). Williams has commented on this test and recommended minor changes (Williams, 1986). Although it is promising in principle, it is considered to be too untried to be included among the recommended methods.

A.4 Methods Not Recommended

These methods are in current use, but do not meet the criteria for efficiency set out in section 4.6. The reason is that they attempt to detect all patterns of treatment effect, and are not efficient when applied to data from experiments such as AR experiments in which the objective is to detect a specific pattern (an increasing negative effect on reproduction as the dose level increases).

A.4.1 Least Squares Methods

A.4.1.1 One-way ANOVA

This is the standard method for testing for treatment effects in general, and is discussed in all statistics texts. It compares all treatment means simultaneously using an F-test. The standard ANOVA procedure requires that the within-treatment variance be constant over all treatments. However a weighted ANOVA procedure exists that is applicable to cases where the variance differs from one treatment to another, and is described in Scheffé (1959).

A.4.1.2 General Multiple Comparison Procedures

These are procedures to compare each treatment mean with each other mean. Each of these pairwise comparisons employs a test statistic that is similar to that for a t-test or LSD test. Test statistics are of the form

$$t = D/S_D$$

where D is the difference between two means and S_D^2 is the variance of D.

However the critical values used in general multiple comparison procedures are more conservative than those in single tests, in order that the experiment-wide error rate be equal to 5% (or some other specified value). The critical values are such that if no treatment effects are present, the probability of even one of the pairwise mean comparisons being wrongly declared significant is equal to the specified level. The larger the number of pairwise mean comparisons, the more conservative the critical values must be. The critical values are in special tables which are given in most statistics texts.

The multiple comparison methods currently employed for AR experiments include

- Tukey's test
- Student-Newman-Keuls (SNK) test
- Duncan's multiple range test

The level of conservativeness differs from one method to another, with Tukey's test being the most conservative, then the SNK test, then Duncan's test. (For AR experiments even Duncan's test is too conservative, however.)

Tukey's Test

This is the most widely accepted multiple comparison test for situations where there is a need to be able to compare any pair of treatments. The error rate holds even when the largest and smallest treatment means in the experiment are selected after the fact and compared with each other. Tukey's test requires equal numbers of data values per treatment and equal within-treatment variances. However it has been extended to the case of unequal numbers per treatment by Kramer (1956) and to the case of unequal within-treatment variances by Games and Howell (1976).

The SNK Test

This is a modified, sequentially-applied version of Tukey's test in which the critical values are reduced when comparing means that are close together in rank Thus the test is more liberal for these comparisons. To compare a set of 7 treatment means using the SNK test, for example, the means are first ranked. Tukey's test is then applied to compare the 1st and 7th ranked means. If they are significantly different, Tukey's test is then used to compare the 1st and 6th ranked means and the 2nd and 7th. But for these latter comparisons, the critical value is that for a 6-treatment experiment (while for Tukey's test the 7-treatment value would be used throughout). If the 1st and 6th means are significantly different, the 1st is compared with the 5th and the 2nd with the 6th using the critical value for 5 treatments and so on.

Duncan's Multiple Range Test

This test is applied sequentially in the same manner as the SNK test, but is more liberal and has still lower critical values (Duncan, 1955). These values are based on Duncan's 'special protection levels' rather than on a true experiment-wide error rate. Duncan's test is generally considered to be too liberal in the statistical literature, but it is quite widely employed.

A.4.1.3 Dunnett's Test

This is a multiple comparison method designed for experiments in which a control is to be compared to a number of other treatments (Dunnett, 1955). Treatment means are compared on a pairwise basis. Each non-control mean is compared to the control mean, but non-control means are not compared to each other. As with the general multiple comparison procedures, the test statistic is

$$t = D/S_D$$

where D is the difference between the two means and S_D^2 is the variance of D.

Critical values are set so that if there are no treatment effects, the probability of a significant result in any one of these control-non-control comparisons is equal to the specified confidence level. Special tables are needed for these critical values. The number of data values per treatment must be equal and the within-treatment variance must be constant.

While Dunnett's test involves fewer comparisons than the general multiple comparison tests and is therefore less conservative than them, it is still too conservative for experiments such as AR experiments where the objective is to test for a single pattern of treatment effects.

A.4.2 Methods for Qualitative Data

A.4.2.1 General Chi-Square Test

If a data set for an AR experiment is reduced to a single proportion per treatment, it is in the form of a 2 X N contingency table. The form of such a data set for an experiment with a control C and dose levels D1, D2 and D3 would be:

	C	D1	D2	D3
Eggs that Hatch	x	x	x	x
Eggs that Do Not Hatch	x	x	x	x

The chi-square test is the standard general test for treatment effects in a contingency table and is covered in all standard texts. It is currently common practice for AR experiments to apply this general test. But the chi-square test tests for any pattern of treatment effects, and is too conservative for AR experiments where the objective is to test for one specific pattern.

