CATEGORICAL DATA ANALYSIS IN SURVEY SAMPLING

By

ANIL RAI

A Thesis
submitted to the faculty of Post - Graduate School
Indian Agricultural Research Institute, New Delhi,
in partial fulfilment of the requirements
for the award of the degree of

DOCTOR OF PHILOSOPHY

IN

AGRICULTURAL STATISTICS

INDIAN AGRICULTURAL STATISTICS RESEARCH INSTITUTE
(I.C.A.R.)

LIBRARY AVENUE, NEW DELHI-110012
1991

CERTIFICATE

This is to certify that the work incorporated in the thesis entitled **CATEGORICAL DATA ANALYSIS IN SURVEY SAMPLING** by Shri Anil Rai and submitted in partial fulfilment of the requirement for the award of the degree of **DOCTOR OF PHILOSOPHY** in Agricultural Statistics of the Post-Graduate School, Indian Agricultural Research Institute, New Delhi was done under my guidance and supervision and that no part of this thesis has so far been submitted for any other degree or diploma.

The assistance and the help recieved during the course of this investigation have been duly acknowledged by him.

IASRI, New Delhi
Dec., 1991

(O.P. KATHURIA)
Chairman
Advisory Committee

Advisory Committee :

1. Dr. Randhir Singh (Member) ----------
2. Dr. P.R. Sreenath (Member) ----------
3. Shri M. KUMAR (Member) ----------

DEDICATED TO MY MOTHER

ACKNOWLEDGEMENT

I have a great pleasure in expressing my deep sense of gratitude to Dr. O.P. Kathuria, Principal Scientist, IASRI for his valuable guidance, keen interest, inquisitive ideas and constant encouragement through out the course of this investigation.

I am grateful to members of my advisory committee, Dr. Randhir Singh, Senior Professor, IASRI, New Delhi, Shri R. Gopalan, Head, Division of computer science, IASRI, New Delhi, and Dr. P.R. Sreenath, Principal Scientist, IASRI, New Delhi, who helped me to improve the quality by way of providing suggestions and constructive criticisms in the preparation of this thesis.

I wish to extend special thanks to Dr (Mrs.) Vimlesh Seth, Professor (Pediatrics), AIIMS, New Delhi for her constant encouragement throughout this programme.

I am thankful to Prof. Prem Narain, Director, IASRI, New Delhi for providing necessary facilities at IASRI to carry out this research work.

I am also extremely grateful to Dr. H.V.L. Bhatala and Shri K.K. Kher, IASRI, New Delhi providing the data required for this research work. The help provided by the staff of

computer division, IASRI, particulary Shri S.P. Doshi, Shri P.L. Gupta and Shri V.N. Chakravarti are gratefully acknowledged.

The work was supported by Senior Research Fellowship of IASRI, New Delhi. the help is gratefully acknowledged.

I Wish to acknowledge the association of Dr. P.S. Pandey, Mr. D.V.V. Ramana, Mr. S.K. Diwedi, Mr. R.K. Shukla, Miss. Seema Jaggi, other friends and hostel inmates for their encouragement, cooperation and help at various stages of research work.

I express my heartful thanks to my father, Dr. S.D. RAI, and mother, Late Smt. Vidhya Rai, wife Smt. Suman Rai and brother Shri Arun Kumar for their constant moral support and inspiration during the period of my studies. I am also thankful to my son master Abhishek Rai for his cooperation during this research work.

I thankfully acknowledge the excellent typing assistance recieved from M/s Ankur Electrostat & Stationary.

(ANIL RAI)

CONTENTS

INTRODUCTION

INTRODUCTION

1.1 INTRODUCTION :

Standard statistical methods were developed on the assumption that elements of the population under study from which samples are drawn are identically and independently distributed. These properties can be ensured when samples of elements are either drawn from infinite population or with the help of simple random sampling with replacement, in the case of finite population. The assumption of independent selection of elements (and hence independence of observations) greatly facilitates in obtaining theoretical results of interest. Also, the assumption of independence yields mathematical simplicity that becomes desirable as we move from simple statistics such as mean to complex statistics such as ratios, proportions etc. Independence is often assumed automatically and needlessly, even its relaxation would permit broader conclusions.

Although independence of sample elements is typically assumed, it is seldom realized in the procedures of practical survey work. Randomization of the sample would be unnecessary if the population itself were randomized, but "well mixed urns" are seldom provided by nature or created by man.

One of the most striking features of developments in statistics is the rapid growth of interest in sampling methodology covering both theory and practice. Survey sampling has significantly helped in providing estimates of important

population characteristics for scientific planning. Early landmarks in the literature of sampling theory were papers byTschuprow(1923), Neyman(1934), Mahalanobis(1944), and Yates (1946). Five classics in five years were Yates(1949), Deming (1950) Cochran (1953), Hansen et al.(1953) and Sukhatme (1954) who outlined the boundaries that have broadly defined the developments of methods in this subdiscipline of statistics.

Generally, the literature of survey sampling concentrates on providing estimates of simple statistics like mean and its standard error and consequently confidence intervals. Although recently methods for estimation of complex statistics from complex survey designs were developed but still there is need of further refinement and development. Inferences based on standard errors are acceptable on the assumption that survey samples are large enough to yield the needed approximate normality in spite of non-independence of observations. Standard errors should be computed in accord with the complexity of sample designs, neglect of that complexity may be source of serious bias. On the other hand, trying to obtain more exact but complicated statistics than standard errors would become too difficult for complex selection designs.

1.2 BASIC PROBLEMS IN CATEGORICAL DATA ANALYSIS :

There are mainly two types of problems encountered in categorical data analysis. These are (i) Measures of association, by which the degree of relationship between any

two variables can be measured (ii) Testing of hypothesis, under which various hypotheses of interest can be tested. The development in case of measures of association is confined to some specific areas, where as, testing of hypothesis in categorical data analysis is widely applied. The categorical data analysis in survey sampling is mostly confined to the testing of hypothesis because of its practical utility to study the structure of the population under consideration and drawing inferences accordingly. Broadly, there are three types of inferential problems in case of categorical data analysis from survey sampling.These are :

(i) To test whether the sample proportion is equal to a certain fixed proportion.

(ii) Testing of independence among categorical variables,

(iii) To test the homogeneity among a set of proportions.

The above mentioned problems generally occur in many large scale surveys. For example, and investigator may wish to compare sample proportions for categories of variables such as proportions of farmers growing high yielding varieties of crops with known population proportions from the previous surveys. This can also be used to check the quality of sampling as used by Brackstone and Gosseline (1973). More generally, we might make comparisons among several different surveys from the same population or among different regions of

a country or among similar surveys from different countries of the world. The problem of testing independence among

categorical variables is very common in almost all surveys.

1.3 TESTING THE HYPOTHESIS :

Karl Pearson in 1900 introduced goodness of fit test which had a revolutionary impact on categorical data analysis as it was the first inferential statistical method. This test statistic was susequently used for testing the independence of attributes. Later on likelihood-ratio test also gained popularity for same type of testing problems and it was proved that this test also follows chi-square distribution with the same degrees of freedom as original chi-square test. General method of application of these tests are discussed below.

1.3.1 CHI-SQUARE TEST AS GOODNESS OF FIT :

Consider the null hypothesis (Ho) that k parameters $\{r_i\}$ of a multinomial distribution are equal to certain fixed value $\{r_{io}\}$, where $\sum_{i=1}^{k} r_{io} = 1$. Under Ho, $m_i = n\ r_{io}$, $i=1,2,\cdots,k$ where m_i is the expected number of observations in the i-th category. For the sample counts $\{n_i\}$, Pearson proposed the following test statistic ;

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - m_i)^2}{m_i}$$

For large samples, X^2 has approximately, a Chi-squared null distribution, with k-1 degrees of freedom. The above statistic is called Pearson Chi-squared statistic.

For a multinomial sampling of sample size n, n_i has the binomial distribution with index n and parameter r_i. For large n, by normal approximation to binomial, n_i (and $p_i = n_i/n$) has approximate normal distribution. More generally, by the central limit theorem, the sample proportions $p = (n_1/n, \ldots, n_{k-1}/n)'$ have approximate multivariate normal distribution. Let Σ_o denote the covariance matrix under null hypothesis of $\sqrt{n}\,p$, and let $r_o = (r_{1o}, \ldots, r_{k-1,o})$ Under, Ho, since $n(p-r_o)$ converges to $N(0, \Sigma_o)$ distribution, the quadratic form.

$$n\,(p-r_o)'\,\Sigma_o^{-1}(p-r_o)$$

has distribution converging to Chi-square with k-1 degrees of freedom. The covariance matrix of $\sqrt{n}\,p$ has elements

$$\sigma_{ij} = -r_i r_j \quad \text{if } i \neq j$$
$$= r_i(1-r_i) \quad \text{if } i=j$$

The matrix Σ_o^{-1} has (i,j)th element $\frac{1}{r_{ko}}$ when i#j and $\left(\frac{1}{r_{io}} + \frac{1}{r_{ko}}\right)$ when i = j. This can be varified by putting $\Sigma_o\Sigma_o^{-1} = I$ With this substitution direct calculation simplifies to X^2.

1.3.2 CHI-SQUARE TEST OF INDEPENDENCE :

In two way contigency tables with multinomial sampling, the null hypothesis of statistical independence is Ho: $r_{ij} = r_{i+}\,r_{+j}$ for all i and j. To test Ho, we could use Pearson X^2 statistics, with $m_{ij} = n\,r_{ij} = nr_{i+}\,r_{+j}$; where m_{ij} is the

expected number of observation in the (i,j) - th cell. Let $\hat{m}_{ij}$ = $np_{i+}\ p_{+j}$. The X^2 statistics then equals

$$X^2 = \sum_{j=1}^{J} \sum_{i=1}^{I} \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}$$

Pearson (1900, 1922) claimed that replacing $\{m_{ij}\}$ by estimates $\{\hat{m}_{ij}\}$ would not affect the distribution of X^2. Since, there are k=IJ categories for cross classification, he argued that X^2 would have asymptotically Chi-square distribution with df = IJ-1. On the contrary, since $\{m_{ij}\}$ are determined by estimating $\{r_{i+}\}$ and $\{r_{+j}\}$, the Chi-square distribution has

$$df = (IJ-1) - (I-1) - (J-1) = (I-1)\ (J-1)$$

The dimensions of $\{r_{i+}\}$ and $\{r_{+j}\}$ reflect the constraint $\sum_i r_{i+} = 1 = \sum_j r_{+j}$ Pearson's error was not pointed out until 1922 by R.A.Fisher, in an important article that helped to clarify geometrically the notion of degree of freedom.

1.3.3. LIKELIHOOD - RATIO TEST :

The likelihood ratio test is a general purpose way of testing a null hypothesis Ho against an alternative hypothesis Ha. In this test, the likelihood is maximized under Ho, and also under the condition that Ho or Ha is true. Let Δ denote the ratio of maximized likelihoods, which can not exceed. 1. Wilks (1935, 1938) showed that - 2log Δ has limiting chi-square distribution under Ho as $n \to \infty$. The degrees of freedom equal the difference in the dimensions of the

parameter spaces under HoUHa and under Ho.

For the multinomial sampling in a contingency table, the kernel of likelihood is $\prod_{ij} r_{ij}^{n_{ij}}$, where $r_{ij} > 0$ and $\Sigma_i \Sigma_j r_{ij} = 1$.

Under Ho, independence (all $r_{ij} = r_{i+} r_{+j}$), the likelihood is maximized when $\hat{r}_{i+} = \frac{n_{i+}}{n}$ and $r_{+j} = \frac{n_{+j}}{n}$, so $\hat{r}_{ij} = \frac{n_{i+} n_{+j}}{n^2}$

In the general case, the likelihood is maximized when $\hat{r}_{ij} = n_{ij}/n$

The ratio of the likelihoods equals

$$\Delta = \frac{\prod_i \prod_j (n_{i+} \; n_{+j})^{n_{ij}}}{n^n \prod \prod n_{ij}^{n_{ij}}}$$

It follows that Wilk's statistics, denoted by G^2, is

$$G^2 = -2 \log \Delta = 2 \sum_i \sum_j n_{ij} \log \frac{n_{ij}}{\hat{m}_{ij}}$$

where $\{\hat{m}_{ij} = \} \frac{n_{i+} \; n_{+j}}{n}$ are estimated expected frequencies under the assumption of independence. This statistic is called the likelihood ratio chi-square statistic. The larger the value of G^2, the more evidence there is against the null hypothesis. For large samples, G^2 has a chi-squired distribution under Ho with df = (I - 1) J - 1). When independence holds, the pearson statistic X^2 and likelihood ratio statistic G^2 have asymptotic chi-square distribution with df = (I-1) (J-1). In fact, X^2 and G^2 are asymptotically equivalent . In that case $X^2 - G^2$ converges in probability to zero.

1.4 METHODS OF SAMPLING IN CATEGORICAL DATA ANALYSIS :

All the methods of sampling in the case of categorical data can be broadly classified into three groups given below:

i) METHOD I : Many of the tables obtained from survey data are obtained by what is sometimes known as method I sampling. Here, the only frequency which is pre-set is the total frequency. Other frequencies in the table become known only after data have been collected and tabulated.

ii) Method II : In this method of sampling the sample sizes to be taken from several populations are fixed, i.e. marginal frequencies for one of the variables involved are fixed and classification on the other variables

becomes known later. Stratified sampling and cluster sampling give rise to tables of this type.

iii) Method III : Method III sampling occurs particularly, in controlled comparative trials. Marginal frequencies for one of the two variable typically a 'type of treatment' variable, are fixed in advance, but the selection of which units are to receive which treatments is made at random by the researcher, that is, classification on this variable is not an inherent property of the units, but something controlled by the researcher. This method of sampling might occur in studies of survey methodology, for example in the study to compare face-to-face and

telephone interviews, when we might fix the number of interviews of each type in advance and determine randomly, exactly who is to be given each type of interview.

The sampling scheme has a bearing on the theoretical model describing the table and on which methods of analysis are appropriate. Sampling method I is also known as multinomial sampling model, and sampling method II as product-multinominal, though strictly these terms apply only to simple random sampling with replacement.

1.5 EFFECT OF CLUSTERING, STRATIFICATION AND WIEGHTING :

In most surveys the assumption of independence is far from realistic as discussed above. Any large scale survey involves either stratified multistage or cluster sampling and correlations between units in the same cluster or stratum can have substantial impact. There are two reasons which are mainly responsible for wide spread application of cluster sampling. Although, the first intention may be to use the elements as sampling units, it is found in many surveys that no reliable list of the elements of the population is available and that it would be prohibitively expensive to construct such a list. Even when the list of individual units is available, economic considerations may point to the choice of a larger cluster unit. When cost is balanced against precision, the larger unit may prove superior. For example, a sample of farms independently and randomly selected is likely

to be scattered over the entire area, and thereby provides a better cross-section of the population than an equivalent sample of the same number of farms clustered together in few villages for which observations are likely to be correlated. On the other hand, it will cost more to survey a widely scattered sample of the farm than an equivalent sample of clusters of farms, since the additional cost of surveying a neighbouring farm is small as compared to the cost of both locating and surveying a second independent farm. Survey designs involving two or more stages of selection, in which the universe is divided into primary units is followed by additional stages of selection, may be considered to be clustered samples as well. Hence, the sampled primary units forming the clusters, are used for the purpose of estimating sampling variability. 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 list of the households of the village may be available and a sample from this may represent simple random sample for all practical purpose. Data collected on a household basis could appropriately analysed 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. As pointed out earlier due to the clustering any standard technique which will be applied to the data will not only give the distorted picture of the situation but also mislead the researcher appreciably.

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, the 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 distribnution, 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 many occasions, samples are selected at different sampling rates from individual strata. A common rationale of differential probabilities of selection is to increase the reliability for special subgroups of the population. In order to represent consistantly 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 designs is found in terms of replicate observations weighted data presents no additional difficulty.

1.6 DESIGN-BASED INFERENCE FOR LOG-LINEAR MODELES :

Common analytic models, such as linear regression, log-linear models, and generalized linear models were initially developed in the context of explicit stochastic models, for example, the normal or multinomial distributions. Developments in "robust" estimation avoid specific distributional requirements, but often maintain assumptions not typically encountered in survey sampling, for example, that the error terms of the models are independent and selected from a symmetric population. Many researchers familiar with one or more of these analytic models have applied them directly to sample survey data without recognition of the possible consequences of the sample design on the validity of inferences based on the usual distributional assumptions.

Log-linear models, which expresse the logarithm of the expected frequencies for categorical responses as a linear function of unknown parameters, encompass both factoral models for cross-classfied categorical data and logistic models for one or more dependent categorical or continuous predictors. Bishop, Fienberg and Holland (1975) provided one of the earliest books in this rapidly expanding field. Many log-linear models, particularly those for fully cross-classfied categorical data, involve a large number of parameters. The three most typical problems of inference are.

i) To compute standard errors and confidence intervals for the individual estimated parameters.

ii) To test the significance of the contribution of specific

sets of parameters to fit the model.

iii) To test the overall goodness-of-fit of the model. In the context of simple random samples, standard results in maximum likelihood theory provide an answer to these questions, although Pearson Chi-square test rightfully enjoys greater popularity than the likelihood-ratio chi-square test as a solution to the third problem.

The effects of complex sample designs on the analysis of categorical data have received considerable attention. Complex sampling designs typically seriously affect the Pearson or likelihood ratio chi-square tests for categorical data models. A number of alternative tests have been proposed under various sets of assumptions about the nature of complex design. Three approaches described below represents general solution:

1.6.1. THE WALD TEST : The wald test was proposed by Koch, Freeman and Freeman in 1975. It incorporates an estimate of the covariance matrix of the estimated cell frequencies into both estimation under the model and testing. Although this approach was the first of the three general solutions to be introduced, the specific manner in which the estimated covariance matrix of the estimated cell probabilities is incorporated leads to appreciable instability in many applications. Recently some modification to the wald test was proposed to lessen the effect of this source of variability.

1.6.2. ADJUSTMENT TO ORIGINAL CHI-SQUARE :

This procedure was proposed by Fellegi (1980) and Rao and Sott (1981, 1984, 1987) and employ the standard estimation methods of maximum likelihood estimation applied to the weighted. cell estimates as if they were cell counts from a multinomial distribution and adjustments to the usual chi-square test was done accordingly. Two principal forms of these procedures are available. In the first, relationships between the variance under the multinomial distribution and the sampling variance under the complex design are examined for both the cells and margins. These relationships are then incorporated into an adjustment factor by which chi-square tests are divided. The procedure compares the resulting chi-square test to the chi-square distributed on the same number of degrees of freedom as appropriate under multinomial sampling. The method is particularly useful in applications to the published tables if the required information about the variances under the complex design for the cells and marginal table is also available.

The second method proposed by Rao and Scott incorporates an estimate of the covariance matrix under the complex design for the cells of the estimated cross-classification. In practice this method requires returning to the original data to compute the estimated covariance matrix, since such matrices are rarely published. The method is again based on an adjustment to the original chi-square tests, but in this case both the test statistics and the original degrees of freedom are altered and interpreted according to an approximation due

to Satlerthwaite (1946). The second method requires more extensive calculations than the first, but its performance is sufficiently superior that it clearly represents the preferred method over the first, when such calculations are possible.

1.6.3. THE JACKKNIFED CHI-SQUARE TEST :

It was proposed by Fay(1985) and employs replication to determine the effect of complex sampling design on the original chi-square or likelihood tests. This method also employes standard maximum likelihood estimators applied to the observed cell estimates, as do the methods of Rao and Scott. The behaviour of the chi-square test recomputed according to a replication method reflecting the complex sample design is used to derive a new test of significance. In practice, an application of the Jackknifed test requires access to the original data in order to form the necessary replicate samples.

1.7 VARIANCE ESTIMATION AND INFERENCE :

Complex sample designs usually involve selection schemes that lead to statistical dependence among the selected sample units. Often, this dependence takes the form of a positive correlation between members of the sample. This positive correlation leads to negative bias in estimates of variance based upon the assumption of simple random sampling. To produce at least approximately unbiased estimate of variance, aspects of sample design must be taken into account in estimating variance. The following design elements have an

impact on the estimation of variance and will be discussed :

i) With or without replacement sampling

ii) Stratification

iii) Unequal probability of selection

v) Multiple stages of selection

v) Certainty selection of clusters

In spite of the fact that most complex sampling designs have a net increase in variance over simpler sample designs, certain aspects of the sample design can actually result in a decrease in variance. The use of without replacement sampling is one of these, under simple random sampling, the variance is reduced by a factor equal to one minus the sampling rate. In more complex sample designs, an analogous situation exists. For simplicity, estimates of variance are calculated under the assumption that with replacement sampling has been used because in without replacement sampling unbiased estimates of variance require known joint probabilities of selection of all sampled units. Formula based upon with replacement sampling require knowing only the probability for selection of each sampled units. Additionally, sampling past the first stage is not explicitly taken into account in with replacement formulas, since this is not necessary for obtaining unbiased variance estimates when with-replacement sampling of first stage units is used.

The simplicity of variance estimation under the assumption of with-replacement sampling has a price. When

with-replacement formulae are used for without-replacement designs, the variance estimates are biased. This bias is actually equal to two times the reduction in variance brought about by using without replacement sampling. This results in an interesting paradox. Without-replacement sampling is used to lower variance, but easiest approach to variance estimation results in treating estimates as if they were less precisely measured than would have been the case had with replacement design been used. Often, the first-stage sampling rate will be low enough, or first stage component of variance small enough, that is reasonable to assure this effect will not be too great.

Often, prior to selecting the sample, the population is grouped into strata based upon characteristics thought to be related to variable or variables of interest. This stratification usually reduces the variance of the population estimates. The stratification of the primary sampling units can be taken to the point where only two are selected from each stratum. This allows maximum stratification to be used while still allowing for the estimation of variance. If the stratification is ignored when variance is estimated, a positive bias will be introduced because the resulting estimate of variance will contain a "between Strata variance component. When stratification has been carried out to the point where only one unit is selected per stratum this "between strata" component is not easily avoided. In this situation similar strata are combined so that a "within

strata" variance component can be estimated with what is hoped to be a little, albeit positive bias.

The above discussion is for the estimation of variance for linear statistics. Most statistics of the interest are not linear. Generally, we apply approximation method for the variance estimation in the case of non-linear statistics. In summary, approximate variance estimators are required for the following reasons.

i) No explicit variance estimator is available because. design does not allow for one in the case of systematic sampling or one PSU per stratum designs.

ii) Adjustments have been made to sample weight

iii) The variance of non-linear estimator is desired

iv) It is too much trouble to use one of the exact formula.

1.7.1 ALTERNATIVE APPROXIMATE ESTIMATORS :

Two different approaches are currently in widespread use for the estimation of survey sampling errors for complex parameter estimators, namely linearization and replication. The method of linearization provides a general approach through the use of linear approximation to the nonlinear estimation of interest. Explicit formulas for the estimate of variance for these linear approximations can be desired. Variance estimation is achieved by estimating the variance of a linear combination of simple estimators whose variance is close to that of the complex estimator of interest. It is

important to keep in mind that the linearization approach does not actually yield an estimate of variance. Instead the linearization approach provides a linear approximation to the quantity for which variance is to be estimated, after which the usual text book formula for variance of a linear statistics is applied.

Replication (sample re-use) methods repeat the estimation process on a sequence of subsets of the full summary data set; and then compute the variance from the variation among these subsample estimates. The available replication methods differ as to their specification of sample subsets or replicates and subsequent variance estimation formulae. Three general approaches in use are known as balanced repeated replication (BRR) or balanced half-sampling, Jackknifing or Jackknife repeated replication, and bootstrapping. Each method has variations of application which affect the number of replicate estimates derived in a given case.

