Plant Injury Analysis: Contingency Tables as an Alternative to Analyses of Variance

P. M. McCOOL, T. YOUNGLOVE, and R. C. MUSSELMAN, Statewide Air Pollution Research Center, University of California, Riverside 92521, and R. R. TESO, California Department of Food and Agriculture, Statewide Air Pollution Research Center

In research dealing with assessment of plant injury, problems often occur in data analysis when specific treatments or controls result in little or no injury. For example, ratings of leaf or root injury from pathogen attack may have large numbers of zero or low injury scores when low inoculum level treatments are used. This results in a highly skewed distribution of the injury scores, with a large proportion of the scores at zero. For higher inoculum levels, the injury response distribution tends toward the normal distribution. This leads to another problem when the variance of the skewed low treatments is significantly less than the variance for the higher treatment levels. Additional difficulties may arise when a rating system has unequal intervals within the rating scale that could skew the distribution of injury scores. Proper analysis of the data is vital to accurate assessment of treatment effects.

Analysis of variance

In the analysis of variance (ANOVA) of injury data, the assumptions of normality and equal variances are violated when the experiment contains control or low level treatments that cause little or no injury. The solution usually offered is to transform the data. When the skewed, low injury distribution is transformed, the result is a skewed, truncated, transformed distribution with a large number of transformed zero injury scores at one end. Variable transformations are not useful for dealing with either problem because a transformation will not redistribute the large number of zero injury scores.

The ratio of mean square (treatments)/mean square (error) is used to test the hypothesis of no significant treatment differences in an ANOVA. ANOVAs are frequently used on highly nonnormal data with the assumption that the probabilities are close to correct (7). With such data, however, the significance levels of the ratios from F tables are not correct because the normality assumption is violated. The result is that the mean squares are not distributed as independent chi-square random variables so that the calculated ratio is not distributed as an F ratio (9). For sums of squares for treatments and sums of squares error to be independent chi-square random variables, Cochran's theorem requires that the observations be normally distributed (5). Small deviations from normality are not thought to have much effect on the probabilities (8), but injury data are often highly nonnormal. Whereas nonnormal data affect identification of treatment effects, unequal variances can lead to misidentification of differences between treatment means.

When unequal treatment variances occur, the distribution of the sample statistic used to test for differences between treatment means no longer follows a Student's t distribution. The distribution of the test statistic can be estimated, but the test is no longer exact (4).

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. § 1734 solely to indicate this fact.

Contingency table analysis

There are many alternatives to ANOVA, including weighted ANOVA, nonparametric one- and two-way ANOVA, assorted nonparametric tests, and log-linear modeling. Contingency table analysis by log-linear modeling provides a better procedure than a standard ANOVA for analyzing nonnormal data because more reliable probabilities result. Log-linear modeling may be used to test many of the same hypotheses that can be tested with an ANOVA. Testing for treatment effect is possible by testing independence of injury from treatment. In contingency table analysis, independence of injury from treatment is equivalent to nonsignificance of treatment effects. Interactions among factors may be evaluated by testing the conditional independence of the factors. A drawback to contingency table analysis is the lack of equivalent mean separation tests, such as Duncan's new multiple range test. However, the odds of given injury classes within each treatment can be estimated for models found to best fit the data. Additionally, confidence intervals can be formed around the odds ratios, allowing comparison of treatments (3). In both the ANOVA and the log-linear models, the null hypothesis tested states that injury is independent of treatment effects.

A distinct advantage in the use of contingency table analysis is that the analysis is not affected by the nonnormal distribution of the data because the hypothesis testing is not based on the normality of the sample distribution. In log-linear modeling, the contingency table is considered as a sample from a multinomial distribution or as samples from several multinomial distributions (2). Contingency tables consist of discrete "cells" of data, each cell a location within the table. Also, the probabilities from all cells are assumed to add up to one. Thus, each observation must fall into one of the cells, and the sample size must be sufficiently large that each cell will have at least one observation.