However, the situation is more complicated than that. It is not uncommon for these contingency-table data sets to have been formed by a simple pooling of all the pen proportions within each treatment. But applying the chi-square test to such data sets is not statistically valid, as this would ignore the possibility of real differences between pens within a treatment (as discussed in section 3.6). The result is that the treatment proportions so formed are not as accurate as the sample size would suggest, so that if a chi-square test is applied to these data sets the test is more liberal than the confidence level of 5% suggests.

Thus there are two factors to consider, one causing the test to be too conservative and the other causing it to be too liberal. It is not clear to what extent these factors would cancel each other out in a given case.

A.4.3 Non-Parametric Methods

A.4.3.1 The Kruskal-Wallis Test

This test is the rank-based counterpart of the one-way ANOVA, and is described in Kruskal and Wallis (1952) and in a number of texts on non-parametric methods (e.g. Conover, 1971). An overall rank is assigned to each data value, and the test statistic is based on the total of the ranks for each treatment. It is too conservative for AR experiments, for the same reason as the one-way ANOVA.

APPENDIX B - AUXILIARY STATISTICAL PROCEDURES

Table of Contents

B.1 Combining Pen Proportions by Rao and Scott's Method

B.1.1 Combining Actual Proportions

Let P_{ij} be the pen proportion for pen j in treatment i. Each P_{ij} is of the form

$$P_{ij} = y_{ij}/n_{ij}$$

where n_{ij} is the number of subjects per pen for the variable being analysed. Consider the problem of how to combine all the P_{ij} within a treatment into a single proportion.

The most obvious procedure for combining the P_{ij} within a treatment is to pool them. The pooled treatment proportion PP_i for treatment i is given by

$$PP_i = \Sigma_j y_{ij}/\Sigma_j n_{ij} = Y_i/N_i$$

Although PP_i is a valid estimate of the overall proportion for treatment i (though not necessarily an efficient estimate), it is not statistically valid to analyse it as if it were a simple proportion of Y_i successes in N_i trials (as discussed in section 3.6). Its variance would be underestimated as it is not as accurate as a proportion obtained from N_i independent trials.

However Rao and Scott have derived a method that overcomes this underestimation of the variance (Rao and Scott, 1992), based on the fact that PP_i is a weighted mean of the P_{ij} with weights of n_{ij}:

$$PP_i = \Sigma_j n_{ij} P_{ij}/\Sigma_j n_{ij}$$

Consequently an unbiased estimate V_i of the variance of PP_i can be obtained from the variation among the P_{ij} within treatment i. The formula for this is:

$$V_i = (m_i/(m_i-1))\ (1/N_i)^2\ \Sigma_j n_{ij}^2 (P_{ij}-PP_i)^2 \qquad (C1)$$

where m_i is the number of pens in treatment i.

Rao and Scott make use of this variance to obtain an 'effective denominator' $(N_i)_{eff}$ for PP_i, which is defined as the value that it gives the correct variance for PP_i if entered as the denominator in the binomial variance formula.

Since the binomial variance for a proportion P with denominator N is P(1-P)/N, the value of $(N_i)_{eff}$ is obtained from V_i and PP_i by solving the equation

$$V_i = PP_i(1-PP_i)/(N_i)_{eff}$$

$(N_i)_{eff}$ is in general smaller than N_i, since the variance estimate V_i obtained using equation (C1) is greater than the binomial variance estimate that would be obtained if PP_i were a simple proportion with denominator N_i.

Once $(N_i)_{eff}$ is obtained, an effective numerator $(X_i)_{eff}$ is calculated for PP_i by defining it as

$$(X_i)_{eff} = PP_i(N_i)_{eff}$$

This allows PP_i to be expressed as

$$PP_i = (X_i)_{eff}/(N_i)_{eff}$$

Since its denominator now corresponds to its variance, PP_i can be entered into statistical formulas as if it were the simple proportion $(X_i)_{eff}/(N_i)_{eff}$ according to Rao and Scott.

Note: In general $(X_i)_{eff}$ and $(N_i)_{eff}$ are not integers. This could restrict the methods employed to analyse the PP_i (for example it would appear that Fisher's exact test would not be applicable).

It may appear that this procedure avoids the issue of the relative size of the binomial and extra-binomial components for the variance of P_{ij} (discussed in section 5.5.5). However this is not the case. The relative size of these components determines the best weighting scheme by which to weight the P_{ij} in calculating treatment proportions. By pooling the P_{ij} into PP_i, and thus employing a weighting scheme with weights equal to n_{ij}, Rao and Scott are making the implicit assumption that the extra-binomial component is small relative to the binomial component. But to be flexible enough for general use, the method must be able to accommodate data sets where the extra-binomial component is moderate or large.

