CONTINGENCY TABLE ANALYSIS FOR COMPLEX SAMPLE DESIGNS: CPLX

Robert Fay, U.S. Bureau of the Census

1. INTRODUCTION

CPLX is a computer program for analysis of cross-classified data from complex sample survey designs. It implements essentially the standard maximum likelihood estimation procedures for factorial, hierarchical log-linear models as if the data were obtained under simple random sampling, but it provides test statistics and standard errors appropriate for a given complex design through use of replication methods. The program was developed at the Census Bureau by the author and has already been employed in a number of analytic studies there.

The purpose of this paper is to explain and summarize the principal features of CPLX, with the intent of enabling the reader to recognize situations in which the methodology implemented in this program represents a possible solution to otherwise difficult statistical problems. The general description here is intended to complement the technical documentation (Fay, 1982) available for the program. While the latter provides the exact instructions for its use, the purpose here is to suggest the scope and applications of the program to the reader.

The second section of the paper gives a general description of log-linear models, as well as the more specific class of models considered by CPLX. Log-linear models represent an area of considerable statistical development, especially for simple random sampling, and this section relies heavily upon citations to some of the wide literature in this field.

The third section attempts to describe and illustrate the sorts of situations in which the designs may be considered "complex" and more suitable for analysis by CPLX than by programs and statistical procedures assuming a simple underlying multinomial distribution. This section also discusses how the sources of variation in such complex designs can frequently be represented by an appropriate series of "replicates."

The weighted least squares approach offers an alternative solution to the problem of drawing inferences about log-linear models in the context of complex samples. The fourth section of this paper compares the scope of CPLX with the weighted least squares method. Although in some situations researchers may justifiably prefer the weighted least squares approach, this section points out that CPLX may be applied to a wide class of problems for which the other method is inappropriate.

The fifth section presents the statistical methodology implemented by CPLX in more detail, including a somewhat fuller description of the replication methods considered in the program. The sixth section indicates in a general manner how the program may be used and the incorporated features. The appendices summarize some of the technical details of the implementation.

2. LOG-LINEAR MODELS CONSIDERED BY CPLX

In the most general form, the log-linear model represents any statement about the expected values for a set of frequency data under some sampling distribution, in which the logarithms of the expected frequencies are given by a model that is linear in the parameters. This class includes the general problem of logistic regression, as well as applications to fully cross-classified data.

Discrete variables of two or more levels each are assumed to form the cross-classifications analyzed by CPLX. To illustrate the general ideas here, consider a cross-classification based upon four variables. A model of mutual statistical independence for all four variables is equivalent to

$$\ln F_{ijkl} = \lambda + \lambda^{I}_{i} + \lambda^{J}_{j} + \lambda^{K}_{k} + \lambda^{L}_{l} \qquad (2.1)$$

where F_{ijkl} gives the expected values for the cross-classified cells. The usual restrictions on the parameters (in order to insure identifiability) are

$$\sum_i \lambda^{I}_{i} = \sum_j \lambda^{J}_{j} = \sum_k \lambda^{K}_{k} = \sum_l \lambda^{L}_{l} = 0 \qquad (2.2)$$

In model (2.1), the parameters (except for λ) each depend on the level of exactly one of the variables. One alternative model for the cross-classification is given by

$$\ln F_{ijkl} = \lambda + \lambda^{I}_{i} + \lambda^{J}_{j} + \lambda^{K}_{k} + \lambda^{L}_{l} + \lambda^{IJ}_{ij} \qquad (2.3)$$

in which conditions (2.2) apply as well as the new conditions

$$\sum_i \lambda^{IJ}_{ij} = \sum_j \lambda^{IJ}_{ij} = 0 \qquad (2.4)$$

Model (2.3) differs from (2.1) in allowing the relationship between two variables, i and j, to be dependent and governed by new parameters λ^{IJ}_{ij}, which do not depend upon the levels of the remaining variables Thus, this model represents a simple generalization of the model of statistical independence, in the sense that (2.3) includes (2.1) as a special case. Choosing between the two models is a matter of determining whether the model of statistical independence is sufficient to explain the observed data or whether the departure is in the specific manner indicated by model (2.3).