Even for relatively simple quantities like means and totals, typical survey estimators involve the use of non-response and ratio adjustment to the weighting, resulting weights that are random quantities, dependent upon the sample actually selected. A question to be addressed when comparing linearization with replication is relatively contribution of these adjustments to the variances of the parameters estimates. While linearization can be undertaken in a manner accounting for this weight variability, such variance

estimation does become cumbersome, whereas it remains relatively straight forward with replication. On the other hand, if the variablity in weights can be safely ignored, for many parameters estimated from surveys, linearization can be undertaken straight forwardly in a much less computationally intensive manner than replication. Kish and Frankel(1974) suggested that the contribution of such variation in weights to variation can reasonably ignored, whereas Lemeshow (1979) cautions against this. Lemeshow's finding from simulation studies suggested that a substantial increase in bias and variance of variance estimates could result from ignoring variability in weights.

1.7.2. INFERENCE :

A number of investigations have been conducted into the properties, both theoretical and empirical of linearization and replication. Though these studies did not investigate the effect of non-response and post-stratification adjustments, the results are still of some interest. Important among the empirical studies are the work of Kish and Frankel (1974) and Frankel (1971), who undertook a large scale empirical study comparing properties of linearization, BRR and the Jackknifing. They concluded that there was evidence that linearization gave some what greater accuracy (as measured by mean square error) in variance estimation, but that replication methods and in particular BRR, gave confidence interval coverage which has slightly closer to the nominal

coverage rate. Subsequently investigations have in the main concerned on the aspects of bias and precision of variance estimation. It was showed that Jackknifing in a number of forms and linearization were almost equivalent, while BRR was not nearly as equivalent to the other two procedures. These results were mainly for multistage designs in which two primary sampling units (PSU's) are selected independently per stratum. For ratio estimator, many studies concluded that all methods performed well when the coefficient of variation for the denominator, was below 10 percent, with a larger coefficient of variation for the denominator, BRR and the bootstrap became substantially positively biased, while linearization and Jackknife variance estimators showed slight negative bias.

One might regard the results of such investigations as indications that the less biased methods of linearization and Jackknifing are superior to BRR in terms of the resulting quality of variance estimation. Since the practical advantages and disadvantages of BRR are similar to those of the Jackknife, if this conclusion is well-founded then it seems that BRR should begin to lose favour. However, it must be remembered that the primary purpose of variance estimation in surveys is for making inferences about the parameters of the population, rather than about sampling errors. Thus, as suggested by Kish and Frankel (1974), the coverage of confidence intervals formed from variance estimates would seem to be of primary importance in assessing the relative merits

of variance estimation techniques. Such assessment involves consideration of the joint properties of the parameter estimate and its variance estimate, making investigation of this issue complex. Investigation suggests, in considering confidence intervals, BRR was some what superior to linearization and Jackknifing. Studies also indicate that use of the confidence interval coefficients derived from an appropriate t-distribution may improve confidence interval coverage,but that the use of the number of strata as the degrees of freedom may not always be appropriate.

Thus, in considering the relative qualities of these different methods of variance estimation, further developments and emperical investigation appear warranted. Such studies should also include consideration of bootstrap methods to assist in determining situations in which these present a better practical alternative than the established methods.

REVIEW OF LITERATURE

2.1 REVIEW OF LITERATURE :

The early literature on categorical data analysis dealt primarily with summary indices of association. The subject sparked heated debate among statisticians such as Karl Pearson and G. Udny Yule about how association should be measured. Pearson (1904, 1913) envisioned continuous bivariate distributions underlying cross-classification tables. He believed that we should describe association by approximating a measure such as the correlation for that underlying continuum. His tetrachoric correlation for a 2 x 2 table was one such measure. Suppose a bivariate normal density is collapsed to a 2 x 2 table having the same margins as observed table.The tetrachoric correlation is the value of the correlation ρ in the normal density that would produce cell probabilities equal to sample cell proportions. Pearson's contingency coefficeint was an attempt for I x J tables to approximate an underlying correlation.

Yule (1900) preferred to work with the defined category structure. Yule's perspective led him to define $G = (\theta-1)/(\theta+1)$ as a measured association for the 2 x 2 table, where θ is the cross product ratio. Yule believed that it was possible to define meaningful coefficients without assuming anything about underlying continuous distributions.

Pearson's approach relates closely to the literature of psychophysical paired -comparison experiments, as reviewed by Bock and Jones (1968). In such experiments each subject is asked to choose that member who is exhibiting the most defined attribute between a pair of stimuli. Each subject may judge one or many pairs, single or multiple times. Responses to a single pair may vary, across subjects or from time to time for a given subject, in accordence with a continuous probability distribution of percieved within pair stimulus differences. Proportions of choices within pairs exihibiting different physical stimulus differences are used to draw inferences about underlying probability distributions of perceptions. Such experiments are attributed to physiologist Vierordt, who utilized a standard stimulus as one member of each pair. They become prominent in the mid-19th century through publication by his student Hegelmayor (1852),and extensive development by Fechner (1860)using the Gaussion distribution. Cross-tabulation of the same subjects responses to each of two stimulies pairs thus gives a double dichotomization of a bivariate contineous distribution which was turned out to be compatible with Pearson's approach although analysis was not conducted from that point of view at the time. Responses of two sets of subjects to different pairs produce 2 x 2 table in which one dimension is non-stochastic due to the experimental design while, within each level, distributions of

the other differing only in location are partitioned at zero. The analysis of these 'constant method ' experiments was considered by other statisticians of the time, notably Spearman (1908) and Urban (1908) who gave least square, solution. Twenty years later Thurstone (1927 a, b, c, 1928) provided a rigorous mathematical underpining to psychological scaling, in which underlying continuous distributions generating quantal response data were fundamental. These and subsequent writings of Thurstone have been compiled (Thurstone, 1959), and further development in the area of paired comparison experimentation is reviewed by Bock and Jones (1968), David (1963) and Bradley (1976). The Bradley - Terry - Luce (Bradley and Terry, 1952; Luce, 1959) linear model for logistic transform of choice proportions has played a major role in this evolution.

If the stimulus difference from standard is identified with same measures of dose of a pharmocologic agent, and the paired-choice replaced by a biological quantal response such as death, the statistical issues underlying paired choice psychophysical experimentation are seen as quite close to those of quantal bioassay. There an attempt is made to measure the potency of a drug in the face of considerable biological variation in responsiveness of which only a quantal indicator, usually of destructive nature, is observable. The concept of percieved stimulus level transmutes to that of tolerance', a hypothetical dose level minimal to elicit the quantal response

from a given animal under experimental conditions.

Thurstone's discriminal process becomes a 'tolerance distribution' across animals. Each of several doses is administered to a (usually different) set of animals, and proportions of responses to each dose are utilized to draw inferences about tolerance distributions. A measure of location of an estimated tolerance distribution, often the medium (called LD50 or ID50 for a 50% lethal or effective dose), is used to summarise the potency. Using the Gaussian distribution of underlying tolerances, Gaddelum (1933), Bliss (1934, 9,b, 1935) and Fisher (1935), introduced probit analysis.Motivated by both theoretical considerations and computational rigour of probit analysis Berkson recommended substitution of very similar logistic for Gaussion tolerance distribution in series of papers from , 1944-1953. Such logit analysis could be accomplished by solution of linear equations as opposed to the iterations necessary for probits, because of similarity of the two tolerance distributions, essentially the same results are almost always obtained. Gurland et. al., (1960) introduced a multicategory logit analysis. Finney (1971) provides acomprehensive discussion of quantal bioassay methods

Log linear model analysis has rapidly become a major tool of statistical practice for deciphering multidimensional contingency tables arising through product-multinomial or

product-Poisson sampling. It is widely accepted that the general multiplicative model provides a natural framework for exploratory examination and testing of various hypothesis of statistical indipendence among variables or related distributional homogeneities, while within this framework many generally interesting parameterizations of patterns of dependence are available. Dyke and Patterson (1952) used logit analysis to model proportions pertaining to good and poor knowledge of cancer facts, in relation to a linear combination of categorical predictors involving exposure to media. This application substantially departed from the earlier literature in that the concept of stimulus or dose no longer proceeded directly from the problem or a natural model for it, but it was simply an index of multiple qualitatively distinct categorical contributors. Truett et al (1967), in an effort to develop multivariate predictive insights from the Framingham heart study data, examined a quantal model where cases and nomals (classified as such after longitudinal observation) are described by vectors from continuous multivariate normal distributions with different means and equal covariance matrices. Walker and Duncan (1967) gave maximum conditional likelihood solution based on the product binomial logistic model. Their method and related extension of Theil (1969) for multinomial data allow qualitative as well as quantitative predictors. Bishop (1969) demonstrated the identity of logistic models with only qualitative predictors and certain log linear models for multiple cross-classifications. Later

Grizzle and Williams (1972) further elucidated the relationship. Nerlove and Press (1973) developed a logistic model for multiple dependent variables based upon an underlying log linear equal association structure, and a linear model for the relationship of log linear main effects to categorical or continuous exogeneous variates. Lachenbruch (1975) and Goldstein and Dillton (1978) gave the details of discrete discrimination.

There was controversy over appropriate definition of association. Bartlett(1935) used the ratio of cross-product ratios in layers of a 2^3 table to measure second order interaction, and Norton (1945) extended this idea to the 2 x 2 table. Lancaster (1951) commenced development of an entirely, different approach, based on chi-squared partitioning using marginal distributions.

Goodman and Kruskhal (1954, 1959, 1963,1972) clarified the situation for two dimensional tables, but the three way problem proved much more difficult. Roy and Kastenbaum (1956) defined second-order interaction in an r x s x t table following Bartlett's approach. Their paper represents one of several important contributions by Roy and his students some of whom among others were Bhapkar, Diamond, Mitra and Sathe. The use of factorial design contrasts among logs of cell probabilities to define all orders of interaction in higher dimensional table is apparently due to Good

(1958,1960,1963,1965) and was further elaborated by Goodman (1964).

Lewis (1962), Bhapkar and Koch (1968 a, b) and Darroch and Speed (1979) discuss aspects of disagreement in this area, much debate has concerned the propriety of :

i) Linear representations of interactions in the tables resulting from multiple samples (as opposed to those originating from the cross-classification of one sample) and

ii) defining interaction in terms of marginal distributions (as opposed to internal slices) of a table.

Lancaster (1957,1969) and Lancaster and Hamdan (1964) have made modern contributions in the original framework of Pearson Though other strategies also see much use, Good's approach has assumed a dominant role, though development by Darroch (1962), Birch (1963, 1964, 1965), Mosteller (1968), Ku and Kullback (1968), Ku et al. (1971), Bishop (1969, 1971), Haberman (1974a) Placket (1974), Bishop et al (1975), Gokhale and Kullback (1978) and others.

An additional general problem of interest is that of standardization of tables, a problem which may arise in several forms. Deming and Stephan (1940) consider a method of adjusting a given observed table to conform to marginal totals

known a priori to describe the population of interest. This has been called the 'external constraints problem' by Gokhale and Kullback (1978), the former authors developed on iterative proportional fitting algorithm, known also as raking, to approximate a least-square solution to the problem. Their algorithm has proven of general utility as it converges (when applied to fit margins of an observed table to a uniform table) to the maximum likelihood fitted cell counts of log-linear model (Dorroch, 1962; Birch (1963). Gokhale and Kullback (1978) discuss both external constraints problem and this latter 'internal constraints problem' extensively. Direct and indirect methods of standardizing rates for comparison, in order to exclude effect of extraneous variables, have been extensively applied in demographic and epidemiologic literatures i.e. Bunker et al. (1969), Fleiss (1973), Shryock and Siegel (1973), Breslow and Day (1975), Bishop et al. (1975), Freeman and Holford (1980). The first reference, the National Halothane Study, reports on U.S.A hospital survey to investigate occurence of a rare event massive hepatic damage subsequent to surgical anesthesia. The significant stimulus that this study provided to the development of categorical data methodology, as well as that provided by the Framinagham Heart Study mentioned earlier, is worthy of historical emphasis. In view of Bishop et al. (1975) and other related references, Fienberg (1978) notes that 'Standardization is basically a descriptive technique that has been made obsolete ----' by log linear analysis. However, standardization is

likely to continue to see frequent application in the research literature. Fienberg's comment not ⌣ with ⌣ standing, occsionally it might be desirable to use log linear modelling in support of it, rather than as a replacement.

The preceding relatively brief sketch of the literature, for which the bibliographics of Killion and Zahn (1976) and Singer (1979) have proved quite useful, should convey a sense of variety of issues which may be addressed directly by log-linear modelling. These share the feature that unconditional or conditional hypothesis of independence or homogeneity are of central concern relative to individual cell probabilities. However, many statistical problems involving contigency tables or more general. Categorical data structures can not be so characterized but comment on the simultaneous use of log-linear models and standardization is meant to suggest that even in such situations, log linear models may play an important role. Thus scientist may be directly interested in issues such as marginal symmetry for repeated measurements or split-plot types of experiment (e.g. Koch et al 1977), measures of agreement in observer reliability studies (e.g. Landis and Koch (1977) or generally in variation of essentially nonlogarithmic functions of cell probabilities across subpopulations (e.g. Bhapkar and Koch, 1968 a, b) Forthofer and Koch (1973).

The philosophy of using log linear model with data set of increasing complexities encourages both intellectuals and computational economics and implemented in practice by numerous statistician in various frame works over a long period, its formal development is motivated by growth curve model concepts as described by Potthof and Roy (1964), Grizzle and Allen (1969), Koch and Greenberg (1971). Tolley and Koch (1974) developed a two stage application of this approach for categorical data analysis. The first stage smoothing procedure involves the selection of a model which parametrizes salient characteristics of the data. The model is then applied to relevent subsets of the data with estimates of parameters descriptive of each subsets, obtained by maximization of a corresponding partial or marginal likelihood function, formulated as (explicit or implicit) functions of the observed data . Applications of this procedure to biological data sets are presented by Koch and Tolley (1975) and Koch, Tolley and Freeman (1976).

The class of multistage procedure for the categorical data analysis is known as functional asymoptotic regression methodology, which provides efficient parameter estimates and consistent covariance estimates from some underlying first stage model, and utilizes weighted least sqare algorithm to examine these at later stages. This use of functional asymptotic regression methodology, described by Koch et al

(1976) and Landis et al (1978), is also implicit in the work of Nelder and Wedderburn (1972) and Haberman (1974 a). More, recently it has been described by Gokhale and Kullback (1978) and Haberman (1978).

Three general strategies for fitting log linear models have been widely promulgated in forms suitable for use by the researchers. These are:

I) Iterative proportional filling of hierarchical analysis of variance models, analogous to factorial cross classifications and subsequent reduced models for continuous data, marginal tables sufficient for model parameters are used to find maximum likelihood, minimum-discrimination estimates and associated likelihood ratio, information, or Pearson Chi-squared test statistics.

II) Weighted least squres fitting of asymptotic regression models to log linear functions of observed cell frequencies, leading to linearized minimum-modified (Neyman) chi-sqared estimates and Wald statistics.

III) Function maximization techniques of a more general mature i.e. Newton-Raphson or iterative weighted least - squares or modifications there of, leading to the same class of solution as method - I.

Method I is just an appropriate application of the raking algorithm of Deming and Stephan (1940), shown relevant and explored for log linear model fitting by Darroch (1962). Birch (1963), Bishop (1967), Fienberg (1970), Haberman (1972, 1973a, 1974a), Daroch and Ratcliff (1972), Gokhale (1971), and described by Goodman (1970), Ku and Kullback (1974), Bishop et al (1975) and elsewhere. Method II, with computational ancestry in logistic models of Berkson, was developed by Grizzle and williams (1972) as an application of the general procedure of Grizzle et al. (1969), who in turn had synthesized earlier work of Wald. (1943), Neyman (1949) and Bhapkar (1961, 1966, 1970). It is also used by Theil (1970). Mantel (1966), Walker and Duncan (1967), Cox (1970), Gokhale (1972), Nelder and Wedderburn (1972), Haberman (1974 a,b), Nelder (1974), Bock (1975), Schmidt and Strauss (1975 a,b) have been prominent advocates of method III.

The paper by Nathan (1969) can be regarded as the starting point of the application of standard techniques of categorical data in survey sampling. Cohen (1976) investigated a very special case of the general problem of testing goodness of fit from complex samples. He assumed a simple random sample of n clusters consisting of two units each. This sample of 2n units is classified into r cells. In the model studied by when if r_i is the probability of unit 1 of a cluster being in cell i, then the probability of the two units being in cells i and

j respectively is

$$r_{ij} = (1-a)\, r_i r_j \qquad \text{if } i \neq j$$

$$r_{ii} = r_i[a + (1-a)\, r_i] \qquad \text{otherwise}$$

for values of ' a ' between 0 and 1

Let there be m categories and denote the number of observations in the sample of n units which falls into the i-th category by $\bar{y}_i$, i=1,2, ----,m where $\Sigma\, \bar{y}_i = n$.To test the null hypothesis Ho, in cae of chi-square statistics for goodness of fit is

$$E\,(\bar{y}_i/n) = r_i \qquad i = 1, 2, ---,m.$$

The statistics is computed as

$$X^2 = \sum_{i=1}^{m} \frac{(\bar{y}_i - n\, r_i)^2}{n\, r_i}$$

Under this model it can be easily shown that

$$V(\bar{y}_i) = (1+a)2n\, r_i(1-r_i), \quad i = 1,2,--,m$$

The design effect as defined by Kish and Frankel (1974) is

$$Deffi = Var\,(\bar{y}_i)/n\, r_i(1-r_i)$$

The design effect as defined by Kish and Frankel (1974) is the ratio of the variance under the sampling design to the variance under simple random sampling with replacement. In other words design effect is a measure of deviation in variance under the sampling design under consideration with respect to simple random sampling with replacement and denoted by "Deff". So the design effect is equal to (1+a) for all i. Cohen shows that the statistic $X^2/(1+a)$ is distributed as chi-square (under Ho) with (m-1) degrees of freedom.

The most sustained work on tests of independente from complex samples has been carried out by Nathan (1969,1971,1972,1973,1975). He also reviewed the work of several other authors, such as Bhapkar and Koch (1968) and Chapman (1966).

Consider the usual contingency table. Let there be I rows and J columns, overall sample size is n, P_{ij}'s are estimated proportions of frequencies in the (i,j)-th cell (i=1,2,---,I; j = 1,2---,J). The statistic

$$X^2 = \sum_i \sum_j \frac{(nP_{ij} - nP_{i.}P_{.j})^2}{nP_{i.}P_{.j}} \qquad (2.1)$$

has approximately the chi-square distribution under the null hypothesis, if n is based on simple random sample and is sufficiently large. The quantities $P_{i.}$ and $P_{.j}$ are obtained by summation over the missing subscripts. Under the null hypothesis the expression

$$P_{ij} - P_{i.}P_{.j} \qquad (2.2)$$

has zero expected value but generally an expected value different from zero if the null hypothesis does not hold. The zero expected value of (2.2) is the result of number of variance and covariance terms in $E(P_{i.}P_{.j})$ cancelling. Under simple random sampling and null hypothesis we have

$$E(P_{i.}P_{j}) = r_{i.}r_{.j} + Var(P_{ij}) + \sum_{(k,k')\neq(i,j)} Cov(P_{kj}, P_{ik})$$

$$= r_{ij} + \frac{1}{n} r_{ij}(1-r_{ij}) - \frac{1}{n} \sum_{(k,k') \neq (ij)} r_{kj}\, r_{ik}$$

$$= E(P_{ij}) \qquad -(2.3)$$

In case of complex surveys the variances and covariances in the above equation have to be multiplied by their respective design effect, therefore may not cancel out, thus the expected value of $P_{ij} - P_{i.} P_{.j}$ may not be equal to zero even under Ho.

The works of both Nathan and Bhapkar and Koch start out with construction of an expression involving estimates of r_{ij}, $r_{i.}$ and $r_{.j}$ which has zero expected value under Ho, even in case of complex samples.For this purpose they both resorted to balanced repeated replication and make use of the fact that the two half-samples of any replicate are uncorrelated under the assumptions. Thus, if $\hat{P}_{ij}^{(k)}$, $\hat{P}_{i.}^{(k)}$, $\hat{P}_{.j}^{(k)}$ are estimated, under a complex sample design, from the first half sample of the k-th replicate and $\tilde{P}_{ij}^{(k)}$, $\tilde{P}_{i.}^{(k)}$, $\tilde{P}_{.j}^{(k)}$ are the corresponding quantities estimated from the second half sample, Nathan's test is based on the expression

$$U_{ij}^{(k)}(N) = \hat{P}_{ij}^{(k)} + \tilde{P}_{ij}^{(k)} - \hat{P}_{i.}^{(k)}\, \tilde{P}_{.j}^{(k)} - \tilde{P}_{i.}^{(k)} \hat{P}_{.j}^{(k)} \qquad (2.4)$$

and Bhapkar and Koch's is based on

$$U_{ij}^{(k)}(B) = \hat{P}_{ij}^{(k)}\, \tilde{P}_{rc}^{(k)} - \hat{P}_{ic}^{(k)}\, \tilde{P}_{rj}^{(k)} \qquad (2.5)$$

Both (2.4) and (2.5) have zero expected values under Ho.

Chapman's test is based on $$U_{ij}^{(k)}(C)=\hat{P}_{ij}^{(k)}-\tilde{P}_{i.}^{(k)}\tilde{P}_{.j}^{(k)} \qquad (2.6)$$

and it does not necessarily have zero expected value even under Ho.

Now if an estimate $\hat{\underset{\sim}{V}}$ can be constructed for the covariance matrix for (I-1) x (J-1) linearly independent quantities among the I x J, U_{ij} values, and if U is corresponding vector of these values, then

$$\underset{\sim}{U}'(\hat{\underset{\sim}{V}})\,\underset{\sim}{U}^{-1} \qquad (2.7)$$

would, for large enough n, and apart from a suitable constant multiplier, be either distributed approximately as F or as X^2, depending on whether $\hat{\underset{\sim}{V}}$ is estimated from a large enough number of degrees of freedom. In the case of simple random sampling each cell of covariance matrix of (2.2) is readily estimated approximately as

$$\hat{V}_{ij,fg} = \frac{1}{n} P_{i.}P_{.j}(1-P_{f.})(1-P_{.g}) \text{ for } (i,j) = (f,g)$$

$$= -\frac{1}{n} P_{i.}P_{.j}P_{f.}P_{.g} \qquad \text{for } (i,j) \neq (f,g)$$

and resulting estimate of $\hat{\underset{\sim}{V}}$ is based on a large number of degrees of freedom so (2.7) would be distributed as chi-square. In the case of complex samples the anologous estimation can not be carried out without some very strong simplifying assumptions.

Alternatively, one may observe that even in the case of complex samples the vectors $U^{(k)}$ in (2.4) - (2.6) are

identically distributed and hope to derive estimates of variances and covariances from

$$\sum_{k=1}^{K} (U_{ij}^{(k)} - \bar{U}_{ij}) (U_{fg}^{(k)} - \bar{U}_{fg}) \qquad - \qquad (2.8)$$

Now if the K replicate values of $U_{ij}^{(k)}$ were independent, the expression (2.8) above, divided by (k-1), would provide an unbiased estimate of $V_{ij,\ fg}$. However, far from independent, they are highly correlated. In the case of $U^{(k)}(N)$ the correlations are very close to one which was also described by Kish and Frankel, that for all replicates the sum of two analogous non-linear statistics, computed respectively from the two half-samples is very nearly the same for all K replicates and is identically the same in the case of linear statistics. In order to correct (2.8) for the correlation involved, one would have to estimate these correlations and that, in turn, again requires strong simplifying assumptions. Moreover, when the correlations are close to one the numerical behaviour of the estimate is very bad.

Thus ,whichever of the two methods of estimating the covariance matrix is attempted, strong simplifying assumptions are needed in case of complex samples. Nathan (1973) is forced to make the assumption, among the others, that for each stratum h there is a number n_h which depends only on the number of final units selected in the stratum h in each of the two primary sampling units (PSU), and if P_{ijha} is the estimate of the proportions in cell (i,j) derived from PSU a (a=1,2) of

stratum h, then n_h P_{ijha} has approximately the multinomial distribution with paramters n_h and r_{ijh}. However, this assumption implies that the expected value of an estimate p_{ijha} derived from any selected PSU conditional on that PSU being in the sample, depends on the stratum only not on the selected PSU. Thus total between PSU component of variance is assumed away. Other assumptions of Nathan, less important to his development assume away the effects of stratification and disproportionate sampling in different strata as well.