This makes it advisable to consider a range of possible weighting schemes. In principle it should be possible to extend Rao and Scott's method to any weighting scheme. Let w_{ij} be the weight assigned to P_{ij} in a general weighting scheme. Then the weighted mean $\bar{P}_{Wi}$ for treatment i for this scheme is

$$\bar{P}_{Wi} = \Sigma_j w_{ij} P_{ij} / \Sigma_j w_{ij}$$

and its unbiased variance estimate V_{Wi} is

$$V_{Wi} = (m_i/(m_i-1))\ (1/\Sigma_j w_{ij})^2\ \Sigma_j w_{ij}^2 (P_{ij}-\bar{P}_{Wi})^2$$

The effective numerator and denominator for $\bar{P}_{Wi}$ could then calculated from V_{Wi} in the same manner as for PP_i.

A possible approach to the question of weighting schemes would be to select one of three simple schemes:

1. $w_{ij} = n_{ij}$ 2. $w_{ij} = n_{ij}^{0.5}$ 3. $w_{ij} = 1$

These would accommodate data sets where the extrabinomial variance component is small, moderate or large, respectively, compared to the binomial component. The procedure proposed by Cochran (1943) could be employed to determine the relative size of the two components. (Cochran's method is discussed in section B.2.2.1.)

B.1.2 Combining Estimated Proportions

Let P_{ij} be an estimated proportion for pen j in treatment i, that is the product of two simple proportions. Then P_{ij} is of the form

$$P_{ij} = Q_{ij}\ R_{ij} = (a_{ij}/b_{ij})\ (c_{ij}/d_{ij})$$

Since there is no single denominator, it is not clear how Rao and Scott's method could be applied to combine the P_{ij} in each treatment into a single proportion. A possible way of proceeding (that has not been examined for validity) would be to designate either b_{ij} or d_{ij} as the denominator, whichever has the larger values. Suppose the b_{ij} are larger. Then weighting schemes corresponding to those of section B.1.1 could be drawn up that are based on b_{ij} instead of n_{ij}. This in effect treats P_{ij} as the proportion

$$P_{ij} = x_{ij}/b_{ij} \qquad \text{where} \qquad x_{ij} = a_{ij}c_{ij}/d_{ij}$$

B.2 Derivation of Weighting Schemes

B.2.1 A Weighting Scheme for Pen Means

To obtain a weighting scheme for the pen means of a measurement variable, it is necessary to model and estimate their variance. Consider an experiment with M treatments, N_i pens per treatment and n_{ij} measurements per pen. Let X_{ijk} be the k'th measurement within pen j of treatment i. The standard linear model for X_{ijk} is

$$X_{ijk} = \mu + T_i + E_{ij} + e_{ijk}$$

where μ is the true population mean, T_i is the effect of treatment i, E_{ij} is the random effect for pen j and e_{ijk} is the random effect of measurement k. Let σ_{Eij}^2 and σ_{eijk}^2 be the variance of E_{ij} and e_{ijk} respectively. The pen mean $\bar{X}_{ij}$ is then modelled by

$$\bar{X}_{ij} = \mu + T_i + E_{ij} + \bar{e}_{ij}$$

where $\bar{e}_{ij}$ is the mean of the e_{ijk} for that pen. The variance σ_{Xij}^2 of $\bar{X}_{ij}$ is given by

$$\sigma_{Xij}^2 = \sigma_{Eij}^2 + \sigma_{eij}^2/n_{ij}$$

where σ_{eij}^2 is the mean of the σ_{eijk}^2.

Assuming that the variances σ_{Eij}^2 and σ_{eijk}^2 have constant values of σ_{Ei}^2 and σ_{ei}^2 within each treatment, σ_{Xij}^2 is given by

$$\sigma_{Xij}^2 = \sigma_{Ei}^2 + \sigma_{ei}^2/n_{ij} \qquad (C2)$$

To estimate σ_{Xij}^2 for each pen mean, estimates of σ_{Ei}^2 and σ_{ei}^2 must be obtained for each treatment. The approach suggested is to base these estimates on the mean squares produced in a one-way ANOVA for each treatment. If a one-way ANOVA is carried out on the data for treatment i, the ANOVA table would be of the form:

Source	df	Expected Mean Square
Variation Between Pens	N_i-1	$\sigma_{ei}^2 + a_i\sigma_{Ei}^2$
Variation Between Measurements Within Pens	$\Sigma_j n_{ij}-N_i$	σ_{ei}^2
Total	$\Sigma_j n_{ij}-1$	

The parameter a_i is derived from the n_{ij}. If the n_{ij} for treatment i are all equal, a_i is this common value. If they are not all equal, the formula for a_i is

$$a_i = (\Sigma_j n_{ij} - \Sigma_j n_{ij}^2/\Sigma_j n_{ij})/(N_i-1)$$

This formula is given in most texts (e.g. Snedecor and Cochran, 1967, p. 290).

The estimates of σ_{Ei}^2 and σ_{ei}^2 are:

$$\hat{\sigma}_{ei}^2 = MS(Measurements)$$

$$\hat{\sigma}_{Ei}^2 = [MS(Pens) - MS(Measurements)]/a_i$$

σ_{Xij}^2 is estimated for each $\bar{X}_{ij}$ by substituting $\hat{\sigma}_{Ei}^2$ and $\hat{\sigma}_{ei}^2$ into equation (C2).