A frequent application of log-linear models arises when one (or more) of the variables may be considered "dependent" in the sense often used in characterization of linear regression. That is, one wishes to build a model to describe, explain, or account for the changes in the distribution

of this variable in terms of the the remaining "independent" variables. For purposes of illustration, suppose that the first variable is composed of two levels and is considered "dependent" for purposes of the analysis. The log-linear model

$$\ln F_{ijkl} = \lambda + \lambda_j^J + \lambda_k^K + \lambda_l^L + \lambda_{jk}^{JK} + \lambda_{jl}^{JL} + \lambda_{kl}^{KL} + \lambda_{jkl}^{JKL} + \lambda_i^I + \lambda_{ij}^{IJ} + \lambda_{ik}^{IK} + \lambda_{il}^{IL} \quad (2.5)$$

with restrictions on the parameters similar to those in (2.2) and (2.4), is equivalent to

$$\ln (F_{1jkl}/F_{2jkl}) = 2\lambda_1^I + 2\lambda_{1j}^{IJ} + 2\lambda_{1k}^{IK} + 2\lambda_{il}^{IL} \quad (2.6)$$

Equation (2.6) is comparable to an equation for linear regression, where the "logits" of the left-hand side of the equation are expressed in in terms of a constant term, $2\lambda_1^I$, and coefficients, $2\lambda_{ij}^{IJ}$ etc. In this model, nothing is asserted about the relationships between the last three variables, while the relative proportions for the first variable, given the remaining three, are expressed in (2.6) as sums of a constant term plus coefficients that each depend on the level of exactly one of the independent variables.

A log-linear model similar to (2.5) is given by

$$\ln F_{ijkl} = \lambda + \lambda_j^J + \lambda_k^K + \lambda_l^L + \lambda_{jk}^{JK} + \lambda_{jl}^{JL} + \lambda_{kl}^{KL} + \lambda_{jkl}^{JKL} + \lambda_i^I + \lambda_{ij}^{IJ} + \lambda_{ik}^{IK} \quad (2.7)$$

This model is equivalent to (2.5) except for the omission of the last term. The corresponding logistic expression of this model

$$\ln (F_{1jkl}/F_{2jkl}) = 2\lambda_1^I + 2\lambda_{1j}^{IJ} + 2\lambda_{1k}^{IK} \quad (2.8)$$

omits a coefficient relating variable i to variable 1. Choosing between models (2.5) and (2.7) is a matter of determining whether the coefficient in (2.5) but excluded from (2.7) is necessary to describe the data. If not, model (2.7) or (2.8) posits no direct relationship between the first and last variables, once the effects of the other two variables are considered.

For purposes of discussion here, there are three major questions of statistical inference that arise in fitting log-linear models to sample data:

1. What are appropriate standard errors or confidence intervals for the estimated coefficients of the model?

2. Does a specific parameter or group of parameters make a statistically significant contribution to a model?

3. Is there evidence of a significant overall lack of fit of the model?

Question 1 presupposes an estimation strategy for the parameters, of course, but asks a question that arises in almost any situation of model fitting. The second question concerns choice of model; this is illustrated in the preceding discussion by the potential choice between models (2.1) and (2.3), or (2.5) and (2.7), as alternative explanations for the same observed cross-classified data.

When the observations have been sampled through simple random sampling from an infinite population, the observed cross-classification is distributed according to the multinomial distribution. Much of the available theory for the log-linear model has been developed for this distribution or other related distributions such as the product multinomial or Poisson. For these distributions, standard maximum likelihood theory provides these answers to the preceding three questions:

1. Standard errors for the maximum likelihood estimates may be based on general asymptotic theory for maximum likelihood estimation. (Some computer programs implement alternative estimators of the standard errors that are satisfactory approximations to the theoretical results in most practical applications.)

2. Tests of the contribution of a group of parameters to the model are given by the standard likelihood ratio chi-square. A similar, but less recommended test may be based on the differences of Pearson chi-square tests tests under the two models.

3. Tests of overall fit are given by the likelihood ratio chi-square test or the Pearson chi-square test.

The purpose of CPLX is to provide analogous answers to these questions when the observed data are from a more complex distribution. In addition to the requirement that each variable be discrete, CPLX fits factorial models of the type illustrated by (2.1), (2.3), (2.5), and (2.7). The models also must be hierarchical, that is, for each parameter in the model subscripted by a set of variables, there must be corresponding parameters in the model including all subsets of the set in question. For example, inclusion of a parameter λ_{ijk}^{IJK} requires inclusion of parameters λ, λ_i^I, λ_j^J, λ_k^K, λ_{ij}^{IJ}, λ_{ik}^{IK}, and λ_{jk}^{JK}. The preceding models all satisfy this property, as do virtually all models that would have practical significance when applied to sample survey data.

CPLX is also able to provide tests of overall fit and of the contribution of specific groups of parameters in some special applications in which a number of specific cells of the table are considered structural zeroes or are removed from analysis for some other reason. The log-linear model is then assumed to apply to the remaining cells.

The other practical limitations of CPLX are discussed more fully in appendix A.3. It is important to emphasize from the start, however, that the scope of practical application of CPLX is not limited to small tables; considerable experience has already been acquired in applications to tables on the order of 500 cells. Further discussion of this point is included in section 4.

As noted earlier, there is a wide literature on the log-linear model. Books on the subject include Goodman (1977), Bishop, Fienberg, and Holland (1975), Fienberg (1977), and Haberman (1978, 1979). Haberman (1974) gives a basic source for theoretical foundations, although the book cannot be read as an elementary introduction. Introductory articles of shorter length have been given by Fienberg (1970) and Madgison (1977).