A good example of a multinomial distribution is the set of six possible outcomes from rolling a die. The probability of any one side being up on a roll is 1/6. Two groups of 100 dice, red and white, would produce a 2 by 6 contingency table with 200 observations generated by rolling each die once. By counting the number of observations falling into each of the 12 cells, the observed frequencies could be tested for the significant effect of color on die outcome.

Plant injury is often scored for severity according to a numerical classification, such as integer values from 0 to 10, where 0 represents no injury and 10 represents 100% necrosis. Injury rating for each plant corresponds to a roll of the die in the example given. To be analyzed with log-linear modeling, injury data must be put into a contingency table with distinct cells. Both the treatment and the response variables must be divided into discrete levels so that each data unit falls into only one cell of the table. For example, injury ratings grouped into categories as 0 to 3, 3 to 7, and 7 to 10 would not be valid because plants at 3 and 7 would fall into two cells. If injury is measured in integer increments, nonoverlapping categories, such as 0 to 3.5, 3.5 to 7.5, and 7.5 to 10, should be chosen to avoid borderline cases.

Categories should be chosen carefully and not repeatedly changed; to keep selecting categories until the analysis shows the desired result is not valid.

Greater numbers of cells in the table increase the accuracy of the log-linear model. However, the number of cells must be balanced against the desire to have at least one observation in each cell. Log-linear modeling can be accomplished with empty cells, but this should be avoided, particularly with small numbers of cells. When injury scores are extremely skewed, all cells may not be filled. When none of the control plants is injured, the injury scores cannot be broken into more than one category without having many empty cells in the control treatment. This can be solved by grouping the control treatment with contiguous low level treatments or by eliminating the control group. When only two treatment levels are being evaluated and the control (or lowest level) treatment shows no injury, the conclusions are obvious and analysis is unnecessary.

Two common statistics used to evaluate treatment significance are the Pearson chi-square and the likelihood ratio chi-square. Either statistic may be used to test for significance of contingency table models, but only the likelihood ratio chi-square is properly used to evaluate the effects of including additional model parameters. Because the likelihood ratio statistics are asymptotically distributed as chi-square (1), a large number of observations are required for the various tests to be accurate. Therefore, the greater the number of observations, the closer the distribution of the likelihood ratio chi-square approaches the distribution of the chi-square. Although not defined by specific rules, what constitutes a large number of observations depends in part on the number of cells in the table. A rule of thumb is to have an average of 10 observations per cell. For a table with six or fewer cells, the number of observations per cell should average 15 or more.