If there is reason to believe that σ_{Ei}^2 and σ_{ei}^2 may be constant over all treatments, tests of homogeneity of variance can be applied to the estimates $\hat{\sigma}_{Ei}^2$ and $\hat{\sigma}_{ei}^2$ to examine this question. A number of tests for homogeneity are referred to in section B.4. If no significant differences are found, it may be reasonable to assume that σ_{Ei}^2 and σ_{ei}^2 are constant over all treatments and to obtain overall estimates $\hat{\sigma}_E^2$ and $\hat{\sigma}_e^2$. σ_{Xij}^2 is then estimated by substituting $\hat{\sigma}_E^2$ and $\hat{\sigma}_e^2$ into equation (C2).

Two possible procedures for deriving overall estimates are:

- calculate $\hat{\sigma}_E^2$ and $\hat{\sigma}_e^2$ as the simple averages of $\hat{\sigma}_{Ei}^2$ and $\hat{\sigma}_{ei}^2$
- carry out a nested ANOVA over all treatments, and derive $\hat{\sigma}_E^2$ and $\hat{\sigma}_e^2$ from the mean squares for treatment, pen and measurement effects.

The first procedure is the one that is recommended, as the second one involves considerable calculation if the numbers of pens per treatment or of measurements per pen are not constant. Nested ANOVA is described in most standard texts, e.g. Snedecor and Cochran (1967), p. 291.

Once σ_{Xij}^2 is estimated for each pen, its inverse is taken as the weight to apply to $\bar{X}_{ij}$ in a weighted least squares analysis.

B.2.2 Weighting Schemes for Pen Proportions

B.2.2.1 Cochran's Method

Let P_{ij} be the proportion for the j'th pen in treatment i. P_{ij} is equal to y_{ij}/n_{ij} where n_{ij} is the number of subjects per pen for that variable. With Cochran's method (Cochran, 1943) the variance σ_{Pij}^2 of P_{ij} follows the linear model set out in section 5.5.5:

$$\sigma_{Pij}^2 = P_{ij}(1-P_{ij})/n_{ij} + \sigma_{ebij}^2$$

The first component is the binomial sampling error due to the deviation of P_{ij} from the true value for that pen, and the second is the 'extra-binomial' component that is due to pen-to-pen variation in the true pen values within a treatment.

Cochran's method involves the identification of a model for the extra-binomial component σ_{ebij}^2 that will permit a reasonably simple weighting scheme. Cochran considers a number of possibilities for both the form of σ_{ebij}^2 and for its size relative to the binomial component:

	Relative Size of σ_{ebij}^2	Form of σ_{ebij}^2
Case 1	small	n/a
Case 2	moderate	constant σ_{eb}^2
Case 3	moderate	$\lambda P_{ij}(1-P_{ij})$ for constant λ
Case 4	large	constant σ_{eb}^2
Case 5	large	$\lambda P_{ij}(1-P_{ij})$ for constant λ

Cochran sets out a method for estimating the approximate size of σ_{ebij}^2 relative to the binomial component, but does not indicate how to decide on its form. However the form $\lambda P_{ij}(1-P_{ij})$ would appear to be the more likely one. The reason is that σ_{ebij}^2 would be expected to decrease to 0 if P_{ij} increases to 1 or decreases to 0, rather than to remain constant.

Deriving Weights for Case 1 (σ_{ebij}^2 small)

For this case σ_{ebij}^2 can be ignored, and σ_{Pij}^2 is approximately equal to the binomial variance:

$$\sigma_{Pij}^2 = P_{ij}(1-P_{ij})/n_{ij}$$

If the P_{ij} are within a range of roughly .2 to .8, the product $P_{ij}(1-P_{ij})$ is approximately constant and σ_{Pij}^2 is proportional only to $1/n_{ij}$. Since weights are to be inversely proportional to the variance, an appropriate weighting scheme for least squares analysis is to set the weight for P_{ij} equal to n_{ij}.

However if some of the P_{ij} are outside of the range of .2 to .8, the product $P_{ij}(1-P_{ij})$ will be quite variable and σ_{Pij}^2 will be affected by the size of P_{ij}. An angular transformation

$$A_{ij} = \arcsin(\sqrt{P_{ij}})$$

will remove this dependence (angular transformations are discussed further in section B.3.1). The variance σ_{Aij}^2 of the A_{ij} is now dependent only on n_{ij}:

$$\sigma_{Aij}^2 = 821/n_{ij}$$

where A_{ij} is measured in degrees. In this case the A_{ij} should be analysed rather than the P_{ij}, with the weight for each A_{ij} being n_{ij}.