3. COMPLEX SAMPLE DESIGNS

CPLX does not represent an improvement on maximum likelihood theory for the case of the multinomial distribution; rather, it allows analysis of data under other sampling distributions for which maximum likelihood theory, if applied mechanically under an incorrect assumption of the multinomial distribution, would give unacceptable answers. This section attempts to give an elementary description of the sorts of situations in which the methodology implemented in CPLX is preferable.

Three basic elements serve to identify the complex samples where CPLX may provide a more appropriate answer: clustering, stratification, and weighting. Except for some special cases, CPLX should be applied when one or more of these elements is present in the sampling mechanism underlying the observed data. This section first discusses characteristics of clustering separately, since this is sometimes the only element of the three found in many instances where its effect on the properties of maximum likelihood theory is nonetheless severe. The issues of stratification and weighting are then discussed in the following section, since they often occur together.

3.1 Situations Involving Clustering

Clustering denotes the wide variety of situations in which the observations may be grouped into sets within which it is not possible to assume independence. Any of the household surveys conducted by the Census Bureau have this property, since groups of neighboring housing units are included in the sample together in order to reduce the costs of travel. Survey designs involving two or more stages of selection, in which the universe is divided into primary units and a first stage of selection of primary units is followed by additional stages of selection, may be considered to be clustered samples as well. Here, the sampled primary units form the clusters for purposes of estimating sampling variablility.

The technical descriptions of large sample surveys usually remark on the complexities of the sample design and would lead the reader to recognize the clustering in the design. A number of common simpler situations lead to clustering as well, and since the clustering is more subtle, some of these will be mentioned here.

A common situation of clustering occurs in otherwise simple random samples when units of the sample may contribute more than one observation to the cross-classification. For example, it would be possible to take a sample from the phone phone book which would represent, for all practical purposes, a simple random sample of households from this universe. Data collected on a household basis could be appropriately analyzed as if from a multinomial distribution. At the same time, collection of data on more than one person in the household would lead to a clustering effect. For example, very strong clustering effects would occur in the collection of information on the public or private school enrollment for all of the children in the household.

Longitudinal or panel surveys often lead to instances in which an analyst may wish to consider including a person in more than one cell of a table. If the table cross-classifies the responses of individuals to the same variable at different points in time, the table may count each individual only once. If several variables are involved, however, and time is simply included as a separate variable, each individual may contribute a count for for each level of the time variable. In these cases, clustering arises from the probable correlation in results for the same person over time.

CPLX may also be of interest in the analysis of designed experiments, such as experiments on human populations, in which the study population is recognizably grouped into larger clusters, even though the principal analytic interest is in a logistic model for the individual outcomes. For example, a number of hospitals or physicians participating in a long-term clinical trial of different medical interventions for the same condition may be considered as giving rise to a clustered sample, even though the analysis may be of survival or health of individual patients using a number of risk factors as covariates. The possible effects of hospital or practitioner lead to dependencies in the observations that are frequently ignored. CPLX represents one approach to handle such difficulties more robustly.

In short, some sampling situations yield exactly or almost exactly a multinomial distribution for the observed proportions. The number of cases in which some clustering is present,

however, is probably much larger than is recognized or accounted for in analysis. CPLX or other methods that explicitly account for such dependencies could provide more robust statistics for many of these problems.

3.2 Stratification and Weighting

Stratification covers a large class of circumstances in which some data is known about the universe prior to sample selection, the universe is grouped into strata on the basis of these data, and samples are selected from the strata separately. In many applications, stratification actually reduces variance relative to simple random sampling.

In one special case, the strata are identical to the levels of one of the variables in the analysis. As long as multinomial samples are selected from each, maximum likelihood theory gives generally the same results as for the simple multinomial distribution, for most log-linear models. In other cases, however, stratification may have effects on the variance that are omitted from the standard maximum likelihood analysis.

On occasion, samples are selected at different sampling rates from the individual strata. A common rationale for differential probabilities of selection is to increase the reliability for special subgroups of the population. In order to represent consistently the original population, weights are typically applied to the observations, usually the inverses of the probabilities of selection or closely related quantities. Differential weights make any of the results from maximum likelihood theory for the multinomial distribution difficult to interpret without further adjustment. Once an appropriate representation of the sample design is found in terms of replicate observations, weighted data presents no additional difficulty to CPLX.

3.3 Representation of Sampling Variability Through Replication

CPLX evaluates the reliability of the sample estimates from a series of replicates provided by the user. The documentation (Fay, 1982) describes these requirements in detail. Basically, three replication options are offered:

1. Standard jackknife - The total sample is divided into a number of clusters which represent equivalent estimates of the same population total (even though they may be individually highly variable).