In the light of the comments above, it is not too surprising to find that the test statistics proposed by Nathan behaves very badly with respect to its achieved significance levels. The simulation results reported in his paper (1973) are flawed, as pointed out by the author in his own subsequent paper, Nathan (1975). The results reported in (1975) refer to stratified cluster sampling with a self-weighted design, so the traditional Chi-square Test can be applied to serve as a measure of comparison.

Fellegi (1980) proposed two tests after the careful examination of the problem of estimating the variance of non-linear statistics from complex samples, in the light of existing literature. The first test was obtained with the help of linear approximation of U_{ij} and it can be shown that

$$E\ (U_{ij}) = O(\frac{1}{n})$$

$$E\ (U^{2}\ ) = Var\ (U_{ij}\ ) + O(\frac{1}{n^{2}})$$

Fellegi proposed the first test statistics as

$$t' = \frac{L-m}{m(L-1)}\ U'(\hat{V})^{-1}U \qquad - (2.9)$$

where L is the number of strata and $m=(I-1)(J-1)$ t'is approximately distributed by F(m,L-m).The second test is more heuristically constructed than the first. As we knowthat

$$E\ (X^{2}) = \sum_{i}\ (1-r_{i})\ deffi$$

Where deffi = Var $(\bar{y}_{i})/n\ r_{i}(1-r_{i})$

Fellegi (1980) suggested the simplest approach which take into account the sample design, to correct Pearson Chi-squared statistics by some form of average design effect. The adjusted Chi-squared statistics was simply calculated using the following formula

$$X_{a}^{2} = X'^{2}/\ \bar{b} \qquad - (2.10)$$

Where $\bar{b}$ is the "average" design effect. Further, it was shown by Fellegi that the average cell design effect can be used as adjustment factor.

Hence,
$$\bar{b} = \frac{n}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{V_{ij}}{P_{ij}\ (1-P_{ij})}$$

where
$$V_{ij} = \frac{1}{4K} \sum_{k=}^{K} (\hat{P}_{ij}^{(k)} - \tilde{P}_{ij}^{(k)})^{2}$$

Rao and Scott (1979) proposed calculating the adjustment based upon the average eigen value of the following matrix

$$D = P_{o}^{-1}\ V$$

Where $h_{ij} = P_{ij} - P_i P_j$

$h = (h_{11}, \text{-------}, h_{IJ})'$

V(h) is the variance-covariance matrix of h and V_o (h) is the variance-covariance matrix for h under the null hypothesis of independence and multinomial sampling.

It can beseen that X^2 has asymptotic mean and variance

$$E(X^2) = \sum_{t=1}^{(I-1)(J-1)} \delta_t E(W_t) = (I-1)(J-1)\delta.$$

$$V(X^2) = \sum_{t=1}^{(I-1)(J-1)} \delta_t^2 V(W_t) = 2\sum_{t=1}^{(I-1)(J-1)} \delta_t^2,$$

respectively. Rao and Scott showed that the expectation and variance can be written in terms of the variance of h_{ij}'s

$$(I-1)(J-1)\delta. = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{V(hij)}{P_{i.}P_{.j}} \quad \text{and}$$

$$\sum_{t=1}^{(I-1)(J-1)} \delta_t^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{i'=1}^{I}\sum_{j'=1}^{J} \frac{[Cov(h_{ij}, h_{i'j'})]^2}{P_{i.}(1-P_{j.})P_{i'.}(1-P_{.j'})}$$

One approach suggested by Rao and Scott is to standardize X^2 so as asymptotic mean equal to (I-1)(J-1), which is what would be expected. if X^2 actually asymptotically follows χ^2 distribution with (I-1)(J-1) degrees of freedom. The adjusted Chi-squared variate takes on the form of (2.10) where

$$\bar{b} = \frac{n}{(I-1)(J-1)}\left[\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{V_{ij}}{P_{i.}P_{.j}} - \sum_{i=1}^{I}\frac{V_{i.}}{P_{i.}} - \sum_{j=1}^{J}\frac{V_{.j}}{P_{.j}}\right]$$

This expression is based upon a linear approximation to hij in terms of P_{ij}'s. It can be seen that this approximation only required estimates of the variance of cell proportions and row proportions.

Rao and Scott have aslo proposed a more complicated approximation that standardizes for both the asymptotic expectation and variance following the approach of Satterthwaite (1946). In this approach Pearson's Chi-square sttistics is standardized to have asymptotic expectation ν and variance 2ν, where

$$\nu = \frac{[\Sigma \delta_t]^2}{\Sigma \delta_t^2} \qquad - (2.13)$$

This is done by modifying Pearson's chi-square statistic to the following :

$$X_a^2 = X^2 \frac{\Sigma \delta_t}{\Sigma \delta_t^2} \qquad - (2.14)$$

Which is then treated as Chi-squared variate with ν degrees of freedom. As with the Chi-squared statistics that has been standardized to have correct expectation, the correction factor can be based upon either $V(h_{ij}, h_{i'j'})$, which can be calculated using replication, or upon the variance-covariance matrix of linear approximation to the h_{ij}'s using P_{ij}'s. The Fellegi and first Rao and Scott statistics are simplest to

calculate, but are thought to be excessively conservative. The second Rao and Scott statistics attempts to compensate for the suspected conservatism shown by the simpler statistics, but it is not clear at present time if this is generally achieved in practice. In general, the Satterthwaite adjusted statistics which does require more computational efforts is likely to be the best approximation.

Rao and Scott (1984) considered the analysis of multidimentional contingency table with the help of two log-linear models for r. where r is a vector of true proportion in the population. The first model is of the form

$$\log_e r = \mu_1(\theta_1)\, 1 + X_1\, \theta_1 \qquad - (2.15)$$

Where $\log_e r = (\log_e P_i)$ which is vector of log-proportions X_1 represents a given design matrix, and θ_1 denotes a vector of unknown parameters. The function $\mu_1(\theta_1)$ takes a value depending on the parameters θ_1 such that the sum of probabilities over the cells of the table is one. The second is

$$\log_e r = \mu(\theta)1 + X\theta \qquad - (2.16)$$

Where $X = (X_1, X_2)$, $\theta = (\theta_1, \theta_2)$. The purpose is to test the improvement of model (2.16) over model (2.15) when the model (2.16) is assumed to be hold i.e. to evaluate the hypothesis $\theta_2 = 0$. As a special case (2.16) may be taken to be the saturated model, that is, the fully parameterized model that fits any sets of positive probabilities exactly. In this case the test represents an evaluation of the over all fit of the

model (2.15).

Let $P_1 = (P_{1i})$ denote the maximum likelihood estimates of the cell proportion s under the model (2.15) based on the multinomial likelihood, and $P_2 = (P_{2i})$ denote the corresponding estimate under model (2.16). The Pearson chi-square test for this comparison is given by

$$X^2 = n \sum (P_{1i} - P_{2i})^2 / P_{1i} \qquad - (2.17)$$

and the likelihood ratio Chi-square by

$$G^2 = 2n \sum P_i \log_e (P_{2i}/P_{1i}) \qquad - (2.18)$$

The method employing Satterthwaite approximation as proposed by Rao and Scott requires the matrix

$P = (P_{ij})$, where

$P_{ii} = P_i - P_i^2$

$P_{ii'} = - P_i P_{i'}$ for $i \neq i'$

and estimated covariance matrix V equal n times the estimated covariance matrix of P under the complex sampling design. Then, let

$$\tilde{X}_2 = (I - X_1 (X_1' P X_1)^{-1} X_1' P) X_2$$

$$M^* = (X_2' P X_2)^{-1} (X_2' V X_2)$$

The sum of eigen values of M^* may be found as the trace of M^*, and the trace of $M^* M^*$ gives the sum of squares of the eigen values of M^*. The satterthwaite approximation is to compute the integer K' nearest to $[Tr(M^*)]^2 / tr (M^* M^*)$, and to compare $X^2 = (K'/Tr(M^*)) X^2$ to the Chi-square distribution with K degree

of freedom.

Fay (1985) proposed a method based upon recomputing the Chi-square test (2.17) and (2.18), or difference of Chi-square tests of two nested models, each compared to saturated model for a series of replicate sample based on sample data. Each replicate is of the form

$$\bar{Y} + W^{(h,j)}, \quad h = 1, \text{-----}, H; \; j=1, \text{-------}, J_h$$

where $\bar{Y} = (y_1, \text{----}, y_{IJ})$, H is the total number of strata, J_h is the number of replications in the stratum h, where

$$\sum_j W^{(h,j)} = 0 \qquad - (2.19)$$

for each h, such that the usual replication based estimator of V^*, the sampling covariance matrix of $\bar{Y}$ under the complex sampling design is given by

$$V^* = \sum_h b_h \sum_j W^{(h,j)}(x)\, W^{(h,j)} \qquad - (2.20)$$

Where (x) is standard outer product i.e. usual cross product matrix.

Let $X^2_{(1)}(\bar{Y})$ denote the value of the Pearson Chi-square test for evaluating the fit of P_1 and $X^2_{(2)}(\bar{Y})$ corresponding to P_2. Define

$$R_{hj} = \{ X^2_{(1)}(\bar{Y} + W^{(h,j)}) - X^2_{(2)}(\bar{Y} + W^{(h,j)}) \}$$

$$- \{ X^2_{(1)}(\bar{Y}) - X^2_{(2)}(\bar{Y}) \} \qquad - (2.21)$$

$$K^{\#} = \sum_h b_h \sum_j R_{hj} \qquad - (2.22)$$

$$V^{\#} = \sum_h b_h \sum_j R_{hj}^2 \qquad - (2.23)$$

$$X^2_J = \frac{\{ X^2_{(1)}(\bar{Y}) - X^2_{(2)}(\bar{Y}) \}^{1/2} - \{K^+\}^{1/2}}{\{ V^{\#}/ 8 \{ X^2_{(1)}(\bar{Y}) - X^2_{(2)}(\bar{Y}) \} \}^{1/2}} \qquad - (2.24)$$

Where K^+ is $K^{\#}$ when letter is positive,0, otherwise. A similar statistics, G_j , is obtained by replacing X^2 by G^2 throughout. The test procedure is to compare X_J or G_J to critical value tabulated in Fay (1983 or 1985).

Some of the relatively less important work in categorical data analysis in survey sampling are, a paper by Shuster and Downing (1976) who proposed methods for testing independence, quasi independence and marginal symmetry in contingency tables which are derived for a wide variety of sampling schemes, Cowan and Binder (1978) analysed the effect of a two stage sample design on the test of independence in a 2 x 2 table, Brier (1980) used Dirichlet-multinomial distribution as a model for contingency tables generated by cluster sampling schemes, Holt, Scott and Ewings (1980) emperically study the

survey design effect on test of goodness of fit,test of homogeneity and test of independence. Similarly other important empirical studies in this regard are by Rao and Hidiroglou (1981), Kumar and Rao (1984), Fay (1984), Thomas and Rao (1984), Thomas and Rao (1985), Singh and Kumar (1986) and Fay (1989) etc. Rao and Scott (1987) obtained the simple upper bounds on δ. For models not admitting direct solutions, also requiring only cell 'deffs' and marginal 'deffs' not depending on any hypothesis. Certain aspects of multivanate analysis of the data from possibly complex survey designs are discussed in terms of a large sample methodology involving weighted least squares algorithms for the computation of Wald statistics by Koch, Freeman and Freeman (1975).

2.2 ORIENTATION OF THE PROBLEM

Methods for analysis of categorical data have been extensively developed under the assumption of multinomial or product-multinomial sampling. In particular the standard Pearson Chi-square statistic and likelihood ratio statistics are used to test hypothesis in contigency tables. For the analysis of multi-way contingency tables log-linear models are applied extensively. These methods are often used by researchers in subject matter area to analyse sample survey data, even though the multinomial assumptions are voilated because of clustering and stratification used in survey design. Ignoring the effect of survey design and using

Chi-square or likelihood ratio statistics could lead to unacceptably high type I error. Alternative statistics that take account of the design have been recently proposed by many authors. Few of them designed for use with published tables, require knowledge of only cell design effects, whereas other statistics require knowledge of the full covariance matrix or access to the original data file.

In the first chapter of this thesis the concepts of the subject matter related to the categorical data analysis are summarized briefly. A comprehensive review of literature related to the categorical data analysis in general and in case of survey sampling in particular is presented in chapter two.

Chapter three deals in detail with the effect of stratification and clustering on asymptotic distributions of standard Pearson Chi-square test statistics for goodness of fit (simple hypothesis) independence of attributes and heterogeneity of proportions for two way contingency table. It has also been shown that all these three Chi-square statistics are asymptotically distributed as weighted sum of χ_1^2, random variables, where weights are related to familiar design effects used by survey samplers. Many simple corrections to the ordinary Chi-square are also presented, few of them require only the knowledge of variance estimates for individual cells whereas others require knowledge of full variance-covariance

matrix of cell proportions.

The impact of survey design on standard multinomial based methods for a multiway contingency table has been investigated thoroughly in chapter four, under nested log-linear models. Here also, the asymtotic null distribution of Pearson Chi-square test statistic is obtained as a weighted sum of independent X^2, random variables, and the weights are then related to familiar design effect further, simple corrections to Pearson Chi-square are discussed for both, whenever the model admits direct solution of the likelihood equations under multinomial sampling or when the model does not admit a direct solution of likelihood equation under multinomial sampling. The adjustments to the ordinary Chi-squance through Jackknifing and balanced repeated replication approaches are also discussed.

In case of categorical data analysis, generally sampling design is ignored while estimating the parameters like cell proportions and variance covariance matrix of proportions which often leads to biased estimation of these parameters. This leads to further distortion of the real situation when standard methods of categorical analysis are used in survey data. So, chapter five deals with various estimators of these parameters like, Horvitz-Thompson estimator, combined ratio estimator and post-stratified estimator. Further, various methods used for estimating variance covariance matrix in case

of complex surveys i.e. linearization, Jackknifing and BRR techniques are compared theoretically for combined ratio estimator.

Finally, all the statistics proposed in previous chapters are compared with each other through Chi-square tests of independence with the help of loglinear models fitted to survey data of a research project.

ANALYSIS OF TWO-WAY CONTINGENCY TABLE

ANALYSIS OF TWO-WAY CONTINGENCY TABLES

3.1 INTRODUCTION :

Methods of analysis of categorical data have been extensively developed under the assumption of multinomial and product- multinomial sampling. Researchers in subject matter areas, have long been using these multinomial based methods to analyze sample survey data, but most of the commonly used survey designs employ stratification or cluster sampling or both and hence do not satisfy the assumption of multinomial sampling in real sense. Operational and cost considerations often dictate the use of a complex cluster design. Such analysis often ignore the sample design and apply standard statistical methods appropriate for random sampling partly, because of the availability of computer packages, and wealth of survey data published in tabular format. It is somewhat unfortunate that this combination of software, incorporating traditional statistical methods, and the published survey data is so readily interfaced. It is therefore, important to study the effect of survey design on the standard statistical methods and suggest modifications accordingly. Rao and Scott (1981), Holt, Scott and Ewings (1980), Fellegi (1980) and others investigated the effect of clustering on traditional Chi-square tests. Their results indicate that distortion of nominalsignificance levels due to clustering could be quite severe. In this Chapter the approaches of Rao and Scott (1981) and Fellegei (1980) have been followed

to discuss the methods of analysis of two dimensional contingency table in survey sampling.

3.2 CHI-SQUARE TEST OF GOODNESS OF FIT:

Consider, first of all, the chi-square statistics as a test of goodness of fit. First, assume a simple random sample. Let there be k, categories and denote the number of observations in the sample of n units that falls into the i- th category $\bar{y}_i$ where i = 1, 2,..., k. So we have $\sum_{i}^{k} \bar{y}_i = n$. Let the null hypothesis (Ho) to be tested be $E(\bar{y}_i/n) = r_i$, i=1, ----,k; where r_i is expected proportions in the ith category. Now the Chi-square test statistic to test the above hypothesis is given as

$$X^2 = \sum^{k} \frac{(\bar{y}_i - nr_i)^2}{nr_i} \qquad (3.1)$$

It is well known that, under Ho, X^2 is distributed asymptotically as Chi-square with (k-1) degrees of freedom. This asymptotic Chi-square distribution does not hold under complex designs.

Let us consider the relatively simple case of stratified sampling. Under Ho we have

$$E \mid n^{-1} \bar{y}_{ih} \mid = r_{ih} \qquad (3.2)$$

Where r_{ih} is the expected proportion in the i-th category of h-th stratum. Consider, $\bar{y}_i = \sum_{h} W_h \bar{y}_{ih}$, as a statistic, in this case to estimate the number of observations in the i-th category

of the population, where W_h is the stratum weight for the h-th stratum

h=1,..., L, taking the expectation of $\bar{y}_i$ we get,

$$E(\bar{y}_i) = \sum_h W_h E(\bar{y}_{ih}) = n \sum_h W_h r_{ih} \qquad - (3.3)$$

Putting the value of $E(\bar{y}_i)$ from (3.3) in to equation (3.1) we get

$$E(X^2) = E \sum_{i=1}^{k} \frac{(\sum W_h \bar{y}_{ih} - n \sum W_h r_{ih} + \sum W_h r_{ih} - nr_i)^2}{nr_i}$$

$$= E \left[\sum_{i=1}^{k} \frac{(\sum_h W_h \bar{y}_{ih} - n\sum_h W_h r_{ih})^2}{nr_i} + \sum_{i=1}^{k-1} \frac{(n\sum W_h r_{ih} - n\dot{r}_i)^2}{nr_i} \right]$$

$$(k-1) = E(X^2_{st}) + \sum_{i=1}^{k} (nr_i)^{-1} \sum_h n_h (r_{ih} - nr_i)^2$$

$$\Rightarrow E(X^2_{st}) = (k-1) - \sum_{i=1}^{k} (nr_i)^{-1} \sum_h n_h (r_{ih} - nr_i)^2 \qquad (3.4)$$

Where X^2_{st} denotes the X^2 statistic for stratified sampling. From equation (3.4) it is clear that expected value of X^2 reduces as a result of departure of the sampling design from simple random sampling with replacement. In case of self weighting designs (i.e. the inclusion probability of each unit in the population is same for a self-weighting sampling design). We have

$$E(X^2) = \sum_{i=1}^{k} \frac{(\bar{y}_i - nr_i)^2}{nr_i} = \sum_{i=1}^{k} \frac{V(\bar{y}_i)}{nr_i}$$

$$= \sum_{i=1}^{k} \frac{(1-r_i)\ V(\bar{y}_i)}{nr_i(1-r_i)} = \sum_{i=1}^{k} (1 - r_i)\ deff_i \qquad (3.5)$$

Kish calls the quantity "deffi" as design effect. The analogous quantities can be defined for covariances. It was found that in the case of multistage stratified cluster sampling "deffi" is always greater than one so equation (3.5) will always be greater than (k-1). However, the "deffi" can be smaller than one, depending upon the sampling design and variable measured. Again consider the case of self-weighting designs, we know that $E(\bar{y}_i) = n\ r_i$. Here expectation is taken over with respect to the sampling design into consideration. For other designs we encounter the problem of non-centrality as X^2 involving non-central distributions, so it is natural to consider a more general statistics for Ho: $r_i = r_{io}$.

$$\bar{X}^2 = n \sum_{i=1}^{k} (p_i - r_{io})/r_{io}$$

Where p_i is an unbiased (or consistent) estimator of r_i under sampling design p(s) and $\sum p_i = 1$. If $np_i = \bar{y}_i$ then this reduces to the usual statistics given by equation (3.1). Koch, Freeman and Freeman (1975) proposed generalized Wald statistics for testing H_o which is given by

$$X^2_w = n\ (p - r_o)'\ \hat{V}^{-1}\ (p - r_o) \qquad (3.7)$$

where $p = (p_1, \ldots, p_{K-1})'$

$r_o = (r_{o1}, \ldots, r_{o,K-1})'$

and $\hat{V}/n$ = estimator of the covariance matrix, V/n of p.

It was shown that X^2_W is distributed as χ^2_{k-1} under Ho for sufficiently large n. However, in many situations it is not possible to estimate covariance matrix and consequently X^2_W, so it is important to study the effect of survey designs on the distribution of $\bar{X}^2$. Further, it was found through empirical investigations that Wald statistic is very inconsistent in many practical situations. In the case of simple random sampling the Wald statistics reduces to the following statistics.

$$X^2 = n\,(p - r_o)'\, P_o^{-1}\,(p - r_o)$$

where P_o is the value of $P = \mathrm{diag}(r) - rr'$ for $r=p_o$ and P/n is the covariance matrix of $\bar{y}/n$ for multinomial sampling, where $\bar{y} = (\bar{y}_1, \text{-------}, \bar{y}_{k-1})$.

3.2.1 ASYMPTOTIC DISTRIBUTION :

It is clear that $\bar{X}^2$ will not follow central chi-square distribution under any sampling design ,except in the case of simple random sampling with replacement. We now look to the problem of asymptotic distribution of $\bar{X}^2$. The population proportions can be written as mean of the variable $y_t^{(i)}$, where $y_t^{(i)}$ is defined as follow:

$$y_t^{(i)} = 1, \text{ if the t-th population element belongs to i-th class}$$

$$= 0, \text{ othrwise.}$$

Hence, the estimator p_i for the proportion r_i of the i-th category can be of the form.

$$p_i = \sum_{t \in s} W_t(s)\, y_t^{(i)}$$

where $W_t(s)$ is some weight attached to the sample. Again

$$l'\, p = \sum_{t \in s} W_t(s)\, y_{t(l)}$$

where $l = (l_1, \text{-----}, l_{k-1})'$ any vector of constants

and $$y_{t(l)} = \sum_{i=1}^{k-1} l_i\, y_t(i)$$

If we have the central limit theorem for means for specified design i.e. if

$$\sqrt{n}\ (p - r) \xrightarrow{L} N(0, V/n)$$

as $n \longrightarrow \infty$

Then

$$\sqrt{n}\ (l'p - l'r) \xrightarrow{L} N(0, l'V\, l/n)$$

as $n \longrightarrow \infty$

where $n\ \mathrm{Cov}\ (p_i, p_j) = \nu_{ij}$ and $V = (\nu_{ij})$. Hence, p is approximately (k-1)- variate normal with mean vector r and covariance matrix V/n, for sufficiently large n. If a consistent estimator, $\hat{V}$, of V is available then generalized Wald statistics $\tilde{X}^2_w$ will be distributed asymptotically as χ^2_{k-1} under Ho. The correct

asymptotic distribution of $\tilde{X}^2$ can be directly obtained from

normality of p with the help of following theorem.

THEOREM :

Under null hypothesis Ho: $r = r_o$, $\bar{\chi}^2$ may be written as $\sum_{i=1}^{k-1} \lambda oi \; Wi^2$, where $Wi \overset{asy}{\sim} N(0,1)$ and Wi's are independent. λoi'sare the eigen values of $D_o = P_o^{-1} V_o$, $(\lambda o_1 \geq \lambda o_2 \geq, \; ----, \geq \lambda o_{,k-1} > 0)$ where V_o/n denotes the covariance matrix V/n for $r = r_o$.