Case 2 (σ_{ebij}^2 moderate and constant)

For data sets that follow case 2, σ_{Pij}^2 has the form

$$\sigma_{Pij}^2 = P_{ij}(1-P_{ij})/n_{ij} + \sigma_{eb}^2$$

The first step is to estimate the constant σ_{eb}^2. Cochran sets out an approximate procedure for this, based on a calculation of the relative size of the binomial and extra-binomial variance terms. This estimate σ_{eb}^2 is then used to estimate σ_{Pij}^2 for each P_{ij}. The inverse of this estimate of σ_{Pij}^2 is the weight assigned to each P_{ij} in a weighted least squares analysis.

Case 3 (σ_{ebij}^2 moderate and of the form $\lambda P_{ij}(1-P_{ij})$)

Here

$$\sigma_{Pij}^2 = P_{ij}(1-P_{ij})/n_{ij} + \lambda P_{ij}(1-P_{ij})$$

for some constant λ. Before estimating λ, an angular transformation

$$A_{ij} = \arcsin(\sqrt{P_{ij}})$$

is suggested in order to remove the effect of the size of P_{ij} on the variance. The variance σ_{Aij}^2 of the transformed proportions A_{ij} now has the relatively simple form:

$$\sigma_{Aij}^2 = 821/(1/n_{ij} + \lambda) \qquad (C3)$$

where A_{ij} is measured in degrees. A value for λ is then obtained by the same procedure used to estimate σ_{eb}^2 for case 2 and used to estimated σ_{Aij}^2 for each A_{ij}. The weight assigned to each A_{ij} in a weighted least squares analysis is the inverse of this estimate.

Case 4 (σ_{ebij}^2 large and constant)

For this case σ_{Pij}^2 is approximately constant, since the binomial component can be ignored. Weights are not required and unweighted least squares methods can be applied to the P_{ij}.

Case 5 (σ_{ebij}^2 large and of the form $\lambda P_{ij}(1-P_{ij})$)

Here also the binomial component can be ignored, and σ_{Pij}^2 is approximately given by

$$\sigma_{Pij}^2 = \lambda P_{ij}(1-P_{ij})$$

If an angular transformation such as

$$A_{ij} = \arcsin(\sqrt{P_{ij}})$$

is carried out, the variance σ_{Aij}^2 of the A_{ij} will be approximately constant. In this case weights are not required and unweighted least squares methods can be applied to the A_{ij}.

B.2.2.2 Regression Method

This method represents an alternative to Cochran's method for data sets that fall into Case 3 of section B.2.2.1. For this case, which is expected to occur quite frequently, both the binomial and extra-binomial variance components of σ_{Pij}^2 are present and the extrabinomial component σ_{ebij}^2 has the form $\lambda P_{ij}(1-P_{ij})$. As shown in equation (C3) in section B.2.2.1, after an angular transformation is applied the variance σ_{Aij}^2 of the transformed pen proportions A_{ij} has the form

$$\sigma_{Aij}^2 = 821/(1/n_{ij} + \lambda)$$

In this method σ_{Aij}^2 is modelled by the more general formula

$$\sigma_{Aij}^2 = a_0 + a_1/n_{ij}$$

An iterative reweighting procedure is employed to estimate a_0 and a_1:

1. Start with some initial estimates $\hat{a}_0$ and $\hat{a}_1$ (e.g. $\hat{a}_0 = 821$ and $\hat{a}_1 = 0$).

2. Use $\hat{a}_0$ and $\hat{a}_1$ to obtain the initial weights w_{ij} to be assigned to the A_{ij}. Each w_{ij} is the inverse of the variance estimate for A_{ij}:

$$w_{ij} = (\hat{a}_0 + \hat{a}_1/n_{ij})^{-1}$$

3. Calculate the weighted mean $\bar{A}_{Wi}$ for each treatment using the w_{ij}:

$$\bar{A}_{Wi} = \Sigma_j w_{ij} A_{ij} / \Sigma_j w_{ij}$$

4. Calculate the deviations D_{ij} of the A_{ij} from $\bar{A}_{Wi}$:

$$D_{ij} = A_{ij} - \bar{A}_{Wi}$$

5. Carry out a linear regression of D_{ij}^2 on $(1/n_{ij})$, and obtain new estimates $\hat{a}_0'$ and $\hat{a}_1'$ from the coefficients of this regression:

$$D_{ij}^2 = a_0 + a_1(1/n_{ij})$$

6. If the new values $\hat{a}_0'$ and $\hat{a}_1'$ are sufficiently close to $\hat{a}_0$ and $\hat{a}_1$, it is assumed that the process has stabilized and $\hat{a}_0'$ and $\hat{a}_1'$ are taken as the final parameter estimates. If $\hat{a}_0'$ and $\hat{a}_1'$ are not sufficiently close to $\hat{a}_0$ and $\hat{a}_1$, replace $\hat{a}_0$ by $\hat{a}_0'$ and $\hat{a}_1$ by $\hat{a}_1'$ and repeat steps 2 to 5. After a few iterations the values should stabilize.