2. Half-sample - The estimated cross-classification is shown estimated for a series of subsamples of one half of the original data, each mimicking the process of sample selection of the original sample.

3. Stratified jackknife - Again a division of the sample into clusters, this time grouped by strata. The separate replicates within a stratum are assumed equivalent representations of the stratum total, but no relationship is presumed among the totals for the separate strata.

The program documentation presents a detailed discussion of how to use these strategies effectively, and an equally broad discussion is beyond the scope of this paper. In short, however, the principle is to confine all dependent sources of variation within replicates, so that comparisons across replicates represent true sources of independent variation in the design, upon which an inference may be based. As an example, in a simple clustered sample, the individual clusters could be chosen as replicates for the simple jackknife method. The user would tabulate the data for each of the clusters for input to CPLX and evoke the simple jackknife option. CPLX will use the variability among the cluster results to measure the reliability of the outcome for the total sample. On the other hand, CPLX will rely on no further assumption about variability within the replicates in drawing inferences; consequently, the dependencies within the clusters may be quite complex without affecting the validity of the conclusions. (Section 6 of the documentation discusses the choice of replication method; section 6 of this paper outlines some of the different approaches that may be used to create the replicate tables for input to CPLX.)

4. COMPARISON OF THE METHODOLOGY OF CPLX WITH WLS

Earlier available software for analysis of log-linear models in the context of complex sample designs appears to have been restricted to the weighted least squares (WLS) approach. (Work of Scott and Rao (1981) and others on applying generalized design effects is omitted here, since implementation of their methods by general software has been relatively recent. Their statistical methodology also does not yet address all of the questions of inference treated both by WLS and by CPLX.) Grizzle, Starmer, and Koch (1969) developed the basic foundation for the WLS applied to the analysis of cross-classified data for simple sample designs; Koch, Freeman, and Freeman (1975) later extended the results to complex samples.

Within its area of proper application, the WLS approach provides all three of the inferential components enumerated in section 2: standard errors of parameters, tests of overall fit, and tests of the contribution of specific groups of parameters to the model. WLS has also been successfully implemented and applied to survey data. (To cite just some of the relevant literature in this area, the GENCAT program (Landis, Stanish, and Koch, 1976) represents a standard developed by some of the principal statistical researchers in this field. A more recent paper of Lepkowski, Bromberg, and Landis (1982) discusses the implementation of the methodology in the OSIRIS system of the University of Michigan. Cohen and Gridley (1982) have reviewed some of the problems and limitations with this type of

analysis.)

As the newcomer in this area, CPLX draws justification for its separate existence on the basis of the significant limitations of WLS in treating tables of moderate to large size. This limitation of WLS arises from two sources: an analogous weakness of WLS relative to maximum likelihood estimation (MLE) in multinomial sampling for an important class of models; and the typical difficulty in estimating a covariance matrix for a table of this size that is neither singular or ill-conditioned for most complex survey designs, in which the number of first stage units is generally not too large. The remainder of this section elaborates on these two weaknesses of WLS with the intent of enabling the reader to recognize situations in which CPLX provides a methodologically superior solution.

4.1 Methodological Limitations of WLS Relative to MLE for Large Tables

For multinomial samples, MLE possesses asymptotic advantages over WLS for all models except the saturated (fully-parameterized), in which case the two methods agree. For small tables, these differences are almost always subtle and, for practical purposes, insignificant. For large tables, however, there is an important class of models and applications for which the gap between the capabilities of the two methodologies widens enormously. Haberman (1977) presented an underlying asymptotic theory for MLE for this class of models, and his paper forms the basis for the comments here. The models considered by Haberman possess the important property that typically the standard errors for all of the estimated parameters, or for a specifically identified subset of the parameters, still perform well under the MLE procedures, even for large, sparse tables. For this same class, likelihood ratio chi-square tests for the contribution of specific sets of parameters are also approximately correct. Even though in these situations tests for overall fit may be unreliable, the ability to estimate safely specific parameters and their standard errors, and to perform tests of significance for their contribution, enables a great deal of interpretation of such data. In the same situations, the WLS methodology is typically unable to provide any useful results.

Exact description of this important class of models is complex, and Haberman (1977) supplies the precise definitions. For practical purposes, however, the examples of section 2 illustrate the models of this class likely to be of most interest. The models fall into two principal types: ones like models (2.1) and (2.3), which avoid any interactions of high order, and the logistic models like examples (2.5) and (2.7), which include the high-order interaction of the independent variables, but simple relationships between the dependent and independent variables.