PROOF :

Let $z = (p - r_o) \rightarrow E(z) = 0$

The variance - covariance matrix of $z = V_o/n$. Now, if central limit throrem holds for $r = r_o$ pdf of z can be written as

$$p_z(z) = (2\pi)^{-k/2} \left| Vo/n \right|^{-1/2} Exp \left[-\frac{1}{2} n \, z' \, V_o^{-1} z \right]$$

Any homogeneous quadratic form in $z_1, ---, z_{k-1}$ may be expressed as

$$Q(z) = n \, z' P_o^{-1} z = n \sum_{i=1}^{k-1} \sum_{j=1}^{k-1} a_{ij} \, Z_i Z_j$$

where $Po^{-1} = (a_{ij})$ as a symmetric matrix. The distribution of Q(z) can be found out as

$$P_r [Q(z) \leq y] = \int_{nz' P_o^{-1} z \leq y} p_z(z) \, d \, z \qquad - (A)$$

This can be reduced to standard convenient form. Since Vo is a positive definite and symmetrical, it can be factored as

$$Vo = L \, L'$$

where L' is a non-singular lower triangular matrix . Since, L is non-singular, the change of vriable $z_{(1)} = L^{-1} z$ is permissible. Since

$$z'_{(1)}\ z_{(1)} = z'\ L^{-1}\ L^{-1}z = z'\ Vo^{-1}\ z$$

and $$\left|Vo\right| = \left|\ L\ \right|^2$$

From the above we can write (A) as

$$P_r[\ Q\ (z) \leq y] = (2\pi)^{-\frac{1}{2}(k-1)} \sqrt{n} \int \text{Exp}\left[-\frac{n}{2}\ z'_{(1)} z_{(1)}\right] dz_{(1)}$$
$$n\ z_{(1)}(\ L' P_o L\ )\ z_{(1)} \leq y \qquad -(B)$$

The matrix $L'\ P_o^{-1}\ L$ is symmetric, so if Δ is a diagonal matrix of eigen values of $L'\ P_o^{-1} L$ and M is associated orthogonal matrix of eigen vectors then $M'L'P_o^{-1}L\ M = \Delta$ Hence, if further linear transformation $W = M'\ z_{(1)}$ is applied (B) becomes

$$P_r\ [Q(z) \leq y] = (2\pi)^{\frac{1}{2}(k-1)} \sqrt{n} \int_{n\ w'\Delta\ w} \text{Exp}\ [-\frac{n}{2}\ W'\ W\ ]\ dW$$

Thus the distribution of Q is same as that of

$$W'\ \Delta\ W = \sum_{i=1}^{k-1} \lambda i\ W_i^2$$

Where the variables W_i 's are independently distributed as N(0,1) and numbers $\lambda_1 \geq \lambda_2 \geq$, ----- , $\geq \lambda_{k-1}$ are the eigen values of $L'\ P_o^{-1}L$ which is equivalent to eigen values of $Do = P_o^{-1}V_o$. It follows that attention can be focussed on the distribution of

$$Q\ (W) = \sum_{i=1}^{n} \lambda i\ Wi^2.$$

In general, then X^2 is distributed asymptotically as weighted sum $\sum_{i=1}^{k-1} \lambda oi\ Wi$, where $Wi \overset{asy}{\sim} N(0,1)$.

COROLLARY 1 : $$\frac{\bar{X}^2}{\lambda_{o1}} \leq \sum_{i=1}^{k-1} W_i^2$$

where $$\sum_{i=1}^{k} W_i^2 = n\,(p - r_o)' V_o^{-1} (p - r_o) \overset{asy}{\sim} \chi^2_{k-1}$$

under Ho.

If the largest eigen value can be specified , (or a reasonable bound can be set), we can obtain, an asymptotic conservative test by treating X^2/λ^* as $\chi^2_{k-1,}$ where $\lambda^* \geq \lambda_{o1}$

COROLLARY 2 :

$\frac{X^2}{\lambda} \approx \chi^2_{k-1}$ for any r_o if and only if $V = \lambda P$ for some constant λ, that is $V(p_i) = \lambda\, r_i(1-r_i)/n$, and $Cov(p_i, p_j) = -\lambda\, r_i r_j/n$.

3.2.2 DESIGN EFFECT :

The general result discussed in the previous theorem can also be put in the form of finite sampling terms. Let us assume that there are two categories i.e.k=2. Now $D = P^{-1}V$ reduces to the ordinary design effect "deff" as

$$\frac{n\, V(p_1)}{p_1\,(1-p_1)}$$

where variance V is taken with respect to the sampling design into consideration. For general k, D can be thought as multivariate extension of 'deff'. In particular

$$\lambda_1 = \sup_c [C'V\,C \mid C'P\,C]$$

$$= \sup_{c} \left[V\left(\sum_{i=1}^{k-1} C_i p_i\right) \,\middle|\, V_{srs}\left(\sum_{i=1}^{n} C_i \bar{y}_i / n\right) \right]$$

where V_{srs} = Variance operator under simple random sampling with replacement i.e. multinomial sampling

If λ_1 = largest possible "deff" taken overall individual p_i s and overall possible linear combinations of p_i's.

The other λ_i's represents the deffs for special linear combinations of p_i's. Thus it can be termed as "generalized deffs" that are consistently estimated by λ_i's, the eigen values of $\hat{D} = \hat{P}^{-1} V_{\cdot}^{\wedge}$

Where $\hat{P} = \text{diag}(p) - p\,p'$

$p = (p_1, \text{-----}, p_{k-1})'$

Now $\dfrac{\bar{X}^2}{\lambda_1} \sum_{i=1}^{k-1} \dfrac{\lambda_{oi}}{\lambda_1} W_i^2 \le \sum_{i=1}^{k-1} W_i^2$ (From coll. 1)

Corollary 2 says that $\bar{X}^2 = \lambda \chi^2_{k-1}$ for any r_o. It means that not only all the individual cells have the same deff λ but the deff of all covariance terms are also equal to λ. Which may be an impractical condition and rarely satisfied.

3.2.3 SPECIAL CASES :

Three special cases are discussed to show the implication of general result discussed above. These are

(I) Simple random sampling without replacement

(II) Stratified random sampling (proportional allocation)

(III) Two stage sampling.

(I) Simple random sampling without replacement :

In this case we have $V = (1 - \frac{n}{N})\, P$ where N is finite population size. Now $X^2_v = n(p - r_o)' \hat{V}_o (p - r_o) \sim \chi^2_{k-1}$ substituting the value for $\hat{V}^{-1}$ we get

$$X^2_v = \frac{n\,(p - r_o)'\, \hat{P}^{-1} (p - r_o)}{(1 - \frac{n}{N})} \sim \chi^2_{k-1}$$

In other way

$$\overline{X^2} = (1 - \frac{n}{N})\, \chi^2_{k-1}$$

as both N and n $\longrightarrow \infty$ in such a way that (N-n) also tends to infinity. Thus $\lambda_i = \lambda_{oi} = (1 - \frac{n}{N})$ for any r_o, so

$$X^2_c = (1 - \frac{n}{N})^{-1} X^2 \sim \chi^2_{k-1}$$

(II) Stratified random sampling (proportional allocation) :

Let us consider L strata and $s = (s_1, ----, s_L)$ is stratified sample from the population under consideration, where s_h is a random sample of size m_h drawn from the h-th strata, h = 1, 2, ---, L: $\Sigma m_h = n$. Suppose m_{hi}, i = 1, --k; are the observed cell frequencies in the stratum h, W_h is the population proportion of elements in the stratum h and r_{hi} is the proportion of elements from stratum h belonging to i-th category.

Then

$$r_i = \sum_{h=1}^{L} W_h \, r_{hi}$$

and

$$p_i = \sum_{h=1}^{L} W_h \frac{m_{hi}}{m_h}$$

But under proportional allocation

$$m_h = n \, W_h$$

so,

$$p_i = \sum_{h} W_h \frac{m_{hi}}{n W_h} = \frac{\bar{y}_i}{n}$$

where, $\sum_{h} m_{hi} = \bar{y}_i$

With this design, p is approximately (k-1)-variate normal, for large m_h, with mean r and covariance matrix V/n, where V can be obtained as given below

Define

y_{hij} = 1, if j-th unit of h-th stratum falls in the i-th category.

= 0, otherwise

$$\bar{y}_{hi} = \frac{1}{m_h} \sum_{j=1}^{m_h} y_{hij} = \frac{m_{hi}}{m_h} = p_{hi}$$

$$\bar{y}_i = \sum_{h=1}^{L} W_h \, \bar{y}_{hi} = \sum_{h=1}^{L} W_h \, p_{hi}$$

$$V(p_i) = \sum_{h=1}^{L} W_h^2 \, V(p_{hi}) = \sum_{h=1}^{L} \frac{W_h^2}{m_h} \, r_{hi} \, (1-r_{hi})$$

but for the proportional allocation $m_h = nW_h$

$$V(p_i) = \frac{1}{n} \sum_{h=1}^{L} W_h r_{hi} (1-r_{hi})$$

$$= \frac{r_i}{n} - \frac{1}{n} \sum_{h=1}^{L} W_h r_{hi}^2$$

In the matrix notation we can write

$$V = P - \sum_{h=1}^{L} W_h (r_h - r)(r_h - r)'$$

$$= P - H \quad \text{(say)}$$

where

$$r_h = (r_{h1}, \cdots, r_{h,(k-1)})',$$

Now consider

$$C' V C = C' [P - \sum_{h=1}^{L} W_h (r_h - r)(r_h - r)'] C$$

$$= C' P C - \sum_{h=1}^{L} W_h [C' (r_h - r)]^2$$

from the above equation we can say that

$$0 \leq \frac{C' V C}{C' P C} = 1 - \sum_{h=1}^{L} W_h \frac{[C'(r_h - r)]^2}{C' P C}$$

or $\lambda_{o1} \leq 1 \quad \forall \quad r_o$

and $0 \leq X^2 \leq \sum_{i=1}^{k-1} W_i^2 \approx \chi^2_{k-1}$

Hence, from the above it can be concluded that the Pearson statistic X^2 is always asymptotically conservative in case of stratified random sampling. The extent of conservatism may be

seen by noting that $\bar{X}^2=0$ if $L \geq K$ and stratification is perfect, that is all elements in a stratum belong to the same category. Again if $L < K$ we have

$$X^2 \geq \sum_{i=1}^{k-L} W_i^2 \sim \chi^2_{k-L} \quad \text{since } R(H) \text{ is}$$

at the most (L-1) and hence at least k-L of the λi's must be equal to one. Thus X^2 is asymptotically well approximated by χ^2_{k-1} if $\frac{(k-L)}{(k-1)} \simeq 1$ or k is large and L is relatively small.

Now consider the special case for L=2, for which the eigevalues λi can be evaluated explicitly. We have

$$\begin{aligned} D &= P^{-1}V = P^{-1}[P-\{W_1(r_1-r)(r_1-r)' \\ &\qquad + W_2(r_2-r)(r_2-r)'\}] \\ &= I - P^{-1}[W_1(r_1-r)(r_1-r)'+W_2(r_2-r)(r_2-r)'] \end{aligned}$$

But $r = W_1 r_1 + W_2 r_2$

So,

$$\begin{aligned} D &= I - P^{-1}[W_1(r_1 - W_1r_1 - W_2r_2)(r_1 - W_1r_1 - W_2r_2)' \\ &\qquad + W_2(r_2 - W_1r_1 - W_2r_2)(r_2 - W_1r_1 - W_2r_2)'] \\ &= I - P^{-1}[W_1\{(1-W_1)r_1 - W_2r_2\}\{(1-W_1)r_1 - W_2r_2\}' \\ &\qquad + W_2\{(1-W_2)r_2 - W_1r_1\}\{(1-W_2)r_2 - W_1r_1\}'] \\ &= I - P^{-1}[W_1W_2^2(r_1-r_2)(r_1-r_2)' + W_1^2W_2(r_1-r_2)(r_1-r_2)'] \\ &= I - P^{-1}[W_1W_2(W_1+W_2)(r_1-r_2)(r_1-r_2)'] \\ &= I - P^{-1}[W_1W_2(r_1 - r_2)(r_1-r_2)'] \qquad (\because W_1+W_2=1) \\ &= I - A \text{ (say)} \end{aligned}$$

Since rank of A is one which implies that k-2 of its eigen values are zero. The remaining non-zero eigen values are found as

$$Tr(A) = W_1 W_2 \sum_{i=1}^{k} \frac{(r_{1i} - r_{2i})^2}{r_i} = \delta^* \text{ (say)},$$

where $0 \le \delta^* \le 1$. Hence $\lambda_1 = \text{---} = \lambda_{k-2} = 1$ and $\lambda_{k-1} = 1 - \delta^*$ so that

$$X^2 = \sum_{i=1}^{k-2} W_i^2 + (1-\delta_o^*)\, W_{k-1}^2$$

$$= \chi^2_{k-2} + (1-\delta_o^*)\, \chi^2_1$$

Where δ_o^* is the value of δ^* at $r = r_o$. Thus unless k is small, X^2 is asymptotically approximated by χ^2_{k-1} in the two strata case.

(III) Two - stage Sampling :

Suppose we have R primary sampling units (PSU's) and M_t denotes the secondary units in the t-th PSU (t=1,2,--,R). So, $\sum_{t=1}^{R} M_t = N$, where N is total number of secondary units in population.Consider the following commonly used sampling design : Let r PSU's be selected with probability proportional to size of M_t with replacement and then a sample of size m was drawn with simple random sampling with replacement from each PSU's. Hence, total number of secondary units selected in the sample is mr. Again let m_{li} be the total number of secondary units falling in the i-th category in the l-th PSU. In this case we have :

$$p = \sum_{l=1}^{r} p_l / r = n/n$$

where $n = (\bar{y}_1, \text{---}, \bar{y}_k)$

$$E(\bar{y}_i) = n\, r_i$$

where $p = (p_{l1}, \text{---}, p_{l,k-1})'$

Also,

$$p_{ti} = \frac{m_{ti}}{m},$$

and

$$r_i = \sum_{t}^{R} M_t r_{ti}/N$$

where r_{ti} is the proportion in the t-th PSU belonging to i-th category. With this design, p is approximately (k-1)-variate normal, for large r, with mean r and covariance matrix V/n where V can be obtained as given below:

Define

$y_{tki} = 1$, if the k-th SSU of t-th PSU belongs to i-th category.

$= 0$, otherwise.

Now

$$p_{ti} = \frac{1}{m} \sum_{k=1}^{m} y_{tki}$$

and

$$p_i = \frac{1}{r} \sum_{t=1}^{r} p_{ti}$$

Hence,

$$E(p_i) = E_1\{\frac{1}{r} \sum_{t=1}^{r} E_2(p_{ti}/t)\}$$

$$= E_1\{\frac{1}{r} \sum_{t=1}^{r} r_{ti}\}$$

$$= \sum_{t=1}^{R} r_{ti} M_t/N$$

$$= r_i$$

Again

$$V(p_i) = V_1E_2(p_i/t) + E_1V_2(p_i/t)$$

$$= V_1(\frac{1}{r} \sum_{t=1}^{r} r_{ti}) + E_1[\frac{r_{ti}(1-r_{ti})}{m}]$$

So finally, we get the variance covariance matrix in the form given below

$$V = r + (m-1) \sum_{t=1}^{R} W_t \, (r_t - r)(r_t - r)'$$

$$= r + (m-1) \, A_1 \text{ (say)}$$

where, $W_t = \frac{M_t}{N}$ and $r_t = (r_{t1}, ---, r_{t,(k-1)})'$

Denote the eigen value of $P^{-1} A$, by $\rho_1 \geq \rho_2 \geq$, --- $-$, $\geq \rho_{k-1}$.

Then $\lambda i = 1+(m-1)\rho_i$ and

$$X^2 = \sum_{i=1}^{k-1} [1+(m-1)\rho_{io}] \, Wi^2$$

Where ρ_{io} is the value of ρ_i at $r = r_o$.

Also

$$\frac{C' \, A_1 \, C}{C' \, P \, C} \leq 1$$

so the $\rho_{o1} \leq 1$ for any ρ_o and we have

$$\sum_{i=1}^{k-1} Wi^2 \leq [1 + (m-1) \, \rho_{o,k-1}] \sum_{i=1}^{k-1} Wi^2$$

$$\leq X^2$$

$$\leq [1+(m-1) \, \rho_{o1}] \sum_{i=1}^{k-1} Wi^2$$

$$\leq m \sum_{i=1}^{k-1} Wi^2 \qquad \text{[Since } \rho_i \geq o \text{]}$$

Thus, from the above we can conclude that $\frac{X^2}{m}$ gives an asymptotic conservative test The ρ_i is called the **generalized measure of homogeneity**, which is analogous to the measure of homogeneity based on intraclass correlation ρ. The measure ρ is "portable" in the sensethat it is less sensitive to cluster size

than the "deff."

3.3 EFFECT OF SURVEY DESIGN ON TEST OF GENERAL HYPOTHESIS :

Consider the following more general hypothesis about r, which is the vector of ultimate cell proportions. Suppose the hypothesis of interest is

$$H_o : h_i(r) = 0, \quad i=1,2,\ldots,b$$

Assume $\frac{\partial h(r)}{\partial r_j}$ is continuous in the neighbourhood of the true r, $j=1,\ldots,k-1$. Also, $H(r) = \left(\frac{\partial h_i(r)}{\partial r_j}\right)$ and rank of H(r) is b. Now with the help of linearization we get

$$h(p) = h(r) + H(r)(p-r)$$

Also by considering the assumption that

$$\sqrt{n}\,(p-r) \xrightarrow{L} (0, V) \qquad \text{as } n \to \infty$$

We get

$$\sqrt{n}\,[h(p) - h(r)] \xrightarrow{asy} N_b\,[0, H\,V\,H']$$

Where $h(r) = [h_1(r), \ldots, h_b(r)]'$

and $H = H(r)$

If $\hat{V}$, is a consistent estimate of V then we can directly use Wald statistic as given below

$$X^2_w(h) = n\,h'(p)(\hat{H}\,\hat{V}\,\hat{H}')\,h(p) \qquad - (3.8)$$

Where $\hat{H} = H(p)$

Here $X^2_w(h) \sim \chi^2_b$ (under Ho.)

Alternatively, if the sampling design permits the calculation of direct estimates of $\hat{V}_h/_n$ of covariance matrix of h (p), say using BRR or Jackknife methods, we could use this estimate in place of linearization estimate $\hat{H}\,\hat{V}\,\hat{H}'$ in equation (3.8) . For the case of

simple random sampling with replacement which is also equivalent to multinomial sampling we replace V by P_o as given below:

$$X\ (h) = n\ h'(p)\ (\hat{H}_o\ \hat{P}_o\ \hat{H}_o')^{-1} h(p) \qquad - (3.9)$$

Where $\hat{H}_o\ \hat{P}_o\ \hat{H}_o'$ is any estimate of H P H' when our null hypothesis in true, under actual sampling design. This can be applied in secondary data analysis from the published report for which no estimate of V or of HVH' is available. The asymptotic distribution of $X^2(h)$ follows directly from standard results of quadratic form.

Theorem :

Under the null hypothesis Ho: h(r) = 0,

$X^2(h) \approx \sum_{i=1}^{b} \delta_{oi}\ W^2_i$, where δi's are eigen values of

$(H\ P\ H')^{-1}\ (H\ V\ H')$. $\delta_1 \geq \delta_2 \geq$ --- $\delta_b \geq o$ and W_1, W_2, ---, W_b are

independent χ^2_1 random variables and δ_{oi} is the value of δi under Ho.The δi's can be interpreted as design effects of linear combinations, L_i, of the components of H p. Obviously, we have $\lambda_1 \geq \delta i \geq \lambda_{k-1}$ for i=1,2,----b, since L_i are particular linear combinations of p_i 's.

3.4 CHI-SQUARE AS A TEST OF INDEPENDENCE :

Now consider the problem of testing the independence of attributes in two way table. Suppose a two way table is having I-rows and J-columns i.e. total number of cells are equal to IJ, then our hypothesis of interest is

$$H_o: \quad h_{ij}(r) = r_{ij} - r_{i+}r_{+j}$$

$$i = 1,2,--,I-1$$

$$j = 1,2,---,J-1$$

Where $r = (r_{11}, r_{12}, ---, r_{IJ-1})'$

$$r_{i+} = \sum_{j=1}^{J} r_{ij} \quad \text{and} \quad r_{+j} = \sum_{i=1}^{I} r_{ij}$$

where r_{ij} is population proportion in (i,j)th cell. The usual Pearson statistics for testing Ho. is

$$X_I^2 = n \sum_{i=1}^{I} \sum_{j=1}^{J} (p_{ij} - p_{i+}\, p_{+j})^2 / \, p_{i+}\, p_{+j} \qquad - (3.10)$$

which can be rewritten as

$$X_I^2 = n\, h\,(p)' (\hat{P}_I^{-1} \otimes \hat{P}_J^{-1}) h\,(p)$$

Here, $\otimes$ denotes the usual direct matrix product, p_{ij} is the estimate of r_{ij} under the sampling design under consideration, h(p) is the column vector of $h_{ij}(p)$'s, and $\hat{P}_I$ and $\hat{P}_j$ are the values of $P_I = \text{diag}\,(r_I) - r_I r_I'$ and $P_J = \text{diag}\,(r_J) - r_J r_j'$ respectively, for $r = p$, where $r_I = (r_{1+}, ---, r_{I-1+})'$ and $r_J = (r_{+1}, ----, r_{+,J-1})'$ The generalized Wald statistic for testing Ho is given by

$$X^2_{I(W)} = nh(p)' \hat{V}_h^{-1} h(p) - (3.11)$$

Where $\hat{V}_h/n$ is an estimator of V/n, covariance matrix of h(p). The statistic given by (3.11) is asymptotically distributed as χ^2_b for sufficiently large n, where b = (I-1) (J-1). The estimator $\hat{V}_{h/n}$ can be obtained either by linearization method or directly by using balance repeated replication or by Jackknifing method, if the sampling design permits calculation of a direct estimate.

3.4.1 ASYMPTOTIC DISTRIBUTION :

The hypothesis Ho: $h_{ij}(r) = r_{ij} - r_{i+}\, r_{+j}$ is a special case of the general hypothesis Ho: h(r)=0 with k=IJ and b = (I-1)(J-1). Thus X^2_I is of the form $X^2(h)$ hence $X^2_I = \sum_{i=1}^{b} \delta_{oi}\, W_i^2$ under Ho, where δ_i's are eigen values of $D_I = (H\ Po\ H')^{-1}\ (H\ V\ H')$ and δ_{io} is the value of δ_i under Ho.

3.5 CHI-SQUARE TEST AS A TEST OF HOMOGENEITY :

In the case of multinomial sampling, it is well known that the test statistics for independence and homogeneity are identical, but this property does not carry over to more complex sampling designs and the effect on asymptotic distribution of Pearson statistic can be very different in two situations. Consider the problem of testing homogeneity for r populations given independent samples from each population. Let the sizes of the samples from r populations are $n_1, n_2, \ldots, n_r$ respectively. Let $p_i = (p_{i1}, \ldots, p_{i(k-1)})'$ denote the vector of estimated proportions for the i-th sample and suppose that

$$\sqrt{n_i}\,(p_i - r_i) \xrightarrow{L} N(0, V_i)$$

as $n_i \longrightarrow \infty \qquad i=1,2,\ldots,r$

Now our hypothesis which is to be tested is

$$H_0: r_i = r, \qquad i=1,\ldots,r$$

The usual Chi-square test of homogeneity is

$$\chi^2_H = \sum_{i=1}^{r}\sum_{j=1}^{k} n_i \frac{(r_{ij} - p_j)^2}{p_j} \qquad - (3.12)$$

Where $p_j = \sum_{i=1}^{k} \frac{n_i}{\sum_i n_i} p_{ij}$ — The above mentioned statistic has $\chi^2_{(r-1)(k-1)}$ distribution under H_0 with independent multinomial sampling in each population.

3.5.1 ASYMPTOTIC DISTRIBUTION :

For finding the asymptotic distribution, suppose sample size increases together in such a way that $\frac{n_i}{\sum_i n_i} = f_i$; $0 < f_i < 1$, $i=1,2,\ldots,r$. Suppose also that $\sqrt{n_i}\,(p_i - r_i) \xrightarrow{L} N(0, V)$ as $n_i \longrightarrow \infty$. Then if we let p_o be $r(k-1)$ dimensional vector defined by

$$p_o' = (p_1', \ldots, p_r') \qquad \text{it follows that}$$

$$\sqrt{n}\,(p_o - r_o) \xrightarrow{L} N(0, V_o) \text{ as } n \longrightarrow \infty$$

where $n = \sum_{i=1}^{r} n_i, \qquad r_o' = (p', p', \ldots, p')$

and $V_o = \bigotimes_{1}^{r} (V_i/f_i)$

under the assumptions above we can write

$$X^2_H = n(p_o - r_o)' B (p_o - r_o)$$

where $B = F \otimes P^{-1}$

$P = diag(r) - (r)(r)'$

$F = diag(f) - ff'$

$f' = (f_1, ---, f_r)$

Now $Rank\ (B) = R(F) * R(P)$

$= (r-1)(k-1)$

So we have

$$X^2_H = \sum_{i=1}^{(r-1)(k-1)} d_i W_i^2 \quad , \quad W_i \sim N(0,1) \quad \forall \quad i$$

Here, $d_1, d_2, ----, d_{(r-1)(k-1)}$ are the eigen value of

$$A = B V_o$$

$$= \begin{vmatrix} (1-f_1)D_1 & -f_1D_2, ---, & -f_1 D_r \\ -f_2 D_1 & (1-f_2)D_2, ---, & -f_2 D_r \\ | & | & | \\ | & | & | \\ -f_r D_1 & -f_rD_2, --- , & -(1-f_r)D_r \end{vmatrix}$$

where $D_i = P^{-1}V_i$, $i=1,2,--,r$.