Let $(\hat{a}_0)_S$ and $(\hat{a}_1)_S$ be the stabilized values of a_0 and a_1. The final weights for the A_{ij} are given by the inverse of the estimate of σ_{Aij}^2 using the stabilized values:

$$w_{ij} = [(\hat{a}_0)_S + (\hat{a}_1)_S/n_{ij}]^{-1}.$$

B.3 Transformations for Proportions

B.3.1 Angular Transformations

Let P be a proportional variable of the form Y/N, where N is a positive integer and Y is a positive integer in the range of 0 to N. If P is binomially distributed, its variance P(1-P)/N is dependent on the size of P. The standard angular transformation to remove the effect of the size of P is

$$A = \arcsin(\sqrt{P})$$

The variance of σ_A^2 of A is dependent only on N:

$$\sigma_A^2 = 821/N \text{ if A is measured in degrees}$$
$$\sigma_A^2 = .25/N \text{ if A is measured in radians}$$

In order to better remove the dependence of the variance on P, a practice recommended in most texts is to replace P values of 0 by .25/N and P values of 1 by (1-.25/N) prior to transformation.

A variation of the angular transformation that does not require end value adjustments was developed by Anscombe:

$$A = \arcsin(\sqrt{R}) \quad \text{where} \quad R = (Y+.375)/(N+.75)$$

A more recent version, called the Freeman-Tukey binomial transformation, also avoids the need for end value adjustments and is becoming more common in toxicological studies. Its form is:

$$A = [\arcsin(\sqrt{P_1}) + \arcsin(\sqrt{P_2})] / 2$$

where $P_1 = Y/(N+1)$ and $P_2 = (Y+1)/(N+1)$

In a study of possible methods for the analysis of proportional data, the Freeman-Tukey transformation showed a distinct advantage over the standard transformation with end value adjustments (Haseman and Kupper, 1979).

B.3.2 Relation of the Angular to the Square Root Transformation

This discussion is relevant to the conversion of a count to a proportion by dividing by a limit L (discussed in section 5.5.2). If a count Y is much smaller than the limit L, the range of Y will not be affected by L and it will have approximately a Poisson distribution. The question then arises as to whether it is appropriate to apply an angular transformation to Y (which is standard for proportions). The answer is that it is appropriate, since applying an angular transformation to Y is equivalent to applying a square root transformation (which is the standard transformation for counts).

To show the equivalence, consider the behaviour of the function arcsin(x). As x decreases to small values, the value of arcsin(x) becomes asymptotically proportional to x. If an angular transformation is applied to P where P is small and equal to Y/L, the transformed value A can be represented by

$$A = \arcsin(\sqrt{P}) = k\sqrt{P} = k\sqrt{Y}/\sqrt{L}$$

where k is a proportionality constant (k is $180/\pi$ if x is measured in degrees and 1 if x is measured in radians). Thus if L is constant, A is proportional to $\sqrt{Y}$ which demonstrates the equivalence of the transformations.

B.3.3 The Logit Transformation

Another transformation for a proportion P, that is similar to but more extreme than the angular transformation, is the logit transformation:

$$G = \mathrm{Log}_e((P+C)/(1-P+C))$$

For values of P that are close to 0 or to 1, the scale is stretched out even more than it is with the angular transformations. If P is binomially distributed with variance P(1-P)/N, the variance of G will be approximately equal to 1/(NP(1-P)). The constant C is a small positive value, added to stabilize the value of G for P values of 0 or 1. One commonly employed value of C is 1/2N (Snedecor and Cochran 1967, p.497), where N is the denominator of P.

B.4 Tests of Homogeneity of Variance

The usual test for testing the equality of two variances is the F-test of the ratio of the larger variance to the smaller. For testing homogeneity among more than two variances, the usual test is Bartlett's test. Both these tests are described in standard texts. However the latter is generally considered to be vulnerable to non-normality in the data.

A number of alternative tests have been developed. Among them are:

Levene's Test: This test involves running an ANOVA on the absolute values of the deviations of the data values from their treatment means (Levene, 1960). The inhomogeneity is considered to be significant if the ANOVA F-test for differences between treatments is significant. The test is recommended in a number of studies, including Miller (1986), because of its robustness when applied to non-normal data.

Normal Score Test: Described in Fligner and Killeen (1976), this test involves the ranking of the absolute values of the deviations of data values from their treatment means. The ranks are then converted to normal scores, and an ANOVA F-test for differences between treatments is run on these scores. The inhomogeneity is considered to be significant if the F-test is significant.

In addition a number of tests are based on special ranking systems for the deviations of data values from their means. These include tests by Freund and Ansari (1957) and by Siegel and Tukey (1960).

A comparative study involving a large number of methods was carried out by Conover, Johnson and Johnson (1981), and they concluded that Levene's test and the normal score test were among the best. These is also a good discussion of these methods in Madansky (1988).

B.5 Identification of Outliers

The assessment of whether or not to accept an extreme data value as a valid member of a data set is a difficult but sometimes a very critical matter in statistical analysis. The following discussion is intended only as a very brief introduction to this complex and difficult subject.