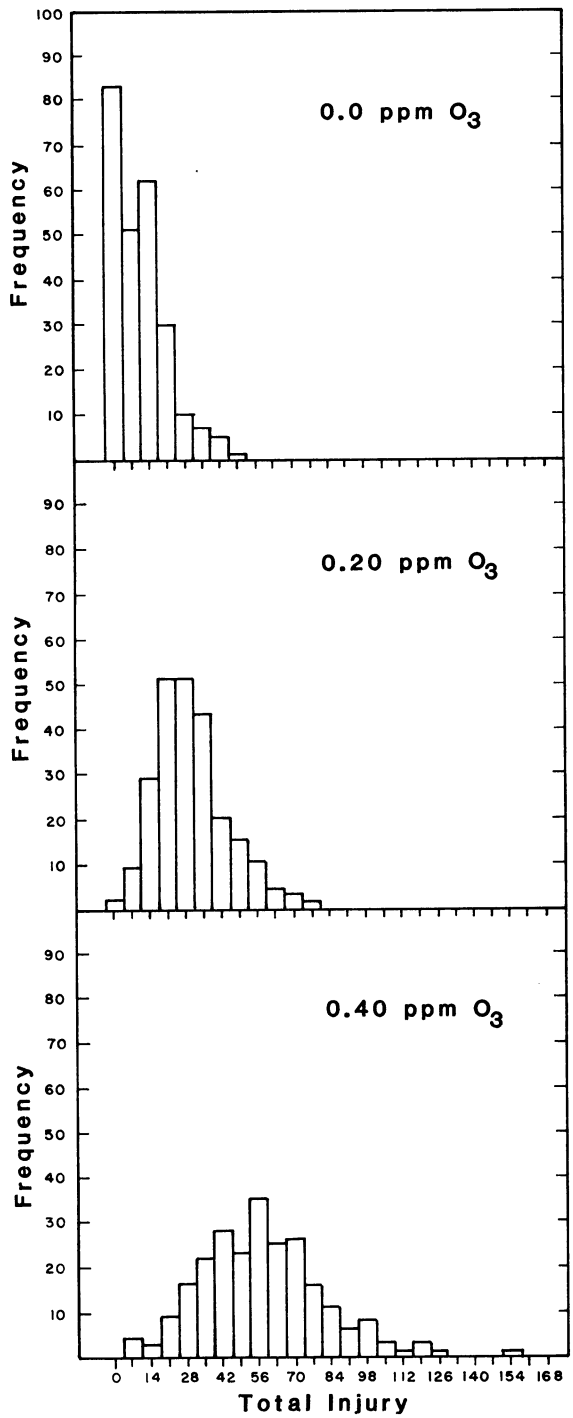

Fig. 1. Frequency distribution of total plant injury for kidney bean exposed to 0, 0.20, or 0.40 ppm of ozone.

An example

The impact of air pollution on beans can be used to illustrate the problems encountered in an ANOVA on highly nonnormal data. Although this example demonstrates the use of contingency table analysis using log-linear modeling, the discussion is applicable to other types of research in which nonnormal data are collected and evaluated.

In an experiment to test the uniformity of ozone exposure chambers, beans (*Phaseolus vulgaris* L. 'California Light Red Kidney') were exposed to three levels of ozone. The experiment was designed as a split plot with three replicate chambers at each level of ozone and four sections (A, B, C, and D) within each chamber. Total injury was calculated by rating each trifoliate leaf on a scale of 0 to 10 and summing the individual leaf ratings for the total plant. The histograms of total plant injury for 0-ppm and 0.20-ppm ozone treatments illustrate the skewness and truncation of the data (Fig. 1). The histogram of total plant injury for the 0.40-ppm treatment appears normally distributed, with a higher variance than the 0-ppm and 0.20-ppm treatments (Fig. 1).

Ignoring the inappropriateness of the data (nonnormal) for ANOVA, the data were analyzed by standard ANOVA techniques for comparative purposes. Division of each of the nine ozone chambers into quadrants (sections) resulted in a restricted randomization of the design. The correct ANOVA for this experimental setup is a split-plot design, with ozone as the main plot treatment and the quadrants as subplots. The ozone × replicate interaction is used for the main plot error, and the ozone × replicate × section interaction is used for the subplot error. The results of the ANOVA are shown in Table 1.

The mean squares and the *F* ratios have been calculated, even though distribution is incorrect because the data are highly nonnormal. The *F* value of 29.46 for ozone is so highly significant that ozone probably did have a significant effect, but the significance level cannot be determined with any certainty. The *F* value of 1.19 for the section factor is not significant, but

Table 1. Split-plot analysis of variance for injury of kidney bean plants exposed to three levels of ozone

Source	df	SS	MS	F^a
Ozone (O)	2	13,145.88	6,572.44	29.46**
Replicate (R)	2	300.01	150.01	0.67 ns
Error A (O × R)	4	892.41	223.10	
Section (S)	3	354.10	118.03	1.19 ns
R × S	6	496.72	82.79	0.84 ns
O × S	6	205.84	34.31	0.35 ns
Error B	12	1,187.46	98.95	
Total	35	16,582.42		

[a]** = Statistical significance ($P = 0.01$), ns = not significant.

with the high nonnormality of the data, the significance cannot be ascertained with any degree of accuracy. Any confidence intervals calculated from this analysis would be highly suspect.