3.6 MODIFICATIONS IN TEST STATISTICS FOR SURVEY DATA :

It is clear from the above discussion that ordinary Chi-square statistic for testing any type of hypothesis i.e. goodness of fit, independence of attributes and homogeneity follows weighted chi-square in case of survey data. Where weights are given by the eigenvalues of the corresponding design

matrix. It was also found through simulation experiment that size of the test considerably increases and consequently power of the test reduces. Hence, applying the Chi-square test statistics as such is quite misleading, results in wrong conclusions. There are several modifications which can be classified into two broad headings; these are (i) first order corrections (ii) second order corrections.

3.6.1 FIRST ORDER CORRECTIONS :

Most of the corrections falling in this category are based on division of original chi-square with the averageofthe eigen values corresponding to the design effect matrix.

(a) Correction for goodness of fit test:

In case consistent estimates of $\hat{\lambda}_i$ or λ_{oi}'s are known we can apply the method of Solomon and Stephens (1977) to find out the distribution of weighted Chi-square asymptotically. However, we can know $\hat{\lambda}_i$ or λ_{oi} only if we know V or V_o under Ho. Further if we know V or V_o we can also apply the Wald statistic which is simple, but not very consistent in some of the particular situations. So, it is better to have some simple approximation to the asymptotic distribution of X^2 that requires only limited information about $\hat{V}$. Simplest approach is to modify the statistic as given below :

$$X^2_c = X / \hat{\lambda}_\bullet . \qquad - (3.13)$$

The X^2_c is distributed asymptotically as

$Y = \sum_{i=1}^{k-1} \left(\frac{\lambda oi}{\lambda_o}\right) W_{i,}^2$ as χ^2_{k-1} random variable under

Ho, where

$$\lambda . = \sum_{i=1}^{k-1} \hat{\lambda}i/(k-1)$$

$$\lambda o. = \sum_{i=1}^{k-1} \frac{\lambda oi}{k-1} \qquad (3.31)$$

and $\hat{\lambda}i$'s are the eigen values of $\hat{D} = \hat{P}^{-1}\hat{V}$

where $\hat{P} = \text{diag}(p) - pp'$

Now we have $E(y) = \sum_{i=1}^{k-1} \frac{\lambda oi}{\lambda o} E(W_i^2) = \sum_{i=1}^{k-1} \frac{\lambda oi}{\lambda o} = (k-1)$

$[\because E(Wi^2) = 1]$

So, $E(y) = E(\chi^2_{k-1}) = k-1$ - (3.14)

Again

$$V(y) = E(y^2) - [E(y)]^2 \qquad - (3.15)$$

Consider $E(y^2) = E\left[\sum_{i=1}^{k-1} \frac{\lambda^2 oi}{\lambda o^2} Wi^4 + \sum_{i \neq i'}^{k-1} \frac{\lambda oi\ \lambda oi'}{\lambda o^2} Wi^2\ Wi'^2\right]$

$$= \sum_{i=1}^{k-1} \frac{\lambda^2 oi}{\lambda o^2} E(Wi^4) + \sum_{i \neq i'}^{k-1} \frac{\lambda oi\ \lambda oi'}{\lambda o^2}$$

$[\because Wi^2 s \text{ are i.i.d}]$

$$= \left[\sum_{i=1}^{k-1} \frac{\lambda^2 oi}{\lambda o^2} \{ V(Wi^2) + [E(W^2 i)]^2 \}\right]$$

$$+ \sum_{i \neq i'}^{k-1} \frac{\lambda oi\ \lambda oi'}{\lambda o^2}$$

$$= 3 \sum_{i=1}^{k-1} \frac{\lambda^2_{oi}}{\lambda_o^2} + (k-1)^2 - \sum_{i=1}^{k-1} \frac{\lambda^2_{oi}}{\lambda_o^2}$$

$$= 2 \sum_{i=1}^{k-1} \frac{\lambda^2_{oi}}{\lambda_o^2} + (k-1)^2$$

Putting the value of $E(y^2)$ in (3.15) we get

$$V(y) = 2 \sum_{i=1}^{k-1} \frac{\lambda_{oi}{}^2}{\lambda_o^2} + (k-1)^2 - (k-1)^2$$

$$= 2 \sum_{i=1}^{k-1} \frac{\lambda^2_{oi}}{\lambda_o{}^2}$$

The above equation can also be rewritten as

$$V(y) = 2(k-1) + 2 \sum_{i=1}^{k-1} \frac{(\lambda_{oi} - \lambda_o)^2}{\lambda_o^2} \qquad - (3.16)$$

So we can see that it is larger than $V(\chi^2_{k-1}) = 2(k-1)$ unless all λ_{oi}'s are equal. The most important point in this modification is that $\hat{\lambda}.$ Only depends upon the estimated variances of cell proportions (or equivalently the estimated cell deffs)

Since,

$$\hat{\lambda}. = \frac{\text{Trace } (\hat{P}^{-1}\hat{V})}{k-1}$$

$$= \sum_{i=1}^{k} \frac{\hat{\nu}_{ii}}{p_i(k-1)} = \sum_{i=1}^{k} \frac{\nu_{ii}}{p_i(1-p_i)} \frac{(1-p_i)}{(k-1)}$$

$$= \sum_{i=1}^{k} \frac{d_i (1-p_i)}{k-1}$$

where $d_i = \frac{\hat{\nu}_{ii}}{p_i(1-p_i)}$ which is estimated design effect of the i-th cell. It can also be noted that $\hat{\lambda}i$ is not generally the

same as average cell deff $\hat{d}_. = \sum_{i=1}^{k} \hat{d}_i / k$ It is also well known that some information about the estimated cell deffi is often available but the knowledge about the covariance term is less common particularly in secondary data analysis from published reports. Similarly, we can modify the test statistic in case of testing the independence of attributes and homogeneity of proportions from several populations. A brief description about them is given below.

(b) **Correction for the test of independence:** Let us consider the first order approximation in the case of testing of independence of attributes, where our null hypothesis is of the form.

$$H_0 : h_{ij}(r) = r_{ij} - r_{i+}\, r_{+j}$$

$$i = 1, 2, \text{----}, I; \quad j = 1, 2, \text{----}, J$$

A modified test statistic is similar to the X_c^2 for goodness of fit problem and is given by

$$X^2_{I(C)} = \frac{X_I^2}{\hat{\delta}_.} \qquad \text{----------(3.17)}$$

where $\hat{\delta}_.$ is given by

$$\hat{\delta}_. = \sum_{i=1}^{I} \sum_{j=1}^{J} \hat{\nu}_{ij}(h)/(b\, p_{i+}\, p_{+j})$$

$$= \sum_{i=1}^{I} \sum_{j=1}^{J} (1 - p_{i+})(1 - p_{+j})\, \hat{\delta}_{ij}/b$$

Hence, $\hat{\nu}_{ij}(h)/n$ is the estimator of variance of $h_{ij}(p)$ and $\hat{\delta}_{ij}$ is the estimated design effect of $h_{ij}(p)$ i.e.

$$\hat{\delta}_{ij} = \frac{n\hat{\nu}_{ij}(h)}{p_{i+}\, p_{+j}\,(1-p_{i+})(1-p_{+j})}$$

But $\hat{\delta}_{ij}$'s are seldom available in the published reports so $\hat{\lambda}.$ is an adequate substitute under this conditions

$$\hat{\lambda}. = \sum_{i=1}^{I} \sum_{j=1}^{J} (1 - p_{ij})\, \hat{d}_{ij}/(IJ-1)$$

However, even $\hat{\lambda}.$ requires the estimated 'deff' $\hat{d}_{ij}$ for all the individual cell estimates p_{ij}, and this information may not be available, especially for large two-way tables.The best we can do for many applications is to have some information on deffs of marginal row and column proportions. Ideally we would like an approximation for $\hat{\delta}.$ based on marginal deffs but it is found to be in consistent with almost any value of $\hat{\delta}.$ There is fair amount of empirical evidence that suggests that $\hat{\delta}.$ tends to be smaller than average deff of either margin in parctice. Thus, it may be possible to find empirically based approximations that work well in practice.

(C) Correction for the test of Heterogeneity : Now first order correction to the test of heterogeneity will be considered briefly because it can be obtained by proceeding on the similar lines as in previous cases. As previously shown that

$$\chi^2_H = \sum_{i=1}^{(r-1)(k-1)} d_i W_i^2 \qquad - (3.18)$$

Where di's are eigen values of A

$$A = \begin{bmatrix} (1-f_1)D_1 & -f_1D_2, & \cdots, & -f_1D_r \\ -f_2D_1 & (1-f_2)D_2 & \cdots, & -f_2D_r \\ \vdots & \vdots & & \vdots \\ -f_rD_r & - - - - & & (1-f_r)D_r \end{bmatrix}$$

where $f_i = \frac{n_i}{n}$ and $D_i = P^{-1}V_i$ is the design effect matrix for the i-th population $i = 1,2, --, r$. As usual the ordinary Chi-square test is conservative with proportionally allocated stratified sampling. For more complex design we can fall back on the modified test statitics $X^2_H / \bar{d}$. Now

$$(r-1)(k-1)\bar{d} = \text{Trace}(A) = \sum_{i=1}^{r} (1-f_i)\ \text{Trace}\ (D_i)$$

so that

$$\bar{d} = \sum_{i=1}^{r} \frac{(1-f_i)}{(r-1)} \bar{d}_i$$

where, $\bar{d}_i$ is the average design effect of the i-th population. Thus d can still be calculated simply from information about cell variances for each population. Notice, that as r becomes large $\bar{d}$ will tend towards the unweighted average of the di's provided no single n_i dominates the others. Notice, also that $\bar{d}$ is simply a weighted average of a population design effects and should stay relatively stable as r increases.

In general we can say that if our null hypothesis is of the form Ho: $h(r) = 0$ and $X^2(h) \approx \sum_{i=1}^{b} \delta_{oi} W_i^2$ then the first order correction can be applied as

$$X^2_c(b) = \frac{X^2(h)}{\hat{\delta}.} \qquad - (3.19)$$

where $\hat{\delta}. = \sum_{}^{b} \delta i/b$ and $\hat{\delta}.$ is consistent estimator of δi under Ho: for example, $\hat{\delta i}$'s could be eigen values of $(\hat{H}_o \hat{P}_o \hat{H}_o')^{-1}(\hat{H}\hat{V}\hat{H}')$ or $(\hat{H}_o \hat{P}_o \hat{H}_o')^{-1}\hat{V}_h$. However, in general $\hat{\delta}.$ requires the knowledge of full $\hat{V}$ or $\hat{V}_h$ unlike $\hat{\lambda}.$ It has been shown by Das Gupta (1953) that $\delta.$ lies between the average of b largest λi's and b smallest λi's so that $\delta.$ should be close to $\lambda.$ if b is large compared to (k-1). The δi's can be interpreted as design effects of linear combinations of L_i of the components of H P. Obviously we have $\lambda_1 \geq \delta i \geq \lambda_{k-1}$ for i=1, -- b, since L_i are particular linear combinations of p_i's.

3.6.2 SECOND ORDER CORRECTION :

As it may be noted from the above discussion that first order correction to the chi-square statistic is based on the principle of standardization of the Chi-square statistics obtained from the survey data such that its expected value should be equal to the standard Chi-square test statistics i.e. Chi-square statistic from the data obtained through simple random sampling or from infinite population. As far as second order corrections are concerned we standardize our Chi-square test statistic such that, its expected value should be equal to its degrees of freedom and its variance should be equal to twice its degrees of freedom as expected in the standard case. For the second order approximation we take the help of Satterthwaite approximation (1946), so this method is also known as

Satterthwaite approximation method. Here, we discuss briefly only the second order correction for testing the goodness of fit, other, second order correction for independence of the attributes will be discussed in the next chapters. Suppose we have the information about $\sum_{i=1}^{k} \lambda_i^2$ then we can apply the Satterthwaite approximation as follows

$$X_s^2 = \frac{X_c^2}{1+\hat{a}^2} \sim \chi^2_{\nu}$$

where $\nu = \dfrac{(k-1)}{(1+\hat{a}^2)}$

and $\hat{a} = \sum_{i=1}^{k} \dfrac{(\hat{\lambda}_i - \hat{\lambda}_.)^2}{[(k-1)\ \hat{\lambda}^2_.]}$ which is square of the coefficient of variation of $\hat{\lambda}_i$'s. Note that

$$\sum_{i=1}^{k-1} \lambda_i^2 = \sum_{i=1}^{k} \sum_{j=1}^{k} \hat{\nu}_{ij}^2 / (p_i p_j)$$

$$[\because V(\sum_{i=1}^{k} \hat{\lambda}_i W_i^2) = 2\sum_{i=1}^{k} \hat{\lambda}_i^2]$$

so that $\hat{a}^2$ and ν can be easily calculated from $\hat{V}$. It is also clear that treating X^2_c as χ^2_{k-1} under Ho will tend to under estimate the upper percentage points of true asymptotic distribution, since $V(y) \geq V(\chi^2_{k-1})$. However, this effect will be small if the coefficient of variation of the λ_i's is not too large.

3.7 CONSTANT DESIGN EFFECT :

For a given sampling design p(s) the design effect (deff)

as defined by Kish is the ratio of variance of an estimate T under the sampling design under consideration with respect to the variance of the estimate under simple random sampling with replacement. Mathematically it is given as

$$deff(T) = \frac{Var\ (T)}{Var_{srs}(T)}$$

where Var_{srs} (T) denotes the variance under simple random sampling with replacement. From the above definition it is clear that "deff" measures the deviation in variance which is due to any sampling design with respect to simple random sampling with replacement.

Let us suppose design effects of p_i's are constant i.e.

$$V(p_i) = \delta\ V_{srs}(p_i) = \delta\ r_{oi}(1-r_{oi})/n$$

where, $i=1,2,--,k$

The distribution of X^2 is $\sum_{i=1}^{k-1} \delta_i W^2_i$ so the expected value of X^2 is

$$E(X^2) = \sum_{i=1}^{k-1} \delta_i \quad E(W^2_i)$$

$$= \sum_{i=1}^{k-1} \delta_i . 1$$

$$= Tr(P_o^{-1}V)$$

where $P_o = diag\ (r_o) - r_o\ r_o'$

so the inverse of P_o is given as

$$P_o^{-1} = [diag(r_o)]^{-1} + \frac{1}{r_{ok}}\ 1\ 1'$$

$$\text{Hence, } P_o^{-1}V = [diag(r_o)]^{-1}\ V + \frac{1}{r_{ok}}\ 1\ 1'\ V$$

So

$$Tr\ (P_o^{-1}\ V) = Tr\ [\{diag(r_o)\}^{-1}V] + Tr[\frac{1}{r_{ok}}\ 1\ 1'V]$$

$$= \sum_{i=1}^{k-1} \frac{\nu_{ii}}{r_{oi}} + \frac{1}{r_{ok}} \sum_{i=1}^{k-1}\sum_{j=1}^{k-1} \nu_{ij}$$

$$= \sum_{i=1}^{k-1} \frac{\nu_{ii}}{r_{oi}} + \frac{1}{r_{ok}}\ Var(1-r_{o1} \text{ ---- } r_{ok-1})$$

$$= \sum_{i=1}^{k-1} \frac{\nu_{ii}}{r_{oi}} + \frac{\nu_{kk}}{r_{ok}} = \sum_{i=1}^{k} \frac{\nu_{ii}}{r_{oi}}$$

Hence, with the help of $Tr(P_o^{-1}V)$ we have

$$E(X^2) = \sum_{i=1}^{k} \frac{\delta\ r_{oi}(1-r_{oi})}{r_{oi}}$$

$$= \delta \sum_{i=1}^{k} (1-r_{oi}) = \delta(k-1)$$

Hence, $\frac{X^2}{\delta}$ has the same first moment as a χ^2 random variable with (k-1) degrees of freedom. Thus when design effects of cells are assumed to be constant, the Pearson statistic will have, as a first order approximation, the distribution of χ^2 random variable with (k-1) degrees of freedom. In other words, for complex sampling designs with fixed design effects, the tests are based on X^2 can be approximately done using χ^2 table. It can be shown the except for case k=2 these assumptions do not imply that V(p) is equal to δP_o.

3.7.1 Constant Design Effect Models:

(I) Altham's Model for Two-stage sampling:

Consider that the survey population consists of R PSU's and let M_t the number of SSU's in the t-th PSU, t=1,2,--,R; such that $\sum_{t=1}^{R} M_t = N$

Define

$Z_{t\lambda i}$ = 1, if λ-th population element in the t-th PSU is in i-th category.

= 0, otherwise

Model approach assumes $Z_{t\lambda i}$ as a random variable with assumed distributional properties and the expectations, are taken with respect to the assumed model. Following Altham (1976). We assume the following general model: The random variables $Z_{t\lambda i}$ in different clusters are independent and

$$\epsilon(Z_{t\lambda i}) = \pi i \qquad -(3.20)$$

and $\epsilon\ (Z_{t\lambda i} - \pi i)\ (Z_{t\mu j} - \pi j) = bij,\ \lambda \neq \mu$.

where ϵ is expectation operator with respect to the model. It can be observed that values in the above model (3.20) are independent of the PSU label, so a form of exchangeability between clusters can be assumed. Also, this model is not suitable if the clusters are different. Let us denote the two stage sample by s, where $s = (s_1, s_2, ---, s_r)$ where s_l denotes the sub-sample of size m such that $\sum_{l=1}^{r} m_l = n$ where r is total no of PSU's selected. Then sample cell frequency can be written as

$$n_i = \sum_{l=1}^{r} \sum_{\lambda=1}^{m_l} Z_{l\lambda i} = \sum_{l=1}^{r} m_{li}, \text{(say)} \quad -(3.21)$$

Hence $$\epsilon\left(\frac{n_i}{n}\right) = \frac{1}{n}\sum_{l=1}^{r}\sum_{\lambda=1}^{m_l}\epsilon(Z_{l\lambda i}) = \pi_i \quad \text{for every s}$$

This indicates that $\hat{\pi}_i = \frac{n_i}{n}$ is a model unbaised for π_i. Let $n = (n_1, n_2, \cdots, n_{k-1})'$. Also we know that

$$n_i = \sum_{l=1}^{r}\sum_{\lambda=1}^{m_l} Z_{l\lambda i}$$

$$V(n_i) = \sum_{l=1}^{r}\sum_{\lambda=1}^{m_l} V(Z_{l\lambda i}) + \sum_{l=1}^{r}\sum_{\lambda\neq\lambda'}^{m_l} \mathrm{Cov}(Z_{l\lambda i}, Z_{l\lambda' i})$$

$$+ \sum_{l\neq l'}^{r}\sum_{\lambda=1}^{m_l} \mathrm{Cov}(Z_{l\lambda i}, Z_{l'\lambda i}) + \sum_{l\neq l'}^{r}\sum_{\lambda\neq\lambda'}^{m_l} \mathrm{Cov}(Z_{l\lambda i}, Z_{l'\lambda i})$$

$$= \sum_{l=1}^{r} m_l(1-\pi_i)\pi_i + \sum_{l=1}^{r} m_l(m_l-1)b_{ii}$$

$$\because V(Z_{l\lambda i}) = \epsilon(Z_{l\lambda i}^2) - [\epsilon(Z_{l\lambda i})]^2$$

$$= \pi_i - \pi_i^2 = [\pi_i(1-\pi_i)]$$

Also clusters are independent of each other.

$$V(n_i) = n\,\pi_i(1-\pi_i) + [\sum_{l=1}^{r} m_l^2 - n]b_{ii}$$

The above equation can be written in the matrix notation as

$$n\,V_s = n\Delta + [\sum_{l=1}^{r} m_l^2 - n]B$$

where

$$V_s = V(n/n)$$

$$\Delta = \mathrm{diag}[(\Pi) - \Pi\Pi']$$

$$B = (b_{ij})$$

$$\Pi = (\pi_1, \pi_2, \cdots, \pi_{k-1})'$$

or $$V_s = \Delta + (m_{os} - 1)B \qquad - (3.22)$$

where $m_{os} = \sum_{l=1}^{r} m_l^2/n$.

From the above equation we can find a close similarity between the two-stage sampling design expression for the variance-covariance matrix . The only difference is that equation in the previous case is obtained with thehelp of sampling design PPSWR and requires no assumptions about population elements, while in this case we need the assumption of exchangeability of model(3.20). Of course, the design and model are usually intimately related and we would normally feel more secure about the results based on the exchangeable model if the sample was choosen with self-weighting design. The general hypothesis about the parameters of the model is given by

$$H_o: \; h_i(\Pi) = 0, \qquad i=1,2,\text{---},b$$

so, the $X^2(h)$ is given by

$$\chi^2(h) = n\, h(\hat{\Pi})' \, (\hat{H}_o \hat{P}_o \hat{H}_o')^{-1} \, h(\hat{\Pi})$$

Since, m_{li}'s in different clusters are assumed to be independent, it follows that

$$\sqrt{n}(\hat{\Pi}-\Pi) \approx N(0, V_s) \text{ for larger } n$$

Hence,

$$\sqrt{n}\; [h(\hat{\Pi}) - h(\Pi)], H\, V_s H')$$

where $H = H(\Pi)$

But as we know

$$X^2(h) = \sum_{i=1}^{b} \bar{\delta}_{oi} \, W_i^{\,2}$$

where $\bar{\delta}_{oi} = 1 + (m_{os} - 1)\, \bar{\rho}_i(h)$

where $\bar{\rho}_i(h)$ = eigen values of $(H \Delta H')^{-1} (H\, B\, H')$

Lemma (Altham):

$$B^{*} = \Delta - B \text{ is non-negative definite.}$$

Proof : Consider $B^{*} = \Delta - B$ and let $\pi_{ij} = \epsilon(Z_{i\lambda i}\ Z_{i\mu j})$

We can write $\pi_{i} = \pi_{ii} + \sum_{j \neq i}^{k} \pi_{ij}$

So,
$$C'B^{*}C = \sum_{i=1}^{k-1} \pi_{ij}\, C_i^2 + \sum_{i<j=1}^{k} \sum \pi_{ij}(C_i - C_j)^2 \geq 0$$

Since, $\pi_{ij} \geq 0$

Hence, with the help of above lemma we can say that $\rho_i^{-}(h) \leq 1$ for i=1,2,--,b. So conservative test can be given as

$$\frac{X^2(h)}{mos} \sim \chi^2_{k-1} \qquad \text{(under Ho)}$$

3.7.2 Constant Design Effect for Stratified two stage Sampling:

Consider the case where whole population is divided into L strata. Let R_h be the number of PSU's in the h-th stratum, M_{ht} is the number of SSU's in the t-th PSU of the h-th stratum, is a stratified two-stage sample in which γ_h is the number of PSU's selected from the h-th stratum and m_h is the number of SSU's selected in each of the selected PSU's.

Let $Z_{ht\lambda i}$ = 1, if (h,t,λ)-th element is in the i-th category.

= 0 otherwise.