Classical methods for the identification of outliers have been based on the probability of extreme values occurring by chance from random variation, using probability theory and assumptions about the distribution of random variation to derive this probability. However more recent methods tend to be more pragmatic and less theoretically oriented.

A simple but widely used procedure is that suggested by Tukey, which involves setting outside limits for valid observations using the 25th and 75th percentile points of the data as a 'yardstick' (Tukey, 1977). Values must be within a certain number of multiples (usually 1.5) of the 25th - 75th percentile range. An assumption is necessary concerning the distribution of valid data values, but this assumption can be approximate in nature. This method is robust in that it can be applied in cases where there may be several outliers. It has been elaborated on by Hoaglin and Iglewicz (1987).

Very simple approaches may also be appropriate, even if they involve a subjective element. For example Miller recommends that the data be plotted on a probit plot and visually examined (Miller 1986, p. 10-14). The use of a probit plot allows for easy visual assessment of the degree of deviation of an extreme value from the distribution followed by the rest of the data set.

REFERENCES

Abelson, R.P. and Tukey, J.W. (1963). Efficient utilization of non-numerical information in quantitative analysis: General theory and the case of simple order. Ann Math Stat 34, 1347 -1369.

Altham, P.M.E. (1978). Two generalizations of the binomial distribution. Applied Statistics 27, 162-167.

Armitage, P. (1955). Test for linear trends in proportions and frequencies. Biometrics 11, 375-386

ASTM (1984). Standard Practice for Conducting Reproductive Studies With Avian Experiments. ASTM Committee E-47.

Bartholomew (1959). A test for homogeneity of ordered alternatives I, II. Biometrika 46, 35-48.

Cochran, W.G. (1943). Analysis of variance for percentages based on unequal numbers. JASA 38, 287-301.

Cochran, W.G. (1954). Some methods of strengthening the common chi-square tests. Biometrics 10, 417-451.

Cochran, W.G. (1977). Sampling Techniques, 3rd ed. John Wiley and Sons.

Cochran, W.G. and Cox, G.M. (1957). Experimental Designs, 2nd ed. John Wiley and Sons, New York.

Conover, W.J. (1971). Practical Non-Parametric Statistics. John Wiley and Sons.

Conover, W.J. (1974). Some reasons for not using the Yates continuity correction on 2 x 2 contingency tables. JASA 69, 374-376.

Conover, W.J., Johnson, M.E. and Johnson, M.M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. Technometrics 23, 351-361.

Crowder, M.J. (1978). Beta-binomial ANOVA for proportions. Applied Statistics 27, 34-37.

Crump, K.S. and Howe, R.B. (1988). A small-scale study of permutation tests for detecting teratogenic effects. U.S. FDA Internal Report.

Duncan, D.B. (1955). Multiple range and multiple F tests. Biometrics 11, 1-42.

Dunnett, C.W. (1955). A multiple comparisons procedure for comparing several treatments with a control. JASA 50, 1096-1121.

Edgington, E.S. (1980). Randomization Tests. Marcel Dekker, Inc.

Edgington, E.S. (1987). Randomization Tests, 2nd edition. Marcel Dekker Inc.

Fligner, M.A. and Killeen, T.J. (1976). Distribution-free two-sample tests for scale. JASA 67, 342-346.

Freund, J.E. and Ansari, A.R. (1957). Two-way rank sum test for variances. Technical Report No. 34, Virginia Polytech Institute, Blacksburg, Virginia.

Games, P.A. and Howell, J.F. (1976). Pairwise multiple comparison procedures with unequal n's and/or variances: a Monte Carlo study. Journal of Education Statistics 1, 113-125.

Gladen, B. (1979). The use of the jackknife to estimate proportions from toxicological data in the presence of litter effects. JASA 74, 278-283.

Haseman, J.K. and Kupper, L.L. (1979). Analysis of dichotomous response data from certain toxicological experiments. Biometrics 35, 281-293.

Haseman, J.K. and Soares, E.R. (1976). The distribution of fetal death in control mice and its implications on statistical tests for dominant lethal effects. Mutation Research 41, 277-288.

Hewett, J.E. and Bair, E. (1986). A two-stage test for ordered means in the Poisson case with an example from mutagenicity testing. Biometrics 42, 647-651.

Hoaglin, D.C. and Iglewicz, B. (1987). Fine-tuning some resistant rules for outlier labelling. JASA 82, 1147-1149.

Hochberg, Y. and Tamhane, A.C. (1987). Multiple Comparison Procedures. John Wiley, New York.

Hollander, M. and Wolfe, D.A. (1973). Nonparametric Statistical Methods. John Wiley and Sons, London-New York.

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

Kempthorne, O. (1979). In dispraise of the exact test: reactions. Journal of Statistical Planning and Inference 3, 199-213.

Klotz, J.H. (1966). The Wilcoxon, ties, and the computer. JASA 61, 772-787.

Kramer, C.Y. (1956). Extension of multiple range tests to group means with unequal numbers of replications. Biometrics 12, 307-310.