The same data set was subjected to a contingency table analysis. First, the contingency table had to be constructed. As an aid in choosing cell cutoff points, histograms were produced of total plant injury separated by treatment to identify the best place to split the injury category (Fig. 1). Too many potential scores for total plant injury prevented giving each possible score an individual cell, so the scores were grouped. A statistical package helped speed production of the histograms. Total plant injury scores were split into two categories, low (0–17.5) and high (17.5 and up). Data range was insufficient for more than two categories, and although some of the original information was lost, using only two categories was necessary to conduct the analysis. The value 17.5 was chosen arbitrarily because it fell near the median for the total population and did not result in any empty cells. The resulting 2 × 3 × 4 table is shown in Table 2.

Model selection. With the contingency table set, the next step was to select the "best" log-linear model for the data. Several computer packages (BMDP, SAS, SPSSX) are available to help select and estimate the best model. The exact method to use depends on the computer package selected.

One frequently used approach to model selection using the BMDP statistical package tests the significance of the variables one at a time, adding variables that significantly increase the fit of the model. In our example, the parameters were S (chamber section), I (total plant injury), and O (ozone treatment level). The BMDP statistical package was used for the example; Table 3 summarizes the statistical analysis. The first step in selecting a model was to determine which parameters of any given interaction order were significant. Significance in the BMDP run was determined by the probability of the parameters equaling zero. Thus, parameters of any order with a probability of 0.05 or less should be significant, i.e., a 95% probability they are not zero. The analysis confirmed that at least one first-order (S, I, or O) and one second-order (SI, SO, or IO) parameter were significant ($P < 0.05$) (Table 3). The third-order parameter (SIO), however, had an approximate 51% chance of being equal to zero and was not significant. Thus, there was not a significant section by injury by ozone interaction.

BMDP also tests for significance of individual interactions in the log-linear model. As before, probabilities less than 0.05 indicate significant interactions. The SIO (third-order) parameter was not significant in the analysis (Table 3). Ozone had a highly significant effect on injury, shown by the high degree of significance of the IO term. This confirms the result found in the ANOVA. Unlike the ANOVA, where section did not approach significance, the SI effect was significant at the 0.05 level. The SO term was not significant, so the appropriate model for the data includes the S, O, I, SI, and IO.

This analysis considers only hierarchical models, i.e., models in which lower order parameters that are part of a higher order interaction included in the model are also included in the model, regardless of significance. Because the SI parameter was significant (Table 3), indicating section by injury interactions, both S and I parameters were included in the model.

The best model included the SI and IO terms, indicating that total plant injury depended on both ozone and section, since the best model included both ozone by injury (IO) and section by injury (SI) interactions. The nonnormality of the data led to the ANOVA completely missing the significant effect of the section. A second method of model selection (not discussed here)

Table 2. Contingency table of total plant injury counts, observed and expected[a]

Ozone	Injury[b]	Chamber section A	B	C	D
0 ppm	Low	54 (54.8)	50 (53.2)	39 (40.3)	53 (47.6)
	High	9 (10.7)	12 (13.4)	19 (15.0)	10 (10.9)
0.20 ppm	Low	11 (11.2)	15 (10.9)	9 (8.2)	5 (9.7)
	High	36 (41.9)	52 (52.6)	61 (58.8)	47 (42.7)
0.40 ppm	Low	3 (2.0)	1 (1.9)	2 (1.4)	1 (1.7)
	High	56 (48.5)	63 (60.9)	62 (68.1)	46 (49.4)

[a]Total plant injury counts: sum of each trifoliate leaf injury, rated 0–10 (0 = no injury, 10 = 100% necrosis). Observed frequencies from original data; expected frequencies calculated for comparison by log-linear models.
[b]Low score = 0–17.5, high score = 17.5 and up.