Further, assume that model (3.20) holds for each stratum seperately Then, if $n_h = (n_{h1}, \text{--------} n_{hk-1})'E(n_h) = n_h \Pi_h$. and the covariance matrix of n_h can be expressed as

$$\bar{n}_h V_h = \bar{n}_h \Delta_h + \bar{n}_h(m_h - 1) B_h$$

where $\bar{n}_h = \gamma_h m_h$

$\Delta_h = \text{diag}\ (\Pi_h) - \Pi_h\ \Pi_h'$

Let $\Pi = \sum_{h=1}^{L} W_h\ \Pi_h$

Then $\hat{\Pi} = \sum_{h=1}^{L} W_h\ n_h/\bar{n}_h$

where W_h is the weight of the h-th stratum

$E\ (\hat{\Pi}) = \Pi$

The covariance matrix of Π is given as

$\frac{V_s}{n} = \sum_{h=1}^{L} W_h^2\ V_{hs}/\ \bar{n}_h$

In the case of proportional allocation with

$\bar{n}_h/n = W_h$, V_s reduces to

$Vs = \sum_{h=1}^{L} W_h\ [\Delta_h + (m_h - 1)B_h]$

But as we know $\Delta_h - B_h$ is non-negative definite as in case of previous section.We get $C'V_sC \leq \sum W_h\ m_h\ (C'\Delta_hC) \leq m^*\ C'(\Sigma W_h\ \Delta_h)C$

$= m^*C\ [\Delta - \sum_h W_h\ (\Pi_h - \Pi)(\Pi_h - \Pi)']C$

$\leq m^*C\ \Delta\ C$

where $m^* = \max\ (m_h)$

Thus the largest eigen value of $(H\ \Delta\ H')^{-1}(H\ V_s\ H')$ does not exceed m^*. Hence $\frac{X^2(h)}{m^*}$ provides an asymptotic conservative test.

When $B_h = \rho\Delta_h$, $m_h = m \quad \forall\ h$

$C'\Delta\ C = [1+(m-1)\rho]\ C'\ (\sum\ W_h\ \Delta_h)C$

$\leq C'[1+(m-1)\rho]C$

$\leq [1+(m-1)\rho]\ C'\Delta\ C$

Hence, $X^2/[1+(m-1)\rho]$ is asymptotic conservative test provided a value for ρ can be specified.

ANALYSIS OF MULTIDIMENSIONAL CONTINGENCY TABLE

ANALYSIS OF MULTIDIMENSIONAL CONTINGENCY TABLE

4.1 INTRODUCTION:

In categorical data analysis, when the observations have more than one characteristics of interest, it is often the case that we would like to study how these characteristics interrelate. The study of these associations and interactions can be nicely formulated using log-linear models. For the secondary analysis from published reports containing multiway tables, the researchers may not have access to the necessary information (e.g. the full estimated covariance matrix of cell estimates) for implementing the methods such as Wald statistics or developed by Nathan (1975) etc. At best report might contain some information about variance estimates (design effects) for marginal totals or cells. Consequently, it is of importance to assess the impact of survey design on standard multinomial based methods and suggest simple corrections requiring only minimal information on the design effects(abbreviated "deffs"). Even when the necessary information is available, it is not clear that methods based on Wald sttistics would necessarily perform well in finite samples, especially when the number of cells in the table increases . This also leads to unstable sample estimates of the covariance matrix. It would be desirable to obtain improved corrections to standard methods utilizing the detailed information and study their finite sample properties

relative to those of Wald statistics and others. In this chapter we discuss some of the methods particularly the papers by Rao and Scott(1984) using log-linear for multiway classifications. These methods have been used in developing the relevant theory and test procedure in subsequent chapters.

4.2 NOTATION AND BACKGROUND :

Suppose that we have an r-dimensional contingency table with independent variables $x_1, x_2, \ldots, x_r$, each having respectively $\nu_1, \nu_2, \ldots, \nu_r$ categories. When $r = 3$, the indices i,j,k can be used to denote a given cell in a table. For example r_{ijk} will denote the expected proportions in the cell i,j,k. This notation can be generalized by using a single symbol usually θ, to denote complete set of subscripts. Thus, r_θ will be the expected proportion of an elementary cell θ.

We generally consider hierarchical models as defined by Birch(1963). This means that the cell expectations are permitted to be log-linearly related in such a way that a suitable set of marginals , usually called the minimal set of fitted marginals, is sufficient for the parameters. Tables of sums of none elementary cell will be caled configurations and will be denoted by letter C (Bishop et.al.(1975)) . For example, in a three-way contingency table, the table of partial sums $x_{ij+} = \sum_k x_{ijk}$ obtained by summing over the third variable, will be denoted by C_{12}. As the third variable has been removed by summing, the subscripts of C refer only to the remaining two variables.

Configurations corresponding to minimal set of fitted marginals, as defined above, will be called the sufficient configurations.

Bishop et al. (1975, page 68) outlined a method to derive sufficient configurations for comprehensive, unsaturated and hierarchical models. For such models, if sufficient configurations are given, it is trivial to write down the log-likelihood function, log m_θ. Also , it can be shown that number of independent parameters in the model can be expressed in terms of numbers of cells in the sufficient configurations.

Indeed, when only C_θ is the sufficient configuration of the model, it is clear that the number of independent variables in the model is equal to the number of cells in C_θ. In other words, if u_θ is the set of all linearly independent μ - terms whose subscripts are subsets of θ. (which is, in this case, the set of all linearly independent parameters for the model) then the cardinality of u_θ is equal to the number of cells in C_θ . This result implies that if the sufficient configurations are C_{θ_i}, i=1,...,k and u_{θ_i}'s sets are defined as above then $u_\theta = u_{\theta_1} \cup u_{\theta_2},..,u_{\theta_l}$ will be a set of a of all linearly independent parameters and cardinality ofu can be fouund using the inclusion-exclusion principle. For instance, the three dimensional contingency table with no three factor effect,

$$\log m_{ijk} = \mu + \mu_1(i) + \mu_2(j) + \mu_3(k) + \mu_{12}(ij) + \mu_{13}(ik) + \mu_{23}(jk) \quad \text{with } i=1,...,I;\ j = 1,2,...,J;$$

and k = 1,2,...,K, has C_{12}, C_{13}, and C_{23} as sufficient

configurations. Let u be the set of all linearly independent parameters then

$$\begin{aligned} card(u) &= 1+ (I-1)+(J-1)+(k-1)+(I-1)(J-1) \\ &\quad + (I-1)(K-1)+(J-1)(K-1) \\ &= IJ + IK + JK - I-J-K+1 \\ &= card(u_{12}) + card(u_{13}) + card(u_{23}) - card (u_{12} \cap u_{13}) \\ &\quad - card (u_{12} \cap u_{23}) - card(u_{13} \cap u_{23}) + card(u_{12} \cap u_{13} \cap u_{23}). \end{aligned}$$

The formula for the number of independent variables will be simpler if the hierarchichal log-linear model is **decomposable.** A hierarchical model with sufficient configurations $C_{\theta i}$, i=1,2,---,I is decomposable if and only if the class $\{\theta i\}$ can be observed in such a way that each θi is composed of one set of elements which are missing in all θs for $s>i$ and one set ϕ_i which is contained in some θ_r, for some $r>i$. In other words, we have

$$\theta i = \theta i^{*} \cup \phi i$$

with $\theta i^{*} \cap \phi i = 0$, $\theta i^{*} \cup_{j>i} \theta j = 0$ and $\phi i \subset \theta_s$ for some $s > i$ .....(A)

Further if such an ordering is possible, a version may be found in which any prescribed set is the last one. For example, three way contingency table with sufficient configuration C_{12}, C_{13} and C_{23} is not decomposable, since the subscripts of any C can not be decomposed into two disjoint subsets satisfying (A) but the seven dimensional hierarchical log-linear model with sufficient configurations C_{123}, C_{124}, C_{235}, C_{136} and C_{57} is decomposable. An ordering of θi which has {5,7} as the last set is

{1,2,4}, {1,3,6}, {1,2,3}, {2,3,5}, {5,7}

where the bold elements do not belong to any set that follows. An ordening that has {1,3,6} as the least set is

{5,7}, {2,3,5}, {1,2,4}, {1,2,3}, {1,3,6)

Usually, to obtain a particular ordering, it will be easier to start with the last set and work backwards.

4.3 MULTINOMIAL SAMPLING :

Here, some basic results about the log-linear model under the multinomial sampling will be discussed just to introduce the application of log-linear models in the categorical data analysis. The standard results for these models are given in Bishop et al. (1975), Fienberg(1980) and Agrestti (1990). Let $r = (r_1,\ldots,r_T)$ be the vector of cell proportions such that $\sum_{i=1}^{k} r_i = 1$. The observedcounts in each cell from the random sample of size n is given by $n = (n_1, n_2 \ldots, n_T)'$. Now, as we know that n has a multinomial distribution, where $\sum n_i = n$. Let $p = n/n$ and define

$$\mu = \log r$$

The log-linear model assumes that for a parameter vector $\theta = (\theta_1,\ldots, \theta_r)'$, we have

$$\mu(\theta) = \tilde{\mu}(\theta)1 + X\theta \qquad \text{---} \qquad (4.1)$$

Where X is known T x r matrix of full rank $r(\leq T-1)$ and $X'1 = 0$, 1 is a T-vector of 1's If r = (T-1), we have a saturated model. For instance general log-linear model for 2x2x2 may be written as

$$\mu_{ijK} = \tilde{\mu} + \mu_1(i) + \mu_2(j) + \mu_3(k) + \mu_{12}(ij) +$$

$$\mu_{13}(ik)+\mu_{23}(jk)+\mu_{123}(i,j,k)$$

Where $i=1,2;\ j=1,2;\ k=1,2.$

Now we impose the following constraints.

$$\sum_{i=1}^{2} \mu_1(i)= 0 \rightarrow \mu_1(1)= -\mu_1(2)$$

$$\sum_{j}^{2} \mu_{12}(ij) = 0 = \sum_{i} \mu_{12}(ij) = \sum_{ij} \mu_{12}(i,j) \text{ etc.}$$

similarly, $\mu_{12}(21) = -\mu_{12}(11)$ etc.

Here

$$\theta= \begin{bmatrix} \mu_1(1) \\ \mu_2(1) \\ \mu_3(1) \\ \mu_{12}(11) \\ \mu_{13}(11) \\ \mu_{23}(11) \\ \mu_{123}(11) \end{bmatrix}, \quad \mu= \begin{bmatrix} \mu_{111} \\ \mu_{112} \\ \mu_{121} \\ \mu_{122} \\ \mu_{211} \\ \mu_{211} \\ \mu_{212} \\ \mu_{221} \\ \mu_{222} \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 1 & -1 & -1 \\ 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ -1 & 1 & 1 & -1 & -1 & 1 & -1 \\ -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ -1 & -1 & -1 & 1 & 1 & 1 & -1 \end{bmatrix}$$

So we have

$$\log r = \tilde{\mu}(\theta)1 + X\theta \qquad (4.2)$$

$$r = \exp\{\tilde{\mu}(\theta)1\}\exp\{X\theta\}$$

$$1'r = 1'\exp\{\tilde{\mu}(\theta)1\}\exp\{X\theta\}$$

$$1 = 1'\exp\{\tilde{\mu}(\theta)1\}\exp\{X\theta\}$$

$$\tilde{\mu}(\theta) = \log\left(1/\,1'\exp X\theta\right)$$

This is also known as normalizing factor. Under multinomial sampling, it is well known that likelihood equations are given by

$$X'\hat{r} = X'n/n \qquad -(4.3)$$

where $\hat{r} = r(\hat{\theta})$ is the maximum likelihood estimator (MLE) of r under the model $\sum_t r_t = 1$. The method of Iterative proportional fitting (IPF) is often used to determine (4.3) whenever (4.3) does not admit explicit solution. The MLE $\hat{r}$ are easily obtained for hierarchical model.Now, from Bishop, Fienberg and Holland(BFH,1975) Section 14.8.1 we have the following results.

$$\sqrt{n}\,(\hat{\theta} - \theta) \xrightarrow{asy} N\,[o\;\;(X'P\;X)^{-1}] \qquad -(4.4)$$

$$\sqrt{n}\,(\hat{r} - r) \xrightarrow{asy} N\,[0\;\;P\,X\,(X'P\,X)^{-1}X'P] \qquad -(4.5)$$

These results are asymptotic in distribution.

Suppose now that linear expression $X\theta$ can be decomposed as $X_1\theta_1 + X_2\theta_2$ where X_1 and X_2 are full rank, X_1 is Txs and X_2 is Txu accordingly θ_1 is sx1 and θ_2 is ux1 (s+u=r) Consider the problem of testing

$$Ho:\ \theta_2 = 0$$

against the alternative

$$H_1 : \theta_2 \neq 0$$

Let $\hat{\theta}_1, \hat{\theta}_2, \hat{r}$ etc. be the maximum likelihood estimates under the full model H_1. Alternatively, let $\hat{\hat{\theta}}_1$ and $\hat{\hat{r}}$. denote the estimate under Ho. The likelihood ratio statistic for the above hypothesis $G^2 = 2n \sum \hat{p}_t \log (\hat{p}_t/\hat{\hat{r}}_t) - 2n \sum \hat{p}_t \log (\hat{p}_t/\hat{r})$

Under Ho, this statistic has asymptotically χ^2 distribution with u degrees of freedom. This statistic is also asymptotically equivalent to the Pearson statistic.

$$X^2_P = n(\hat{r} - \hat{\hat{r}})' D_\pi^{-1} (\hat{r} - \hat{\hat{r}}) \qquad - (4.6)$$

and the Wald statistic

$$X^2_w = n \hat{\hat{\theta}}' X_2' \hat{P} X_2 \hat{\hat{\theta}} \qquad -(4.7)$$

where D_π = diag(r)

$$P = D_\pi - r r' \qquad - (4.8)$$

4.4 EFFECT OF SAMPLING DESIGN :

Suppose a sample, $\tilde{s}$, of n ultimate units is drawn according to a specified survey design, $p(\tilde{s})$, and let $\hat{p}$ denote a consistent estimater of r under $p(\tilde{s})$, $\sum \hat{p}_i = 1$. Assume a central limittheorem for the specified design is available which ensures that $\sqrt{n}\,(\hat{p} - r)$ converges in distribution to a $N_T(0,V)$ random vectors, say y, as $n \longrightarrow \infty$, i.e. $\hat{p}$ is approximately T-variate normal with mean vector 0 and singular covariance matrix V/n for sufficiently large n.

Let $\mathcal{S}_T$ be set of all possible T-dimensional probability vector i.e. $\mathcal{S}_T = \{r: r_i \geq 0$ and $\sum_{i=1}^{T} r_i = 1\}$ and Θ is a subset of r-dimensional Euclidean space called parameter space as θ ranges

over the values of $\odot$, $r(\hat{\theta})$ ranges over a subset M of f_T. Assume the following regularity conditions which are given by Birch (1964) along with the assumption that model is correct i.e. $r=r(\theta)$. Also throughout we assume $r < (T-1)$.

(i) The point θ is an interior point of $\odot$, so that θ is not on the boundary of $\odot$ and there are r-dimensional neighborhood of θ that is completely contained in $\odot$.

(ii) $r_i = r_i(\theta) > 0$ for all i=1,2,--,T. Thur r is an interior point of f_T and does not lie on the boundary of f_T.

(iii) The mapping $r : \odot \longrightarrow f_T$ is total differentiable at θ, so that partial derivatives of r_i with respect to $\hat{\theta}_j$ exist at θ and $r(\hat{\theta})$ has a linear approximation at θ given by.

$$r_i(\hat{\theta}) = r_i(\theta) + \sum_{j=1}^{r} (\hat{\theta}_j - \theta_j) \frac{\partial r_i(\theta)}{\partial \hat{\theta}_j}$$

$$+ O(\| \hat{\theta} - \theta \|); \text{ as } \hat{\theta} \longrightarrow \theta$$

(iv) The Jacobian matrix $\left(\frac{\partial r}{\partial \hat{\theta}}\right)$, Whose (i,j)-th element is $\frac{\partial r_i(\theta)}{\partial \hat{\theta}_j}$ is of full rank (i.e. rank r). Thus $r(\hat{\theta})$ maps a small r-dimensional neighborhood of $r(\theta)$ in M.

(v) The inverse sampling r^{-1}: $M \longrightarrow \odot$ is continuous at $r(\theta) = r$. In particular, for every $\epsilon > 0$ there exists a $\delta > 0$ such that $\|\hat{\theta} - \theta\| \geq \varepsilon$, then $\| r(\hat{\theta}) - r(\theta) \| \geq \delta$.

(vi) The mapping $r: \odot \longrightarrow f_T$ is continuous at every point θ in $\odot$.

Under the regularity conditions given above and assuming $\hat{r} = r(\hat{\theta})$

where $\hat{\theta}$ is estimated by MLE

$$\hat{\theta} = \theta + (A'A)^{-1} A' D_{\prod}^{-1/2} (\hat{p} - r) \qquad - (4.9)$$

[Section 14.8.1.BFH(1975)]

where A = T x r matrix whose (i,j)-th element is

$$r_i^{-1/2} (\partial r_i / \partial \theta_j)$$

and $D_{\prod}$ = diag (r)

from regularity condition (iii) we get

$$\hat{r} - r = D_{\prod}^{1/2} A(\hat{\theta} - \theta) \qquad - (4.10)$$

Now consider the log-linear model

$$\mu = \tilde{\mu}(\theta)\ 1 + X\theta$$

With the help of above model we get

$$A = D_{\prod}^{-1/2} P X$$

where $P = D_{\prod} - r\, r'$

Now

$$A'A = X'P'D_{\prod}^{-1/2} D_{\prod}^{-1/2} P X = X'P X \quad [\because P D_{\prod}^{-1} P = P]$$

$$A' D_{\prod}^{-1/2} (\hat{p} - r) = X'P D_{\prod}^{-1/2} D_{\prod}^{-1/2} (\hat{p} - r) = X'(\hat{P} - r)$$

Since the asymptotic covariance matrix of $\hat{p}$ is $D(\hat{p}) = V/n$. From (4.9) we have the asymptotic covariance matrix of $(\hat{p})$ as

$$D(\theta) = E (\hat{\theta} - \theta) (\hat{\theta} - \theta)'$$

$$= \left[(X'P X)^{-1} X' (\hat{p} - r)(\hat{p} - r)' X (X'PX)^{-1} \right]$$

$$= (X'P X)^{-1} X' E\left[(\hat{p} - r) (\hat{p} - r)' \right] X (X'P X)^{-1}$$

$$\hat{D}(\theta) = \frac{1}{n} (X'P X)^{-1} X'V X (X'P X)^{-1} \qquad - (4.10)$$

$[\because D(\hat{p}) = V/n]$

Hence, the covariance matrix of $\hat{r}$ is given from(4.5) as

$$D(\hat{r}) = E\left[(\hat{r} - r)(\hat{r} - r)'\right] = E\left[PX(\hat{\theta} - \theta)(\hat{\theta} - \theta)'X'P'\right]$$

$$= P\,X\,E(\hat{\theta} - \theta)(\hat{\theta} - \theta)'X'P'$$

So,

$$D(\hat{r}) = P\,X\,D(\theta)X'P \qquad - (4.11)$$

In the case of multinomial sampling we have

$$D(\hat{\theta}) = \frac{1}{n}\,(X'P\,X)^{-1}(X'P\,X)(X'P\,X)^{-1} = \frac{1}{n}(X'P\,X)^{-1}$$

$(\because V = P)$

As we have from (4.4) and (4.5)

$$\hat{\theta} \sim \theta + (X'P\,X)^{-1}X'(\hat{p} - r) \qquad - (4.12)$$

$$\hat{r} \sim r + P\,X(\hat{\theta} - \theta) \qquad - (4.13)$$

From (4.12) and (4.13) we get

$$\hat{r} - r = P\,X\,(X'P)^{-1}X'(\hat{p} - r)$$

Now for finding covariance matrix of residuals $\hat{p} - \hat{r}$ we proceed as follows:

$$\hat{p} - \hat{r} = (\hat{p} - r) - (\hat{r} - r)$$

$$= \left[(\hat{p} - r) - PX(X'P\,X)^{-1}X'(\hat{p} - r)\right]$$

$$= \left[I - P\,X(X'PX)^{-1}X'\right](\hat{p} - r)$$

Dispersion of $(\hat{p} - \hat{r})$ is given by

$$D(\hat{p} - \hat{r}) = E\left[(\hat{p} - \hat{r})(\hat{p} - \hat{r})'\right]$$

$$= E\left[I - P X (X' P X)^{-1} X'\right](\hat{p} - r)(\hat{p} - r)' \left[I - P X (X' P X)^{-1} X'\right]'$$

$$= \left[I - P X (X' P X)^{-1} X'\right] E(\hat{p} - r)(\hat{p} - r)' \left[I - P X (X' P X)^{-1} X'\right]$$

$$= \left[I - P X (X' P X)^{-1} X'\right] D(\hat{p}) \left[I - P X (X' P X)^{-1} X'\right]$$

$$= \frac{1}{n} \left[I - P X (X' P X)^{-1} X'\right] V \left[I - X' (X' P X)^{-1} X' P\right]$$

$$\left[\therefore D(\hat{p}) = \frac{1}{n} V\right] \qquad - (4.14)$$

In the case of multinomial sampling it reduces to follwing form as

$V = P$

$$D(\hat{p} - \hat{r}) = \frac{1}{n}\left[I - P X(X' P X)^{-1} X'\right] P \left[I - X(X' P X)^{-1} X' P\right]$$

$$= \frac{1}{n}\left[P - PX(X' P X)^{-1} X' P\right]\left[I - X(X' PX)^{-1} X' P\right]$$

$$= \frac{1}{n}\left[P - PX(X' PX)^{-1} X' P\right] - \frac{1}{n}\left[PX(X' PX)^{-1} X' P\right]$$

$$+ \frac{1}{n}\left[PX(X' PX)^{-1} X' PX(X' PX)^{-1} X' P\right]$$

$$= \frac{1}{n}\left[P - PX(X' PX)^{-1} X' P\right] - \frac{1}{n}\left[PX(X' PX)^{-1} X' P\right]$$

$$+ \frac{1}{n}\left[PX(X' PX)^{-1} X' P\right]$$

$$= \frac{1}{n}\left[P - PX(X' PX)^{-1} X' P\right] \qquad - (4.15)$$

The diagonal elements of (4.14) provide the asymptotic variance of the residuals which are useful in detecting model deviations.

4.5 NESTED MODELS :

These nested models are discussed earlier in the case of multinomial sampling. Here an attempt will be made to analyse the sample survey data problem with the help of these nested

models. Let $X = (X_1, X_2)$, accordingly $\theta = (\theta_1, \theta_2)$ 'where X_1 is T x s and X_2 = T x u correspondingly θ_1 is s x 1 and θ_2 is u x 1 where s + u = r, $X_1'1=0$, and $X_2'1 = 0$. As stated previously we are interested in the null hypothesis H: $\theta_2 = 0$ so that under H we get reduced model M_2, where as first model (full model) is denoted by M_1. Let $\hat{\theta}_1$ and $\hat{r} = r(\hat{\theta}_1)$ denote the "psuedo MLE" of θ_1 and r respectively, under M_2 obtained from the likelihood equation $X_1'r(\hat{\theta}_1) = X_1'\hat{p}$.The consistency of $\hat{p}$ ensures that of r under $M_2(\sum \hat{r}_t = 1)$.The Pearson Chi-square statistic is given by

$$X^2 = n \sum (\hat{r}_t - \hat{\hat{r}}_t)^2 / \hat{\hat{r}}_t = n(\hat{r}-\hat{\hat{r}})' D_{\hat{\pi}}^{-1} (\hat{r}-\hat{\hat{r}}) \qquad - (4.16)$$

and the likelihood ratio statistic is as given below

$$G^2 = 2n \sum \hat{r}_t \log_e (\hat{r}_t / \hat{\hat{r}}_t) = 2n \sum \hat{p}_t \log_e (\hat{r}_t / \hat{\hat{r}}_t)$$

$$= 2n \sum \hat{p}_t \log_e (\hat{p}_t / \hat{\hat{r}}_t) - 2n \sum \hat{p}_t \log_e (\hat{p}_t / \hat{r}_t)$$

$$= G_2^2 - G_1^2 \quad \text{(say)} \qquad (4.17)$$

BFH (1975) proved in Lemma 14.9.1 that for multinomial and product multinomial sampling X^2(or G^2) IS asymptotically distributed as χ_u^2 . But it is clear that same situation will not present if our data is taken from survey sampling with some complex survey design due to clustering and stratification.