Kupper, L.L. and Haseman, J.K. (1978). The use of a correlated binomial model for the analysis of certain toxicological experiments. Biometrics 34, 69-76.

Kruskal, W.H. and Wallis, W.A. (1952). Use of ranks in one-criterion variance analysis. JASA 47, 583-621.

Lehman, S.Y. (1961). Exact and approximate distributions for the Wilcoxon statistic with ties. JASA 56, 293-298.

Levene, H. (1960). Robust test for equality of variances. Contributions to Probability and Statistics (ed. by I. Olkin, S.G. Ghurye, W. Hoeffding, W.G. Madow and H.B. Mann), Stanford University Press, 278-292.

Madansky, A. (1988). Prescriptions for working statisticians. Springer-Verlag, New York.

Mehta, C.R., Patel, N.R. and Tsiatis, A.A. (1984). Exact significance testing to establish the equivalence of two treatments being compared on the basis of ordered categorical data. Biometrics 40, 819-825.

Miller, R.G. Jr. (1974). The jackknife - a review. Biometrika 61, 1-15.

Miller, R.G. Jr. (1986). Beyond ANOVA, Basics of Applied Statistics. John Wiley and Sons.

Mineau, P., Boersma, D.C. and Collins, B. (1994). An analysis of avian reproduction studies submitted for pesticide registration. Ecotoxicology and Environmental Safety (in press).

OECD (1981). One-Generation Reproduction Toxicity Test. OECD Guideline for Testing of Chemicals #415.

OECD (1983). Two-Generation Reproduction Toxicity Test. OECD Guideline for Testing of Chemicals #416.

OECD (1984). Avian Reproduction Test. OECD Guideline for Testing of Chemicals #206.

Pack, S.E. (1986). Hypothesis testing for proportions with overdispersion. Biometrics 42, 967-972.

Paul, S.R. (1982). Analysis of proportions of affected foetuses in teratological experiments. Biometrics 38, 361-370.

Petrondas, D.A. and Gabriel, K.R. (1983). Multiple comparisons by rerandomization tests. JASA 78, 949-957.

Piccirillo, V.J. and Quesenberry, R.P. (1980). Reproductive capacities of control mallard ducks (anus platyrhynchos) during a one-generation reproduction study. Journal of Environmental Pathology and Toxicology 4, 133-139.

Rao, J.N.K. and Scott, A.J. (1992). A simple method for the analysis of clustered binary data. Biometrics 48, 577-585.

Rice, W.R. (1988). A new probability model for determining exact p-values for contingency tables when comparing binomial proportions. Biometrics 44, 1-22.

Ruberg, S.J. (1989). Contrasts for identifying the minimum effective dose. JASA 84, 816-822.

Satterthwaite, F.E. (1946). An approximate distribution of estimates of variance components. Biometrics 2, 110-114.

Scheffé, H. (1959). The Analysis of Variance. John Wiley and Sons.

Shirley, E. (1977). A nonparametric version of Williams' test for contrasting increasing dose levels of a treatment. Biometrics 33, 386-389.

Shirley, E. and Hickling, R. (1981). An evaluation of some statistical methods for analysing numbers of abnormalities found amongst litters in teratology studies. Biometrics 37, 819-829.

Siegel, S. and Tukey, J.W. (1960). A nonparametric sum of ranks procedure for relative spread in unpaired samples. JASA 55, 429-445.

Snedecor, G.W. and Cochran, W.G. (1967). Statistical Methods, 6th ed. The Iowa State University Press.

Tukey, J.W. (1977). Exploratory Data Analysis. Addison-Wesley.

U.S. EPA (1986) Avian reproduction test. Hazard Evaluation Division, Office of Pesticides Programs, Washington, D.C. EPA 540/9-86-139.

Upton, G.J.G. (1982). A comparison of alternative tests for the 2 x 2 comparative trial. JRSS A 145, 86-105.

Van der Laan, P. and Verdooren, L.R. (1987). Classical analysis of variance methods and nonparametric counterparts. Biom. J. 29, 635-665.

Weil, C.S. (1970). Selection of a valid number of sampling units and a consideration of their combination in toxicological studies involving reproduction, teratogenesis or carcinogenesis. Fd. Cosmet. Toxicol. 8, 177-182.

Welch, B.L. (1947). The generalization of Student's problem when several different population variances are involved. Biometrika 34, 28-35.

WHO (1984). Principles for evaluating health risks to progeny associated with exposure to chemicals during pregnancy. Environmental Health Criteria #30. Geneva.

Williams, D.A. (1971). A test for treatment differences when several dose levels are compared with a zero-dose control. Biometrics 27, 103-117.

Williams, D.A. (1972). The comparison of several dose levels with a zero-dose control. Biometrics 28, 519-531.

Williams, D.A. (1975). The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. Biometrics 31, 949-952.

Williams, D.A. (1986). A note on Shirley's nonparametric test for comparing several dose levels with a zero-dose control. Biometrics 42, 183-186.