In models (2.1) and (2.3) of section 2, the parameters pertain simply to the levels of one of the variables or, in the case of example (2.3), to an interaction between just two variables. In many cases, each of the marginal tables fitted under the model, including the two-way cross-classification of variables 1 and 2 in model (2.3), may be well filled by observed values, even though zeroes and other small values may be present in the complete cross-classification. (This situation becomes even more typical when a larger number of variables are present in the cross-classification.) In this case, the MLE estimates of all of the parameters in these models and their estimated standard errors approximate standard asymptotic behavior. Likelihood ratio test of comparisons of models of this type, such as the test of the contribution of the parameters for the interaction between variables 1 and 2 by comparing the fitted values under models (2.1) and (2.3), are also reliable in this case. Thus, one can study refinements on the independence model, such as additions of simple relationships between pairs of variables, even though the test of overall fit suffers in such sparse tables.

In the case of logistic modeling, as illustrated in section 2, the log-linear model includes the complete interaction of all independent variables. Quite often, the corresponding marginal table (formed by adding across the dependent variable or variables, only) will contain a number of zero cells. At the same time, the typically simpler two- or three-way marginal tables corresponding to parameters relating the dependent variable(s) to the independent variable(s) may be well filled. (This case is more complex than the preceding in that some sparse marginal tables, those pertaining to the independent variables by themselves, are involved.) In this case, the parameters relating the independent variables to each other may be poorly behaved, while at the same time the critical parameters relating the dependent and independent variables (corresponding to the coefficients of the logisitic regression) will be well behaved and have reliably estimated standard errors. The tests of significance using the likelihood-ratio chi-square for the contribution of these latter parameters (such as the comparison of models (2.5) and (2.7) earlier) will also be reliable.

Although the chi-square test of overall fit may deteriorate in large tables, maximum likelihood theory still permits most useful analysis to continue through estimation of parameters, standard errors, and tests of the contribution of specific groups of parameters to the model. WLS, on the other hand, may not be safely applied in these instances or to any other large table with any significant number of sample zeroes. Thus, it is essentially impossible for WLS to be successful in analogous situations in complex samples. CPLX combines the MLE procedures with replication methods in order to exploit the advantages of MLE while at the same time recognizing the implications of a complex sample design.

4.2 Problems in Estimating Covariances in Large Tables

The other fundamental problem with WLS applied to moderate to large scale tables is a practical one. The WLS methodology requires inversion of an estimated covariance matrix for the observed sample values. At a minimum, therefore, the estimated covariance matrix must be non-singular or essentially so. (Possible modifications to the methodology could recognize the effect of fixed totals, etc.) Furthermore, strict algebraic non-singularity of the estimated covariance is not sufficient; ill-conditioned matrices would tend to produce aberrant statistical behavior. Thus, when using the WLS method, the cautious practitioner may prefer to employ two or three times the number of replicates as cells in order to insure a stable estimate of the covariance.

When the number of cells in the table is small, these objectives can usually be achieved readily, but as the size of the table grows, so too does the problem of obtaining a sufficiently stable estimate of the covariance. In many applications, the degrees of freedom will not be present in the design to meet these requirements for analysis of large tables, although in the same contexts WLS may be perfectly adequate for the analysis of much smaller tables.

4.3 Summary Comparison of WLS and the Methodology of CPLX

The previous comparisons have emphasized the difference between the two methodologies principally in terms of the size of the table under analysis. In order to quantify the net effect of these comparisons, table 1, at the risk of some over-simplification, attempts to convey an idea of where these boundaries might be in practical applications. The three categories chosen, 4 to 20, 21 to 80, and 81 to 1000, denote a range where the WLS methodology is generally successful, sometimes successful, and rarely successful, respectively. Although CPLX also does best for the smallest category, the deterioration in performance with table size is much less severe. (Even the value of 1000 cells is not necessarily the upper limit of application of CPLX.)

5. METHODOLOGY IMPLEMENTED IN CPLX

This section summarizes a more complete discussion in section 7 of the program documentation.

The program converts the cross-tabulated data provided by the user for the fractions of the sample or the series of half-sample replicates into a common form. Representing the cross-classification for the total sample as $\underset{\sim}{Y}$, a series of replicates $\underset{\sim}{Y} + \underset{\sim}{W}^{(i,j)}$ is considered with

$$\sum_j \underset{\sim}{W}^{(i,j)} = 0 \qquad (5.1)$$

for each i, and constants b_i are determined such that

$$\mathrm{Cov}^*(\underset{\sim}{Y}) = \sum_i b_i \sum_j \underset{\sim}{W}^{(i,j)} \otimes \underset{\sim}{W}^{(i,j)} \qquad (5.2)$$

i.e. an appropriately weighted sum of cross-products of $\underset{\sim}{W}^{(i,j)}$ with itself, gives the standard unbiased estimate of the covariance matrix for the estimate $\underset{\sim}{Y}$. (In fact, the computation (5.2) is never performed in CPLX, but is given here in order to identify the b_i's.)