4.6 EFFECT OF SURVEY DESIGN ON NESTED MODELS :

As discussed above the clustering and stratification affect the nested model considerably so here we will try to find the asymptotic null distribution of X^2 or G^2 for any sampling design $p(\tilde{s})$. We have equation (4.12) as

$$\hat{\theta} - \theta \tilde{=} (X'PX)^{-1} X'(\hat{p}-r)$$

and equation from (4.13) as

$$\hat{r} - r \tilde{=} PX(\hat{\theta} - \theta)$$

An analogous result is given as

$$\hat{\hat{r}} - r \tilde{=} PX_1(\hat{\hat{\theta}}_1 - \theta_1) \qquad \text{[under Ho]}$$

so

$$\hat{r} - \hat{\hat{r}} = (\hat{r} - r) - (\hat{\hat{r}} - r)$$

$$= PX(\hat{\theta} - \theta) - PX_1(\hat{\hat{\theta}}_1 - \theta_1)$$

$$= P\left[(X_1 \; X_2)\begin{pmatrix}\hat{\theta}_1 - \theta_1 \\ \hat{\theta}_2 - \theta_2\end{pmatrix} - X_1(\hat{\hat{\theta}}_1 - \theta_1)\right]$$

$$= P\left[X_1(\hat{\theta}_1 - \theta_1) + X_2(\hat{\theta}_2 - \theta_2) - X_1(\hat{\hat{\theta}}_1 - \theta_1)\right]$$

$$= P\left[X_1(\hat{\theta}_1 - \theta_1) + X_2\hat{\theta}_2 - X_1(\hat{\hat{\theta}}_1 - \theta_1)\right]$$

[under Ho: $\theta_2 = 0$]

We can have analogous result from the equation (4.12) as

$$\hat{\hat{\theta}}_1 - \theta_1 = (X_1'P\,X_1)^{-1}X_1'(\hat{p} - r) \qquad - (4.18)$$

Now expressing $X'PX$ as partition matrix we have

$$X'PX = \begin{bmatrix} X_1' \\ X_2' \end{bmatrix} P[X_1 X_2] = \begin{bmatrix} X_1'PX_1 & X_1'PX_2 \\ X_2'PX_1 & X_2'PX_2 \end{bmatrix}$$

Using the standard formula for inverse and with the help of (4.12) and (4.18) we get

$$\hat{\hat{\theta}}_1 - \theta_1 = (\hat{\theta}_1 - \theta_1) + (X_1'P X_1)^{-1}(X'PX_2)\hat{\theta}_2$$

so

$$\hat{r} - \hat{\hat{r}} = P \left[X_1(\hat{\theta}_1 - \theta_1) + X_2\hat{\theta}_2 - X\{(\hat{\theta}_1 - \theta_1) + (X_1'P X_1)^{-1}(X_1'P X_2)\hat{\theta}_2\} \right]$$

$$= P \left[X_2\hat{\theta}_2 - X_1(X_1'P X_1)^{-1}(X_1'P X_2)\hat{\theta}_2 \right]$$

$$= P \left[X_2 - X_1 (X_1'P X_1)^{-1}(X_1'P X_2) \right] \hat{\theta}_2$$

$$= P \tilde{X}_2 \hat{\theta}_2$$

where $\tilde{X}_2 = \left[I - X_1(X_1'P X_1)^{-1} X_1'P \right] X_2$

$\tilde{X}_2$ is also known as projection of X_2 on the orthogonal complement of the space spanned by the columns of X_1, where the inner product is defined with respect to P ($\tilde{X}_2 PX_1 = 0$).

Now as we have

$$X^2 = n(\hat{r} - \hat{\hat{r}})' D_{\Pi}^{-1} (\hat{r} - \hat{\hat{r}})$$

$$= n \hat{\theta}_2' \tilde{X}_2' P D_{\Pi}^{-1} P \tilde{X}_2 \hat{\theta}_2$$

$$= n \hat{\theta}_2' (\tilde{X}_2' P \tilde{X}_2) \theta_2 \quad [\text{Note that } P D_{\Pi}^{-1/2} P = P] \quad -(4.20)$$

Now considering (4.10)andthe formula for the inverseof the partitioned matrix we get

$$D(\hat{\theta}_2) = \frac{1}{n} (X_2' P X_2)^{-1} (X_2' V \tilde{X}_2) (\tilde{X}_2' P \tilde{X}_2)^{-1} \quad - (4.21)$$

Hence, we can say that $\hat{\theta}_2 \tilde{=} N_u [0, D(\hat{\theta})]$ and using the standard result of the distribution of Quadratic equations we get that $X^2 \tilde{=} \sum \delta_i W_i^2$where $W_i^2 \sim \chi_1^2$,independentχ_1^2variable and δ_i are the eigenvalues of $D(\hat{\theta}_2)$ $(\tilde{X}_2' P \tilde{X}_2)n$ which is equal to $(\tilde{X}_2' P \tilde{X}_2)^{-1}(\tilde{X}_2' V \tilde{X}_2)(\tilde{X}_2' P \tilde{X}_2)^{-1}(\tilde{X}_2' P \tilde{X}_2)$ or $(\tilde{X}_2' P \tilde{X}_2)^{-1}(\tilde{X}_2' V \tilde{X}_2)$. In case of multinomial sampling we have V=P and the above formula reduces to I. Hence, in case of multinomial sampling $\delta_i = 1$, i=1, ---, u and we get $X^2 = \chi_u^2$ under Ho.

It can be noted that $\tilde{X}_2' \hat{p}$ is a vector of contrasts in the $\hat{p}_i$'s since, $\hat{X}_2' 1 = 0$ because $X_1' 1 = 0$ and $X_2' 1 = 0$. The covariance matrix of $\tilde{X}_2 \hat{p}$ is $(\tilde{X}_2' V \tilde{X}_2)/n$ for the survey design used, while $(\tilde{X}_2' P \tilde{X}_2)/n$ is the corresponding covariance matrix for multinomial sampling. Thus $\delta_1, ---, \delta_u$ are the generalized design effects as discussed in the case of two way contingency table. For the contrast vector $\tilde{X}_2' \hat{p}$ the largest eigen value, δ_1 say, is the largest possible deff taken over all linear combinations of the elements of the vector $\hat{X}_2' \hat{p}$.

4.7 WALD STATISTICS :

Let us consider a consistent estimator of V/n as $\hat{V}/n$, the covariance matrix of $\hat{p}$. Again, let C is any T x u matrix of rank such that $C' X_1 = 0$ and $C' 1 = 0$. Also $C' X_2$ is a non-singular matrix, in particular if $X_1' X_2 = 0$ a convenient choice of C would be X_2. Then our null hypothesis $\theta_2 = 0$ is equivalent to $H' : \emptyset = C'\mu = C' X_2 \theta_2 = 0$. Under Ho: $\theta_2 = 0$ this reduces further as $C'\mu = 0$. Hence the proposed wald statistics for testing H is based upon $\hat{r}$, is given by

$$X^2_{\omega} = \hat{\emptyset}' [\hat{D}\ (\hat{\emptyset})]^{-1}\ \hat{\emptyset} \qquad - (4.22)$$

where $\hat{\emptyset} = C'\hat{\mu}$ and $\hat{D}\ (\emptyset)$ is the estimated asymptotic covariance matrix of $\hat{\emptyset}$. Also, $\hat{\mu}$ is the log-probability vector of T x 1. Now since we have $\sqrt{n}\ (\hat{r} - r) \longrightarrow N\ [0, D\ (\hat{r})]$

and $\log \hat{r} \tilde{=} \log r + \left[\frac{\partial \log r}{\partial r} \right] (\hat{r} - r) + 0\ (\| r - r \|)$

which can be written as

$$\hat{\mu} - \mu \tilde{=} D_{\Pi}\ (\hat{r} - r)$$

so

$$D(\hat{\emptyset}) = E \left[C'\ (\hat{\mu} - \mu)\ (\hat{\mu} - \mu)'\ C \right]$$

$$= E \left[C' D_{\Pi}^{-1}\ (\hat{r} - r)\ (\hat{r} - r)' D_{\Pi}^{-1}\ C \right]$$

$$= C'\ D_{\Pi}^{-1}\ D\ (\hat{r})\ D_{\Pi}^{-1}\ C = \Sigma_{\emptyset}\ (\text{say}) \qquad - (4.23)$$

where $D(\hat{r})$ is given by (4.11). Now by replacing r by $\hat{r}$ and V by

$\hat{V}$ we get the estimator of D $(\hat{\emptyset})$.The Wald statistics given by equation (4.22) is independent of the choice of C matrix and if no estimator of V/n is available we simply replace it by multinomial covariance matrix in $D(\theta)$.

$$D(\hat{\emptyset}) = \frac{1}{n} C' D_{\Pi}^{-1} PX (X'PX)^{-1} X' P D_{\Pi}^{-1} C$$

$$= \frac{1}{n} C' X (X'PX)^{-1} X' C = \Sigma_o$$

$[\because P D_{\Pi}^{-1} C = C]$

With the help of the above equation we get the test statistic alternative to X^2 or G^2.

$$\tilde{X}^2_{\omega} = n \hat{\emptyset}' [C'X (X'PX)^{-1} X'C]^{-1} \hat{\emptyset}. \qquad - (4.24)$$

The true asymptotic distribution of $\tilde{X}^2$ under Ho is a weighted sum of independent χ^2_1 random variables

$$\tilde{X}^2_{\omega} \tilde{=} \sum_{i=1}^{u} r_i W_i^2$$

where $r_1, \ldots, r_u$ are the eigen values of $\Sigma_o^{-1} \Sigma_{\emptyset}$. Now we will try to show that $\tilde{X}^2_{\omega}$ is equivalent to X^2 under the null hypothesis, which implies that $X^2_{\omega} = \sum \delta_i W_i^2$ and $[\delta_1, \ldots, \delta_u]$ is identical to $[r_1, r_2, \ldots, r_u]$. Under the null hypothesis $H : \theta_2 = 0$ We have $\hat{\emptyset} = \hat{\emptyset} - \emptyset = C' (\hat{\mu} - \mu) \sim C' D_{\Pi}^{-1/2} (\hat{r} - r)$

$$\sim C' D_{\Pi}^{-1} P X (\hat{\theta} - \theta) \qquad [\text{using (4.13)}]$$

Also note that

$$C' D_{\Pi}^{-1} PX = C'X = (0, C'X_2)$$

$[\because C' D_{\Pi}^{-1} P = C']$

Hence,

$$\hat{\phi} = C' X (X' P X)^{-1} X' (\hat{p} - r)$$

Again using the formula for the inverse of the partioned matrix we get

$$\hat{\phi} = C' X_2 \hat{\theta}_2$$

and $C' X (X' P X)^{-1} X' C = C' X_2 (\tilde{X}_2' P \tilde{X}_2)^{-1} X_2' C$

Hence, with the help of (4.24) we can write as

$$\tilde{\chi}^2_\omega = n \hat{\theta}_2' X_2' C \left[C' X_2 (\tilde{X}' P X_2) X_2' C \right]^{-1} C' X_2 \hat{\theta}_2$$

$$= n \hat{\theta}_2' (\tilde{X}_2' P \tilde{X}_2) \hat{\theta}_2 \sim \chi^2 \qquad - (4.25)$$

[$\because C' X_2$ is non-singular]

Now another multinomial Wald statistics which is possible is based on a weighted least square estimator (WLS), $\tilde{\theta}$, of θ and denoted by $\tilde{\chi}^2_\omega(1)$. We will try to show the equivalence of χ^2 and $\tilde{\chi}^2_\omega(1)$.

Let F be any matrix of order (T-1) x T of rank (T-1) such that F1 = 0. Let $\tilde{f} = F \tilde{\mu}$, where $\tilde{\mu}$ is a vector of log-probabilities of order T x 1 where $\tilde{\mu}_t = \log \hat{p}$. Now assume the cental limit theorem we have

$$\sqrt{n} (\tilde{f} - f) = N(0, V_f)$$

where $f = F\mu = F X \theta$,

$$V_f = F D_\Pi^{-1} V D_\Pi^{-1} F' \qquad [\text{By.}\delta\text{-method}]$$

If we assume the multinomial sampling, we have

$$V_f = F D_\Pi^{-1} P D_\Pi^{-1} F' = F D_\Pi^{-1} F$$

The weighted least square estimator of θ under the model $f = F X \theta$ is given by

$$\tilde{\theta} = (X' F \tilde{V}_f^{-1} F X)^{-1} X' F \tilde{V}^{-1} F \tilde{\mu} \qquad --- \quad (4.26)$$

where $\tilde{V} = F D_p^{-1} \hat{V}_f D_p^{-1} F'$

The asymptotic covariance matrix of $\tilde{\theta}$ is given by

$$D(\tilde{\theta}) = \frac{1}{n}(X'F'V_f^{-1}F\,X)^{-1} = \frac{1}{n}V_{\theta\upsilon}(\text{say})$$

Now partitioned $\tilde{\theta}'$ as $(\tilde{\theta}_1' \; \tilde{\theta}_2')$ and correspondingly V_θ as

$$V_\theta = \begin{bmatrix} V_{\theta 11} & V_{\theta 12} \\ V_{\theta 21} & V_{\theta 22} \end{bmatrix}$$

Hence, a Wald statistic which is asymptotically χ^2_u under H: $\theta_2 = 0$ is given by

$$\chi^2_\omega(1) = n\,\tilde{\theta}_2'\,\tilde{V}^{-1}_{\theta 22}\,\tilde{\theta}_2 \qquad -(4..27)$$

provided $\hat{V}$ is available. Here, one thing we can note that $\tilde{\theta}$ is independent of the choice of F. To prove this let us consider G which is a non-singular matrix, let

$$f^* = F^*\mu$$

$$F^* = G\,F$$

so $$V_f^* = G\,V_f\,G'$$

Hence,
$$F'^*\,V_{f^*}^{-1}\,F^* = F'G'(G\,V_f\,G')^{-1}G\,F$$
$$= F'G'G'^{-1}V_f^{-1}\,G^{-1}G\,F$$
$$= F'\,V_f^{-1}F.$$

so from the above derivation it is clear that $\tilde{\theta}$ is independent of the choice of F. Now, let us consider a particular choice of F as

$$F = \begin{vmatrix} \bar{} & X'P^{-} \\ \_ & F_1 \_ \end{vmatrix}$$

Where F_1 is a $(T-r-1) \times T$ matrix of rank $(T-r-1)$ with the properties $F_1 X = 0$, and $F_1 1 = 0$ with this choice we have

$$F\,X = \begin{vmatrix} \bar{} & X'PX^{-} \\ \_ & 0 \_ \end{vmatrix}$$

But as we know that $V_f = F\,D_\Pi^{-1}V\,D_\Pi^{-1}F'$, with this particular

value given by F we have

$$V_f = \begin{bmatrix} X'P \\ F_1 \end{bmatrix} D_\Pi^{-1} V D_\Pi^{-1} \; (P \; XF_1')$$

$$= \begin{bmatrix} X'P\,D_\Pi^{-1}V\,D_\Pi^{-1} \\ \\ F\,D_\Pi^{-1}V\,D_\Pi^{-1} \end{bmatrix} \quad (PX \;\; F_1')$$

$$= \begin{bmatrix} X'V\;D_\Pi^{-1}P\;X & X'V\,D_\Pi^{-1}F_1' \\ \\ F_1 D_\Pi^{-1}V\;D_\Pi^{-1}P\;X & F\,D_\Pi^{-1}V\,D_\Pi^{-1}F_1' \end{bmatrix}$$

$$= \begin{bmatrix} X'V\;X & X'V\,D_\Pi^{-1}F_1' \\ \\ F_1 D_\Pi^{-1}VXF_1 D_\Pi^{-1}V\,D_\Pi^{-1}F_1' \end{bmatrix}$$

Now putting V = P for the case of multinomial sampling we have

$$V_f = \begin{bmatrix} X'P\;X & X'P\,D_\Pi^{-1}F_1' \\ F_1 D_\Pi^{-1}PXF\;D_\Pi^{-1}P\,D_\Pi^{-1}F_1' \end{bmatrix}$$

$$= \begin{bmatrix} X'PXX'F_1' \\ \\ F_1X \;\; F_1 D_\Pi^{-1}F_1' \end{bmatrix}$$

$$= \begin{bmatrix} X'PX & 0 \\ \\ 0 & F_1 D_\Pi^{-1}F_1' \end{bmatrix} \qquad -(4.28)$$

Thus by replacing $\hat{V}$ by $\tilde{P} = D_{\tilde{p}}^{-1} - \hat{p}\hat{p}'$ in (4.26) and (4.27) we get the statistics when $\hat{V}$ is not available and it depends upon the assumption of multinomial sampling. Consider the equation (4.26) and replace $\hat{V}$ by $\tilde{P}$ we get

$$\tilde{\theta} = \left[X'F'(F D_{\tilde{p}}^{-1} \tilde{P} D_{\tilde{p}}^{-1} F')^{-1} FX \right]^{-1} X'F'(F D_{\tilde{p}}^{-1} \tilde{P} D_{\tilde{p}}^{-1} F')^{-1} F\tilde{\mu}$$

$$= \left[X'F'(F D_{\tilde{p}}^{-1} F')^{-1} FX \right]^{-1} X'F'(F D_{\tilde{p}}^{-1} F')^{-1} F\tilde{\mu}$$

$$= \left[(X'\tilde{P}X \quad 0)(F D_{\tilde{p}}^{-1} F')^{-1} \begin{pmatrix} X'\tilde{P}X \\ 0 \end{pmatrix} \right]^{-1} (X'\tilde{P}X \quad 0) \times (F D_{\tilde{p}}^{-1} F')^{-1} \begin{pmatrix} X'\tilde{P} \\ 0 \end{pmatrix} \tilde{\mu}$$

$$= \left[X'\tilde{P}X(F D_{\tilde{p}}^{-1} F')^{-1}(X'\tilde{P}X) \right]^{-1} (X'\tilde{P}X)(F D_{\tilde{p}}^{-1} F')^{-1} X'\tilde{P}\tilde{\mu}$$

$$= (X'\tilde{P}X)^{-1}(F D_{\tilde{p}}^{-1} F')(X'\tilde{P}X)^{-1}(X'\tilde{P}X)(F D_{\tilde{p}}^{-1} F')^{-1} X'\tilde{P}\tilde{\mu}$$

$$= (X'\tilde{P}X)^{-1} X'\tilde{P}\tilde{\mu} \qquad - (4.29)$$

Hence,

$$\tilde{\theta} - \theta = (X'\tilde{P}X)^{-1} X'\tilde{P}(\tilde{\mu} - \mu)$$

Now using $\tilde{\mu} - \mu \simeq D_{\Pi}^{-1}(\hat{r} - r)$ we have

$$\tilde{\theta} - \theta = (X'\tilde{P}X)^{-1} X'\tilde{P} D_{\Pi}^{-1}(\hat{r} - r)$$

From the equation (4.13) we have

$$\tilde{\theta} - \theta = (X'\tilde{P}X)^{-1} X'\tilde{P} D_{\Pi}^{-1} \tilde{P} X(\hat{\theta} - \theta)$$

$$= (X' \tilde{P} X)\ (X' \tilde{P} X)\ (\hat{\theta} - \theta)$$

$$= (\hat{\theta} - \theta) \qquad - (4.30)$$

Moreover, V_θ reduces to $(X' P X)^{-1}$ so that $V_{\theta 22}\ (\tilde{X}_2'\ P\ \tilde{X}_2)^{-1}$. Thus under Ho: $\theta_2 = 0$ we have

$$\tilde{X}^2_{\omega}\ (1) = n\tilde{\theta}_2' (\tilde{X}_2' \tilde{P} \tilde{X}_2)^{-1} \tilde{\theta}_2$$

$$\sim\ n\ \hat{\theta}_2'\ (\tilde{X}_2' P\ \tilde{X}_2)^{-1}\ \tilde{\theta}_2 \sim \chi^2.$$

[By (4.25)]

In the special case where M_1 is a saturated model we have s + u = T-1. Also, in the case of saturated models we have $\hat{r} = \hat{p}$ and $D(\hat{r}) = V/n$ under the given survey design $p(\tilde{s})$. For the case of multinomial sampling $D(\hat{r}) = P/n$. Noting that $C' D_\Pi^{-1} P\ D_\Pi^{-1} C = C' D_\Pi^{-1} C$. We get the result that δi's are eigen values of

$\Sigma_0^{-1}\Sigma_\emptyset = (C' D_\Pi^{-1} C)^{-1} (C' D_\Pi^{-1} V D_\Pi^{-1} C)$. In this waywe can observe that we can obtain δi's without calculating the projection matrix $\tilde{X}_2$.

Consider the approximation given by Satterth⌒Waite (1946). Under the null hypothesis the distribution of X^2 is

$$X^2_s = \frac{X^2}{(1+a^2)\delta.} \sim \chi^2_v\ ,\ v = \frac{u}{1+a^2}$$

Where $\delta. = \frac{1}{u} \sum_i \delta i$ and $a^2 = \frac{\sum (\delta i - \delta.^2)}{[u\ \delta^2\ ]}$ which is nothing but square of the coefficient of variation of the δi,s. But we can prove that it is not necessary to evaluate all δi. We can

see that $\bar{E}(X^2) = \sum_i \delta i$ and $\bar{V}(X^2) = 2\sum \delta i^2$, where $\bar{E}$ and $\bar{V}$ denote the asymptotic expectation and asymptotic variance operators.

Consider again

$$X^2 = \sum_t \frac{n(\hat{r}_t - \hat{\hat{r}}_t)^2}{r_t}$$

$$\sqrt{n}\ (\hat{r} - \hat{\hat{r}}) \simeq y \sim N_T(0, B)$$

where $B = (b_{tl}) = P \tilde{X}_2 D(\hat{\theta}_2) \tilde{X}_2' P$ and $D(\hat{\theta}_2)$ is given by (4.21)

Hence, $E(X^2) = \sum_t b_{tt} / r_t$

similarly $V(Y_t^2) = 2b_{tt}^2$

and $Cov(Y_t^2, Y_l^2) = 2b_{tl}^2$, $t \neq l = 1, ---, T$

So variance is given by

$$V(X^2) = 2 \sum_t \sum_l b_{tl}^2 / (r_t\ r_l)$$

If the estimated full covariance matrix $\hat{V}/n$ of cell estimates available, then x_s^2 is employed as an alternative to wald statistics by estimating $\hat{a}$ and $\hat{\delta}$ obtained from $\hat{B}$ in place of 'a' and 'δ'. The asymptotic sygnificance level of χ^2 for desired nominal level of α can be found and consequently the design effect may be obtained as

$$SL(X^2) = P\left[x^2 \geq \chi^2_{u(\alpha)}\right] = P\left[\chi^2_\nu \geq \{1+a^2\ \delta.\}^{-1}\ \chi^2_{u(\alpha)}\right]$$

is computed to α , where χ^2_μ is the upper α - point of χ^2_μ.

4.8 MODIFICATIONS TO X^2:

The first order correction to X^2(or G^2) can improve the X^2 statistics adequately i.e. we can treat $X^2/\hat{\delta}.$ or $G^2/\hat{\delta}.$ as χ^2_u under the null hypothesis H_1 where $\delta.$ may be written as

$$u\delta_{\cdot} = \text{Trace}\left[(\tilde{X}_2' P \tilde{X}_2)^{-1} (\tilde{X}_2' V \tilde{X}_2) \right]$$

$$= \text{Trace}\left[(X' P X)^{-1} (X' V X) \right]$$

$$- \text{Trace}\left[(X_1' P X_1)^{-1} (X_1' V X_1) \right]$$

$$= (s + u)\lambda_{\cdot} - s\lambda_{1\cdot} \text{ (say)} \qquad -(4.31$$

In the case of saturated model M_1 the above equation can be written as

$$(T - s - 1)\,\delta_{\cdot} = (T-1)\lambda_{\cdot} - s\,\lambda_{1\cdot}$$

[$\because s+u = T-1$.]

where $(T-1)\,\lambda_{\cdot} = \sum_t (1-r_t)d_t$

where $d_t = \dfrac{\nu^t_{tt}}{r_t(1-r_t)}$ = cell design effect of $\hat{p}_t$

$V = (\nu_{tt})$. For $T \gg s$, $\delta_{\cdot} = \lambda_{\cdot}$ and it is expected that $X^2/\hat{\lambda}_{\cdot}$ will also perform well in the large tables if s is fairly small. It can be noted that $\lambda_{\cdot}$ is independent of H as the first order correction proposed by Fellegi (1980) where $\hat{d}_{\cdot} = \sum \hat{d}_t / T$ where as $\hat{\delta}_{\cdot}$ is full dependent on the null hypothesis under consideration.