Table 3. Summary of BMDP model selection analysis: simultaneous test that all interactions of order *k* are zero and significance tests for specific interactions

Effects[a]	df	Pearson chi-square	Probability
***k* order**			
1 (S, I, or O)	6	35.64	0.000
2 (SI, SO, or IO)	11	379.22	0.000
3 (SIO)	6	5.25	0.512
Specific interactions			
SI	3	8.99	0.030
SO	6	5.99	0.424
IO	2	391.30	0.000
SIO	6	5.50	0.482

[a]I = injury, S = section, O = ozone.

Table 4. Log-linear and exponentiated parameters from BMDP for interaction terms of IO, SI model

Chamber section	Log-linear (lambda) injury Low	High	Exponentiated [exp (lambda)] injury Low	High
A	0.134	−0.134	1.143	0.875
B	0.004	−0.004	1.004	0.996
C	−0.191	0.191	0.826	1.210
D	0.053	−0.053	1.054	0.948

Odds of low injury

Cross product	Odds ratio
A vs. B = (1.143 × 0.996)/(1.004 × 0.875) =	1.296
A vs. C = (1.143 × 1.210)/(0.826 × 0.875) =	1.914
A vs. D = (1.143 × 0.948)/(1.054 × 0.875) =	1.175

Table 5. Log-linear and exponentiated parameters from BMDP for interaction terms of IO, SI model

Ozone level	Log-linear (lambda) injury Low	High	Exponentiated [exp (lambda)] injury Low	High
0 ppm	1.30	−1.30	3.670	0.273
0.20 ppm	−0.178	0.178	0.837	1.195
0.40 ppm	−1.122	1.122	0.326	3.071

Odds of low injury

Cross product	Odds ratio
0 ppm vs. 0.20 ppm = (3.67 × 1.195)/(0.837 × 0.273) =	19.19
0 ppm vs. 0.40 ppm = (3.67 × 3.071)/(0.326 × 0.273) =	126.64
0.20 ppm vs. 0.40 ppm = (.837 × 3.072)/(0.325 × 1.194) =	6.60

utilizing all possible models can be accomplished through BMDP (1). Although two- or three-way tables are practical, four-way or larger tables include too many models to be manageable.

Parameter estimation and evaluation. The next step in the analysis after selection of the best model was estimation of the parameters. Using BMDP, an additional analysis specifying the SI and IO models was necessary to determine the expected cell frequencies and the log-linear parameters. Examination of the expected cell frequencies helped determine which cells did not fit the model. The expected values under this model (Table 2), as determined for the example by BMDP, were quite close to the observed frequencies (Table 2), indicating a good fit to the data.

The next step after estimating the log-linear parameters was determining the effect of the interactions. The simplest way to identify interaction effects is to generate odds ratios that can give insight into treatment effects. Odds ratios are produced from exponential log-linear parameters (lambda values) generated in the analysis (Table 4). To determine the effect of section on injury (SI interaction), the odds of low injury occurring in section A vs. those in sections B, C, and D were calculated from the appropriate cross product ratios. When section A vs. section B was examined for the odds of low injury, the cross product of only these sections was used. The odds for low injury occurring in section A vs. section C and section A vs. section D were calculated similarly. The odds of occurrence of high injury could have been determined identically except that the low and the high injury columns would have been reversed before calculating the cross products. In the example data set, all of the odds ratios were greater than one, so a plant in section A was more likely to be in the low injury category than a plant in any other section. From the ratios, the odds of low injury occurring among plants in section A were 1.296 times those of plants in section B, 1.914 times those of plants in section C, and 1.175 times those in section D (Table 4). Therefore, section C was least likely to have plants in the low injury category. All other sections had approximately equal odds ratios for low injury. Other section odds ratios (BC, BD, and CD) were not necessary for identifying the high and the low injury sections and thus were not calculated. With other data sets, calculation of all odds ratios might be necessary to determine treatment effects.