As an example of this notation, in the case of the simple jackknife, the user provides a series of tables $\underset{\sim}{Z}^{(j)}$ with

$$\underset{\sim}{Y} = \sum_j \underset{\sim}{Z}^{(j)} \qquad (5.3)$$

In the general notation, i is fixed at 1 for the simple jackknife option. The program computes, in effect

$$\underset{\sim}{W}^{(1,j)} = (n-1)^{-1} (\underset{\sim}{Y} - n\underset{\sim}{Z}^{(j)}) \qquad (5.4)$$

and $b_1 = (n-1)/n$. Although these expressions do not give the standard representation of the simple jackknife, (5.2) yields the same estimate of covariance as usually computed.

If $\underset{\sim}{\lambda}(\underset{\sim}{Y})$ represents an estimated log-linear parameter based on the full sample, and $\underset{\sim}{\lambda}(\underset{\sim}{Y} + 1/2\ \underset{\sim}{W}^{(i,j)})$ the corresponding result for for a modified replicate table, the program computes

$$P_{ij} = \underset{\sim}{\lambda}(\underset{\sim}{Y} + 1/2\ \underset{\sim}{W}^{(i,j)}) - \underset{\sim}{\lambda}(\underset{\sim}{Y}) \qquad (5.5)$$

$$\mathrm{Var}^*(\lambda) = 4 \sum_i b_i \sum_j P_{ij}^2 \qquad (5.6)$$

and, optionally, a jackknifed estimate for the parameter as

$$\lambda_J = \underset{\sim}{\lambda}(\underset{\sim}{Y}) - 4 \sum_i b_i \sum_j P_{ij} \qquad (5.7)$$

(The factors of 1/2 and 4 appear in the formulas for related reasons: the factor of 1/2 avoids creating replicate tables with additional zeroes not present for the original cross-classification, and the factor of 4 compensates for this effect in the estimate of the covariances.)

When zeroes are present in the fitted values, the entire set of computations are carried out by

adding small constants to the cells of the tables before fitting the models. The computations are done in such a way that any first-order bias induced in the estimates is removed in the computation (5.7).

For tests of overall fit or of the contribution of specific parameters, CPLX computes jackknifed chi-square tests described in Fay (1980, 1981). For the jackknifed test of overall fit of a model, if $X^2(\underset{\sim}{Y})$ denotes the Pearson chi-square test for the full sample, and $X^2(\underset{\sim}{Y} + \underset{\sim}{W}^{(i,j)})$ the corresponding test for each replicate, CPLX computes

$$P_{ij} = X^2(\underset{\sim}{Y} + \underset{\sim}{W}^{(i,j)}) - X^2(\underset{\sim}{Y}) \qquad (5.8)$$

$$K = \sum_i b_i \sum_j P_{ij} \qquad (5.9)$$

$$V = \sum_i b_i \sum_j P_{ij}^2 \qquad (5.10)$$

$$X_J = \frac{(X^2(\underset{\sim}{Y}))^{1/2} - (K^+)^{1/2}}{(V/(8\ X^2(\underset{\sim}{Y})))^{1/2}} \qquad (5.11)$$

The last expression gives the jackknifed chi-square test statistic, which may be compared to tabulated critical values provided in either of the previous sources, or to a more extensive set included in the program documentation.

A similar computation is employed to test the contribution of specific parameters to a given model. If $G^2(\underset{\sim}{Y})$ represents the likelihood ratio chi-square test statistic based on the entire sample, $G^2(\underset{\sim}{Y} + \underset{\sim}{W}^{(i,j)})$ the test based on a replicate table for a model that excludes the set of parameters in question, and $G^{2\prime}(\underset{\sim}{Y})$ denotes the value of the test when the parameters are added to the model, CPLX computes

$$P_{ij} = (G^2(\underset{\sim}{Y} + \underset{\sim}{W}^{(i,j)}) - G^2(\underset{\sim}{Y})) - (G^{2\prime}(\underset{\sim}{Y} + \underset{\sim}{W}^{(i,j)}) - G^{2\prime}(\underset{\sim}{Y})) \qquad (5.12)$$

and carries out the analogous computations (5.9), (5.10), and

$$G_J = \frac{(G^2(\underset{\sim}{Y}) - G^{2\prime}(\underset{\sim}{Y}))^{1/2} - (K^+)^{1/2}}{(V/(8\ (G^2(\underset{\sim}{Y}) - G^{2\prime}(\underset{\sim}{Y}))))^{1/2}} \qquad (5.13)$$

which gives the jackknifed chi-square test for the comparison of two different models. This test statistic may then be compared to the same table of critical values as the overall test of fit (5.11).

6. NOTES ON THE USE OF CPLX

CPLX reads two input files provided by the user: a set of control cards providing basic information about the table and replication strategy, followed by requests for statistical operations; and a separate file containing a series of tables whose contents depend upon the replication method selected. This section will describe these two files separately.