The relative effects of ozone levels of injury (OI interactions) were determined in the same manner as those for the SI interactions, using cross products of the OI interaction parameters. The results indicated that the odds of low injury occurring among plants in the low ozone (0 ppm) chambers were 19.19 times those of plants in the medium ozone (0.20 ppm) chambers and 126.64 times those of plants in the high ozone (0.40 ppm) chambers (Table 5). The odds of low injury occurring among plants in the medium ozone chambers were 6.60 times those of plants in the high ozone chambers (Table 5).

In summary

Using log-linear modeling for highly nonnormal data will give more reliable results than simply running an ANOVA. Log-linear modeling does not require many sacrifices, and the lack of a mean separation test is a fair price to pay for information reported with reliable probabilities. Adequate observations are essential, but conducting contingency table analysis on even marginal numbers of observations is better than performing an inappropriate analysis, i.e., ANOVA. Although with our example data set, discrete statistical separations could not be assigned to the odds ratios as in Duncan's test, the odds ratios were calculated for factors and interactions that had a minimum level of significance ($P < 0.05$). The odds ratios were the "best" estimate of the actual treatment differences.

The variety of injury rating systems with different scalar attributes and the usual nonnormal injury response of plants stressed with both biotic and abiotic factors lead to problems in accurately assessing plant effects and the efficacy of control measures. Therefore, results must be reported reliably and treatment effects or interactions cannot be overlooked or incorrectly identified because of inappropriate analyses. We used determination of uniformity in air pollution injury within fumigation chambers as an example, but contingency table analysis can be utilized for other nonnormal data.

The contingency table analysis available on most computer statistical packages is somewhat more involved than ANOVA procedures. Several programs in addition to BMDP, used for our example, are readily available. For thorough understanding of the analysis and for additional information and examples, we recommend any of the standard texts on analyzing nonnormal data, e.g., the text by Kennedy (6).

LITERATURE CITED

1. Brown, M. 1977. Multiway frequency tables—the log-linear model. Pages 302 and 314 in: BMDP. Biomedical Computer Programs, P Series. W. J. Dixon and M. B. Brown, eds. University of California Press, Berkeley.
2. Gokhale, D. V., and Kullback, S. 1978. Introduction. Page 11 in: The Information in Contingency Tables. Marcel Dekker Inc., New York. 365 pp.
3. Hill, M. 1984. Using 4F and LR to obtain estimates of the odds ratio. Pages 6-11 in: BMDP Statistical Software Communications. Vol. 17. M. L. Musser-Brauner, ed. BMDP Statistical Software, Los Angeles. 16 pp.
4. John, P. W. M. 1971. Introduction. Page 6 in: Statistical Design and Analysis of Experiments. The Macmillan Co., New York. 356 pp.
5. John, P. W. M. 1971. Linear models and quadratic forms. Page 33 in: Statistical Design and Analysis of Experiments. The Macmillan Co., New York. 356 pp.
6. Kennedy, J. J. 1983. Analyzing Qualitative Data: Introductory Log-Linear Analysis for Behavioral Research. Praeger Publishers, New York. 262 pp.
7. Little, T. M., and Hills, F. J. 1972. Transformations. Pages 103 and 120 in: Statistical Methods in Agricultural Research. T. M. Little and F. J. Hills, eds. University of California Cooperative Extension, Berkeley. 242 pp.
8. Montgomery, D. C. 1976. Experiments with a single factor. Page 60 in: Design and Analysis of Experiments. John Wiley & Sons, New York. 418 pp.
9. Rohatgi, V. K. 1976. Sample moments and their distributions. Page 317 in: An Introduction to Probability Theory and Mathematical Statistics. John Wiley & Sons, New York. 684 pp.