6.1 Content of the Control Cards

The user must provide in the control deck two required pieces of information: the dimensions of the table and the replication method chosen. If the stratified jackknife option is used, the number of fractions in each stratum is also required. Additionally, the user may optionally affect the analysis, display of results, or computational methods in any order in any of the following ways:

- provide labels for variables, levels of variables, and/or coefficients
- change the method of defining the parameters of the log-linear model for variables of more than two levels, enabling, for example, testing of specific contrasts
- change the default parameters governing the convergence criteria used in fitting the models
- enter a starting matrix for the iterative proportional fitting algorithm (chiefly to enter a matrix of 1's and 0's to exclude structural zero cells from the model)
- enter comments in the printed output
- restrict the display of parameters to those involving one or more declared "dependent" variables, for use with logistic regression.

The basic statistical requests are

- fitting a model to compute jackknifed chi-square tests of the overall fit and, optionally, display the observed and fitted values under the model
- compare two models previously fitted to test the contribution of the parameters by which one differs from the other (restricted to models in which one of the two implies the other)
- compute and display parameters, with or without the bias correction (5.7), along with standard errors.

6.2 Replicate File

The other file is a series of cross-classifications, each of the size specified in the control deck, where the cell entries vary in FORTRAN order. For the jackknife methods, the tables are estimates based on the separate clusters or fractions of the design; CPLX sums these to derive the table for the total sample. For the half-sample replication methods, the user must provide the table based on the total sample as the first table in the file, and the remaining tables are each interpreted as half-sample estimates.

Actual preparation of a replicate file for CPLX may be accomplished in a number of ways. One is direct tabulation of the required tables by a FORTRAN program, followed by output in the specified format. This method is the only available method to create half-sample or bootstrap replicates known to the author. When the simple or stratified jackknife options are used, however, users may employ systems such as SPSS or SAS to manage their data. Either system will create output files summarizing tabulation results, and the program documentation for CPLX lists short FORTRAN programs that may be used to interface these results to CPLX.

REFERENCES

Bishop, Yvonne M. M., Fienberg, Stephen E., and Holland, Paul W. (1975), Discrete Multivariate Analysis, Cambridge, MA.: MIT Press.

Cohen, Stephen B., and Gridley, Gloria (1982), "Present Limitations in the Availability of Statistical Packages for the Analysis of Complex Data," Proceedings of the Computing Section, 1981, Washington, D.C.: American Statistical Association, 20-24.

Fay, Robert E. (1980), "Further Results in Jackknifing Chi-Square Test Statistics," presented at the annual meetings of the American Statistical Association, Houston, Texas, 1980.

_____ (1981), "On Jackknifing Chi-Square Test Statistics - Part I: Description and Applications of the Method," unpublished manuscript to be revised for resubmission to the Journal of the American Statistical Association.

_____ (1982), "Contingency Table Analysis for Complex Sample Designs (CPLX): Program Documentation," unpublished.

Fienberg, Stephen E. (1970), "The Analysis of Multidimensional Contingency Tables," Ecology, 51, 419-433.

_____ (1977), The Analysis of Cross-Classified Categorical Data, Cambridge, MA: MIT Press.

Goodman, Leo A. (1978), Analyzing Qualitative/ Categorical Data, Cambridge, MA.: ABT Associates Inc.

Grizzle, J. E., Starmer, C. F., and Koch, G. G. (1969), "Analysis of Categorical Data by Linear Models," Biometrics, 25, 489-504.

Haberman, Shelby J. (1974), The Analysis of Frequency Data, Chicago, Ill.: The University of Chicago Press.

_____ (1977), "Log-Linear Models and Frequency Tables with Small Expected Cell Counts," Annals of Statistics, 5, 1148-1169.

_____ (1978), Analysis of Qualitative Data: Vol. 1, Introductory Topics, New York: Academic Press.

_____ (1979), Analysis of Qualitative Data: Vol 2, New Developments, New York: Academic Press.

Koch, G. G., Freeman, D. H. Jr., and Freeman, J. L. (1975), "Strategies in the Multivariate Analysis of Data from Complex Surveys," International Statistical Review, 43, 59-78.

Landis, J. R., Stanish, W. M., and Koch, G. G. (1976), "A Computer Program for Generalized Chi-Square Analysis of Categorical Data Using Weighted Least Squares to Compute Wald Statistics (GENCAT)," Department of Biostatistics Technical Report No. 8, Department of Biostatistics, University of Michigan, Ann Arbor, MI.

Lepkowski, James M., Bromberg, Judith A., and Landis, J. Richard (1982), "A Program for the Analysis of Multivariate Categorical Data from Complex Samples," Proceedings of the Computing Section, 1981, Washington, D.C.: American Statistical Association, 8-15.

Rao, J. N. K., and Scott, A. J. (1981), "The Analysis of Categorical Data from Complex Sample Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables," Journal of the American Statistical Association, 76, 221-230.

APPENDIX

A 1 Source Language and Maintenance

The original version of CPLX is written in FORTRAN 77 and may be compiled on level 9 or 10 of Univac ASCII FORTRAN and presumably, by other FORTRAN 77 compilers for large mainframe computers. The blocked IF-THEN-ELSE structures of the language appear extensively in the code.

The FORTRAN 77 version also contains a small number of special comment cards. A separate preprocessor program, itself in FORTRAN 77, is able to interpret these comment cards as instructions for conversion to FORTRAN IV. The preprocessor reformats and punctuates the CPLX source into an intermediate structured FORTRAN language, SFOR, developed by Emmett Spiers and

Donald Dalzell of the Census Bureau. In turn, the SFOR processor produces a Univac FORTRAN V source that, under these circumstances, is compatible with IBM FORTRAN IV. A source generated in this manner compiles and executes correctly under either FORTRAN G or H at the computing facility of the National Institutes of Health. (The derived source also correctly compiles under Univac FORTRAN V at the Census Bureau and may also compile on FORTRAN implementations of other manufacturers.)

FORTRAN 77 represents a current standard recognized by the Federal government. The syntax of the language facilitates a reasonably structured coding of the original source and offers portability to other systems supporting FORTRAN 77. On the other hand, a FORTRAN IV source compatible with an IBM compiler was a practical necessity, even for some applications within the Census Bureau. The current method of deriving the second set of source code from the first insures consistent additions of enhancements and maintenance on both versions simultaneously.

A.2 Numerical Accuracy

The requirements for numerical precision by CPLX are generally more severe than most applications of log-linear models, since comparatively subtle changes in chi-square tests and parameter values are evaluated over a series of replicates. As a partial response to this problem of numerical accuracy, all real arithmetic and storage of results is in double precision.

The global choice of double precision arithmetic represents a satisfactory solution to many of the numerical problems that would have otherwise been encountered. The most critical remaining question, however, is that maximum likelihood estimation of many (but not all) log-linear models by either the Newton-Ralphson algorithm or iterative proportional fitting does not achieve exact convergence after any finite number of steps. Choice of the stopping point, therefore, is a potentially important numerical issue.

CPLX automatically evaluates the effect of the stopping rule on all jackknifed chi-square tests by conducting all computations and reporting all results under two separate stopping rules, in parallel. The stopping points are defined as the allowance for a maximum deviation in any one step between the fitted marginal tables of the observed data and the current estimated cell values. Under the default option of the program, these two stopping rules differ by a factor of 1000. Consistent results under the two separate rules are a strong assurance that sufficient numerical accuracy is present, whereas significant differences warn the user that further consideration is required. A change of stopping points may be accomplished by appropriate control cards, if more precision is required. In the author's experience, however, the pair of default parameters provided with the program is sufficient to provide a valid test of the numerical precision and to assure, in almost all cases, that acceptable numerical precision is obtained.

The computation of standard errors for the parameter estimates poses relatively less severe numerical problems, and parallel computations under two different stopping rules is not automatically provided. (Instead, convergence is required to the more stringent of the two criteria in force.) The user concerned about this point could accomplish the same effect of parallel computations by computing the parameters under the default option, changing the stopping points, and then recomputing the parameters.

A.3 Limitations

The design of the system of control cards limits the number of levels (categories) of any one variable to 99 and the number of variables to 39. The second limit is beyond any reasonable application; the first limitation may be circumvented by an intrepid user through relatively few modifications of the source code.

Specification of the remaining limitations of the program is somewhat more difficult and less precise. CPLX internally allocates and deallocates all arrays of variable requirements in size from three named common areas, one each of integer, double precision, and character. The sizes of the integer and double precision arrays impose practical limitations on the applications that may be handled by CPLX. The present version assigns 10000 cells to the integer array and 6000 to the double precision; this space is adequate for complete analysis (including estimation of parameters) of most tables of up to 500 cells and for jackknifed chi-square statistics of tables up to about 1000 cells. Provided that the space is available, larger tables may be analyzed by increasing the size of these two arrays as necessary, which involves changing a limited number of lines of the source for CPLX.

Table 1 General Comparison of WLS and CPLX for Tables of Differing Size from Complex Sample Surveys

Size of Table	WLS	CPLX
4 to 20 cells	Generally adequate	Generally adequate
21 to 80 cells	Increasing difficulty: sometimes adequate but often zero cells pose problems	Usually adequate. A relatively large number of zero cells may impose same restrictions as for tables of size 81 to 1000 cells
81 to 1000 cells	Rarely possible	Occasionally entirely adequate, more often restricted to: · tests of contribution of specific subsets of parameters · standard errors of parameters for: · logistic models · parsimonious models