It can be shown that $\delta_{\cdot}$ can be computed knowing only cell deffs, d_t and the deffs of collapse tables (or marginals), whenever model admits explicit solution. For example, the no three factor interaction hypothesis is the only hierarchical hypothesis that does not permit explicit solution in a threeway table. Using the notation of BFH(1975) a hypothesis leading to

direct estimate is of the form.

$$r_\theta = [\prod_i \hat{p}_{\theta i}] / [\prod_j \hat{p}_{\emptyset j}] \qquad - (4.32)$$

where $\hat{p}_{\theta i}$ is the marginal total of $\hat{p}_\theta$ corresponding to θi. Similarly $\hat{p}_{\emptyset j}$ is the marginal total of $\hat{p}_\theta$ corresponding to $\emptyset j$. For example consider the example of three-way table,

$$\theta = (ijk),\ \theta_1 = (i\ k),\ \theta_2 = (j\ k),\ \emptyset_1 = \{k\}$$

Again, consider the hypothesis that variable 1 and 2 are conditionally independent given the level of the variable (3).

Now using the equation (4.32) we get

$$\frac{G^2}{n} = 2 \sum_\theta \hat{p}_\theta \log_e (\hat{p}_\theta / r_\theta)$$

$$= 2\sum_\theta \hat{p}_\theta \log_e \hat{p}_\theta = 2\sum_i \left[\sum_{\theta i} \hat{p}_{\theta i} \log_e \hat{p}_{\theta i}\right]$$

$$+ 2\sum_j \left[\sum_{\emptyset j} \hat{p}_{\emptyset j} \log \hat{p}_{\emptyset j}\right]$$

Now

$$2 \sum_\theta \hat{p}_\theta \log_e \hat{p}_\theta = 2\sum_\theta \hat{p}_\theta \left[\log_e r_\theta + \log_e \left\{1 + \frac{\hat{p}_\theta - r_\theta}{r_\theta}\right\}\right]$$

$$\sim 2 \sum_\theta \hat{p}_\theta \left[\log_e r_\theta + \frac{\hat{p}_\theta - r_\theta}{r_\theta} \frac{(\hat{p}_\theta - r_\theta)^2}{2 r_\theta^2}\right]$$

$$\sim 2 \sum_\theta \hat{p}_\theta \log_e r_\theta + \sum_\theta \frac{(\hat{p}_\theta - r_\theta)^2}{r_\theta} \qquad -(4.33)$$

Similarly

$$2\sum_{\theta i}\hat{p}_{\theta i}\log_e\hat{p}_{\theta i} \sim 2\sum_{\theta i}\hat{p}_{\theta i}\log_e r_{\theta i}$$

$$+\sum_{\theta i}(\hat{p}_{\theta i}-r_{\theta i})^2/r_{\theta i} \quad -(4.34)$$

and

$$2\sum_{\phi j}\hat{p}_{\phi j}\log_e\hat{p}_{\phi j} \sim 2\sum_{\phi j}\hat{p}_{\phi j}\log_e r_{\phi j}$$

$$+\sum_{\phi j}\frac{(\hat{p}_{\phi j}-r_{\phi j})^2}{r_{\phi j}} \quad -(4.35)$$

Putting the values of (4.33), (4.34) and (4.35) in above equation we get

$$\frac{G^2}{n} \sim \sum_{\theta}\frac{(\hat{p}_{\theta}-r_{\theta})^2}{r_{\theta}} - \sum_i\left[\sum_{\theta i}\frac{(\hat{p}_{\theta i}-r_{\theta i})^2}{r_{\theta i}}\right] + \sum_j\left[\sum_{\phi j}\frac{(\hat{p}_{\phi j}-r_{\phi j})^2}{r_{\phi j}}\right]$$

$$\left[\because 2\sum_{\theta}\hat{p}_{\theta}[\log_e r_{\theta}-\sum_i\log_e r_{\theta i}+\sum_j\log_e r_{\phi j}] = 0\right]$$

[from (4.32)]

$$\text{or} \quad \bar{E}\,G_1^2 = \sum_{\theta}(1-r_{\theta})\,d_{\theta} - \sum_i\left\{\sum_{\theta i}(1-r_{\theta i})d_{\theta i}+\sum_j\sum_{\phi j}(1-r_{\phi j})d_{\phi j}\right\} \quad -(4.36)$$

Where

$d_{\theta} = \dfrac{n\,v(\hat{p}_{\theta})}{r_{\theta}(1-r_{\theta})}$, which is the design effect of $\hat{p}_{\theta}$ under

the null hypothesis $d_{\theta i} = \dfrac{n\, V(\hat{p}_{\theta i})}{r_{\theta i}(1 - r_{\theta i})}$, which is the design effect of marginals $\hat{p}_{\theta i}$ Similarly

$$d_{\emptyset j} = \frac{n\, V(\hat{p}_{\emptyset j})}{r_{\emptyset j}(1-r_{\emptyset j})}$$, which is the design effect of marginals $\hat{p}$

The estimated deff $\tilde{d}_\theta$ of d_θ is given by $\hat{d}_\theta \left[\hat{p}_\theta(1-\hat{p}_\theta)/ \hat{r}_\theta (1-\hat{r}_\theta) \right]$

where $d_\theta = \hat{V}(\hat{p}_\theta) \Big/ \left[\hat{p}_\theta (1-\hat{p}_\theta)/n \right]$ is the estimated design effect

of $\hat{p}_\theta$ irrespective of H, and $\hat{V}(\hat{p}_\theta)$ is the estimate of variance of $\hat{p}_\theta$. It is common practice to report $\hat{d}_\theta$ rather than $\tilde{d}_\theta$.

Hence, we have

$$u\,\hat{\delta}. = \sum_\theta \frac{\hat{p}_\theta}{\hat{r}_\theta}(1-\hat{p}_\theta)\hat{d}_\theta - \sum_i \left\{ \sum_{\theta i} (1-\hat{p}_{\theta i})\, \hat{d}_{\theta i} \right\}$$

$$+ \sum_j \left\{ \sum_{\emptyset j} (1-\hat{p}_{\emptyset j})\, \hat{d}_{\emptyset j} \right\} \quad - (4.37)$$

It is an important result which can be used to prepare cotingency Tables from the survey data.

For example consider I x J Table and let our null hypothesis is $r_{ij} = r_{i+}\, r_{+j} \iff \mu_{12}(ij) = 0$, i=1,---,I; and j=1,---,J.

so the equation (4.37) can be written as

$$(I-1)\,(J-1)\delta. = \sum_i \sum_j (1-r_{i+}\, r_{+j})d_{ij} - \sum_i (1-r_{i+})d_i(r)$$

$$- \sum_j (1-r_{+j})d_j\ (c)$$

where r_{i+} and r_{+j} are the row and column marginals, $d_i(r)$ and $d_j(c)$ are the design effect respectively, and d_{ij} is the deff of $\hat{p}_{ij}$. Hence, $X^2/\hat{\delta}.$ can be computed knowing only the cell proportions $\hat{p}_{ij}$ and the design effects of the above mentioned quantities. Similarly for IxJxK table under the hypothesis of complete independence ie $r_{i++}.\ r_{+j+}\ r_{++k}$ we get

$$(IJK - I - J - K + 2)\delta. = \Sigma_i \Sigma_j \Sigma_k\ (1 - r_{i++}\ r_{+j+}\ r_{++k})d_{ijk} - \Sigma_i\ (1 - r_{i++})\delta_i(r) - \Sigma_j (1 - r_{+j+})d_j(c) - \sum_k (1 - r_{++k})d_k(1).$$

Where $r_{i++}\ r_{+j+}$ and r_{++k} are the three way marginals and $d_i\ (r)$, $d_j(c)$ and $d_k(1)$ are the corresponding marginal deffs d_{ijk} is the deff of $\hat{p}_{ijk}$. When the model does not permit explicit solution for $\hat{r}$ and $\hat{r}$, $\delta.$ can not be expressed in terms of only cell deffs and marginal deffs.

Again let us consider that one of the margins of IxJ table is fixed. Let n_{ij} be the observed frequency in (i,j)th cell, asn$_{i+}$ isthefixed row marginal. Take

where $\hat{p}_{ij} = \left(\frac{n_{i+}}{n}\right) \hat{p}_{j(i)}$ and $r_{ij} = \frac{n_{i+}}{n} p_{j(i)}$, where

$P_{j(i)}$ are the cell proportions within the i-th row population and $\hat{p}_{j(i)}$ are the corresponding survey estimates ($\Sigma_j\ P_{j(c)} = \Sigma_j\ \hat{p}_{j(c)} = 1$)

Hence, H: $r_{ij} = r_{i+} r_{+j}$ <=> $P_{j(c)} = \Sigma_i \frac{n_{i+}}{n}\ P_{j(i)} = P_j$ say

[$\because \Sigma_j\ P_{j(i)} = 1$]

Which is equivalentto test of homogeneity across the populations.

Since $\hat{p}_{i+} = \frac{n_{i+}}{n}$, we get $V(\hat{p}_{i+}) = 0$

Hence, $$(1-r_{i+})d_{i(r)} = 0 \qquad - (4.38)$$

Also $$(1-r_{i+} - r_{+j})\, d_{ij} = \frac{n\, V(\hat{p}_{ij})}{r_{i+}\, r_{+j}} = (1-P_j)\, d_{j(i)} \qquad - (4.39)$$

where $d_{j(c)} = \frac{n_{i+}\, V(\hat{p}_{j(i)})}{P_j\,(1-P_j)}$; which j-th cell deff in the i-th row population under the null hypothesis

Similarly

$$(1-r_{+j})\, d_{j(c)} = \frac{n\, V(\hat{p}_{+j})}{r_{+j}} = \Sigma_i \frac{n_{i+}}{n} d_{j(c)}(1-P_j)$$

Note $$\hat{p}_{+j} = \Sigma_i \frac{n_{i+}}{n} \hat{p}_{j(i)} \qquad - (4.40)$$

$$n\, V(\hat{p}_{+j}) = \Sigma_i \frac{n_{i+}}{n} d_{j(i)}\, P_j\,(1-p_j)$$

Assume that sampling is done independently with in each row population. Hence using (4.38), (4.39) and (4.40) in (4.36) we get

$$(I-1)\,(J-1)\delta_{\cdot} = \Sigma_i \Sigma_j\, (1-P_j) d_{j(i)} \left[1 - \frac{n_{i+}}{n}\right]$$

which agrees with the result of Rao and Scott(1981).

4.9 MODELS NOT ADMITTING DIRECT SOLUTION TO MULTINOMIAL LIKELIHOOD EQUATION :

Consider those models which are not admitting direct solution to multinomial likelihood equations. The $\delta_{\cdot}$ Corresponding to a nested model M^{*} closest to the original model

M can be used for modifying the statistics under consideration. Let the model M^* closest to the model M and admitting direct solution of r i.e. a model $r_t = r_t(\theta^{\natural})$ is given by

$$\mu = \tilde{\mu} \begin{pmatrix} \theta^{\natural} \\ 0 \end{pmatrix} \quad 1 + X^* \theta^*$$

where $X\theta = X^* \theta^* + X^{**} \theta^{**}$, X^* is $T \times r^*$ matrix of rank r^*, X^{**} is $T \natural (r-r^*)$ matrix of rank $(r-r^*)$ and $(r-r^*)$ is as small as possible. This method can be further improved by finding an approximate upper bound of δ.. Let $G^{*2} = 2n \sum \hat{p}_t \log_e (\hat{p}_t / \hat{r}_t^*)$ be the likelihood ratio test of goodness of fit of model M^* and r_t^* be the pseudo-MLE of r_t under M^*. We have

$$\bar{E}\, G^2 = (T-r-1)\delta.$$

$$\bar{E}^* G^{*2} = (T-r^*-1)\delta.^*$$

Where $\bar{E}$ is the asymptotic expectation taken over the model M and $\bar{E}^*$ is the asymptotic expectation taken over the model M^* which admits direct solution. Since, model M and M^* is assumed to be close we have $r_t(\theta) - r_t(\theta^*) = (1/\sqrt{n})\, a_t$ applying the method of linearization as used previously we get

$$\bar{E}\, G^{*2} = (T-r^*-1)\delta^* + \sum_{t=1}^{T} a_t^2 / r_t(\theta)$$

Hence note that $G^{*2} > G^2$ consequently we get

$$(T-r-1)\delta. < (T-r^*-1)\delta.^* + \sum_{t=1}^{T} a_t^2 / r_t(\theta) \qquad (4.41)$$

$$\Rightarrow \delta. < (T-r-1)^{-1}(T-r^*-1)\delta^*. + (T-r-1)^{-1} \sum_{t=1}^{T} a_t^2 / r_t(\theta)$$

The second term of the equation can be estimated from the

sample but it can be seen that it is very small relative to the first term and can be neglected. So we get the nearly conservative correction as given below

$$\chi^2 / \left[(T-r-1)^{-1} (T-r^*-1)\hat{\delta}_.^* \right]$$

or $$G^2 / \left[(T-r-1)^{-1} (T-r^*-1)\hat{\delta}_.^* \right] \qquad (4.42)$$

The above mentioned correction is more conservative than $\chi^2/\hat{\delta}_.^*$ or $G^2/\hat{\delta}_.^*$ because of the fact

$$(T-r-1)^{-1} (T-r^*-1)\ \delta_.^* > \delta_.^*$$

An exact upper bound to $\delta_.$ under M can also be obtained from some seperation inequalities for eigen values (Scott and Styan, 1985) we have

$$\delta_. \leq \sum_{i=1}^{T-r-1} \lambda_i / (T-r-1) \qquad (4.43)$$

where, $\lambda_1 \geq \text{-----} \geq \lambda_{T-1} > 0$ are non-zero eigen values of $D_\Pi^{-1}V$. If λ_i's are not available, a simple upper bound of $\delta_.$ depending upon the all deffs d_i can be obtained from (4.43).

$$(T-r-1)\delta \leq \sum_{i=1}^{T-r-1} \lambda_i \leq \sum_{i=1}^{T-1} \lambda i = (T-1)\ \lambda_. = Tr\ D^{-1}V$$

$$= \sum_{i=1}^{T} (1-r_i)d_i \qquad (4.44)$$

where $d_i = \dfrac{nVar\ (\hat{p}_i)}{r_i\ (1 - r_i)}$ is the design effect of t-th cell.

However, the upper bound $(T-1)\lambda / (T-r-1)$ on $\delta_.$ is not likely to be good unless r is small relative to T.

4.9.1 NESTED HYPOTHESIS :

Let us consider the hypothesis $H_{2.1} : \theta_2 = o$ given the following model M, this is also known as nested hypothesis

$$\mu = \tilde{\mu}(\theta)\ 1 + X_1\theta_1 + X_2\theta_2$$

As we know that Pearson statistic is given by

$$X^2\ (2/1) = n \sum_t (\hat{r}_t - \hat{\hat{r}}_t)^2 / \hat{\hat{r}}_t$$

where $\hat{\hat{r}}_t$ is the pseudo MLE under $H_{2.1}$ and $\hat{r}_t$ is pseudo MLE for the full model given above. Similarly, the likelihood - ratio statistics is given by

$$G^2\ (2/1) = 2n \sum \hat{p}_t \log_e (\hat{r}_t / \hat{\hat{r}}_t)$$

$$= G^2\ (2) - G^2(1) = 2n \sum \hat{r}_t \log \left(\frac{\hat{r}_t}{\hat{\hat{r}}_t} \right)$$

Here $$G^2(1) = 2n \sum \hat{p}_t \log_e \left(\frac{\hat{p}_t}{\hat{r}_t} \right)$$

$$G^2(2) = 2n \sum \hat{p}_t \log_e (\hat{p}_t / \hat{\hat{r}}_t)$$

Again it can be shown that

$$G^2(1) \sim \sum_t \frac{(\hat{p}_t - \hat{r}_t)^2}{\hat{r}_t}$$

From the equation (4.14) we can see that

$I - PX\ (X'P\ X)^{-1}\ X'$ is an idempotent matrix

so $$D\ (\hat{p} - r) = \frac{1}{n} A' D_{\Pi}^{-1} A$$

Where $A = I - PX\,(X'PX)^{-1}X'$

Hence $G^2(1) = n(\hat{p} - r)' A' D_{\Pi}^{-1} A\,(\hat{p} - r)$ (4.45)

Let $X = (X_1, X_2)$, $\theta = (\theta_1\ \theta_2)'$

and $(X'PX)^{-1}X'(\hat{p} - r) = \begin{bmatrix} E_1 \\ E_2 \end{bmatrix} X'(\hat{p} - r)$ (say)

Then from the equation (4.12) we have

$$\begin{bmatrix} \hat{\theta}_1 - \theta_1 \\ \hat{\theta}_2 - \theta_2 \end{bmatrix} = \begin{bmatrix} E_1X' & (\hat{p} - r) \\ E_2X' & (\hat{p} - r) \end{bmatrix}$$

Under $H_{2.1}$: $\theta_2 = 0$ we have

$$\hat{\theta}_2 = E_2X'(\hat{p} - r)$$

But as we know that

$$G^2\ (2/1) \sim \sum_t \frac{(\hat{r}_t - \hat{\hat{r}}_t)^2}{\hat{\hat{r}}_t}$$

Now again with the help of equation (4.20)

We can write

$$G^2(2/1) \sim X^2 = n\,\hat{\theta}_2'(\tilde{X}_2'P\,\tilde{X}_2)\,\hat{\theta}_2$$

so $G^2(2/1) = n\,(\hat{p} - r)'X\,E_2'\,(\tilde{X}_2'P\,\tilde{X}_2)\,E_2X'\,(\hat{p} - r)$ - (4.46)

Since, $\sqrt{n}\,(\hat{p} - r) \longrightarrow N(0, V)$ as $n \longrightarrow \infty$

It follows from (4.45) and (4.46) that

$$G^2\ (1) \sim \sum_{i=1}^{T-r-1} \eta_i\, W_i^2, \text{ where } W_i^2 \sim \chi_1^2$$

$$G^2\ (2/1) \sim \sum^{u} \gamma_i\, W_i^{*2}, \text{ where } W_i^{*2} \sim \chi_1^2$$

$i=1$

$\eta_i > 0$ and $\gamma_1 \geq \gamma_2 \geq \text{----}, \geq \gamma_u > 0$

A first order correction to them is given as

$$\frac{G^2(2/1)}{\hat{\gamma}.} \sim \frac{\chi^2(2/1)}{\hat{\gamma}.} \sim \chi^2_u$$

where $\gamma. = \sum_{i=1}^{u} \gamma_i / u$ under $H_{2.1}$

As we know that

$$u\gamma. = \bar{E}\, G^2(2/1) = \bar{E}\, G^2(2) - \bar{E}\, G^2(1)$$

So $\gamma.$ the correction factor computed or its upper bound is found depending upon the fact whether models under consideration i.e. M and Mi admit direct solution or not. The upper bound can be found exactly the same way as discussed previously.

So an upper bound on $\gamma.$ is given by

$$\gamma. \leq \sum_{i=1}^{u} \lambda_i / u$$

where λ_i,s are the eigen values of $D_{\Pi}^{-1} V$. If λ_i,s are not available, a simpler upper bound depending only on the cell deff, d_i. can be obtained as given below

$$u\,\gamma. \leq \sum_{i=1}^{u} \lambda_i \leq (T-1)\lambda. = \sum (1-r_i) d_i.$$

4.9.2 DESIGN EFFECT OF F-STATISTICS :

Consider the case when we use the F-statistic on the data obtained through some survey design. The statistic under consideration for the above discussed problem is given

$$F = \frac{G^2(2/1)/u}{G^2(1)/(T-r-1)}$$

With the help of the above statistic we want to test $H_{2.1}$ by treating this statistic F-variable with degrees of freedom μ and (T-r-1) respectively. Applying the first order correction to $G^2(1)$ and $G^2(2/1)$ we get

$$F \overset{\sim}{=} \frac{\gamma_\cdot \chi^2_u / u}{\eta_\cdot \chi^2_{T-r-1} / (T-r-1)}$$

where

$$\eta_\cdot = \sum \eta_i / (T-r-1)$$

$$\gamma_\cdot = \sum \gamma_i / u$$

Hence, F reduces to F-variable provided that

$\gamma \overset{\sim}{=} \eta_\cdot$ and χ^2_u and χ^2_{T-r-1} are stochastically independent.

From the book of "Linear Models" (Searle Page 59) we have this theorem.

Theorem : Let $X \sim N(\mu, V)$; then the quadratic forms $x' A x$ and $x' B x$ are distributed independently if and only if $A V B = 0$ or equivalently $B V A = 0$

Now with the help of above theorem and equations (4.45) and (4.46) we get the condition that $G^2(1)$ and $G^2(2/1)$ are asymptotically distributed if and only if

$$(A' D_\Pi^{-1} A) V \left[X E_2' (\tilde{X}_2' P \tilde{X}_2) E_2 X' \right] = 0$$

$$\left[\because \sqrt{n}\ (\hat{p} - r) \longrightarrow N(0, V) \right] \qquad (4.47)$$

The condition (4.47) only holds if $V = \lambda P$ for some constant λ.

Since

$$A\,V\,X = \lambda \left[I - P\,X\ (X' P\,X)^{-1} X' \right] P\,X = 0$$

In this case $\gamma_{\cdot} = \eta_{\cdot} = \lambda_{\cdot}$ and F is asymptotically distributed as F'-variate with u and T-r-1 d.f under $H_{2.1}$

4.10 CONSTANT CELL AND MARGINAL DESIGN EFFECT:

Again consider the equation discussed above as

$$E(G_1^2) = \sum_{\theta} (1 - r_{\theta}) d_{\theta} - \sum\sum_{i\theta_i} (1 - r_{\theta_i})\ d_{\theta_i}$$

$$+ \sum\sum_{j\theta_j} (1 - r_{\theta_j})\ d_{\theta_j}$$

where d_{θ}'s, d_{θ_i}'s, d_{θ_j}'s are cell and marginal design effects. When the cell and marginal design effects are all equal to δ, the above equation reduces to

$$\bar{E}(G_1^2) = \delta \left(\sum_{\theta} (1 - r_{\theta}) - \sum_i \sum_{\theta_i} (1 - r_{\theta_i}) + \sum_j \sum_{\phi_j} (1 - r_{\phi_j}) \right)$$

$$= \delta\ (T-1) - \Sigma_i\ (\text{cells in } C_{\theta i} - 1) + \Sigma_j (\text{No of cells in } C_{\phi_j} - 1)$$

$$= \delta \left[T - \Sigma_i\ (\text{No of cells in } C_{\theta i} + \Sigma_j\ \text{No of cells in } C_{\theta_j} \right)$$

$$= \delta\ (\ T - \text{no of independent parameters in } H_1)$$

Similarly

$\bar{E}(G^2\ 2/1)) = \delta$ (T- no of independent parameters in $H_{2.1}$)

Thus $\bar{E}(G^2) = u\delta. = E(G^2(2/1)) - E(G^2(1))$

$= \delta$ [no of independent parameters in H_1

- no of independent parameters in Hb]

$= \delta\ u$

or $\delta. = \delta$

Hence, under Ho:- $\frac{X^2}{\delta.} = \frac{X^2}{\delta}$ has asymptotically χ^2 distribution

with u - degrees of freedom, where u is the difference of the number of independent parameters in two models.

4.11 A JACKKNIFED CHI-SQUARE TEST:

This test was proposed by Fay (1985) and based on a modified Jackknife procedure applied to the Pearson or likelihood ratio test statistics themselves that is closely related to the work of Rao and Scott especially to their proposed statistic X^2/δ. Under some asymptotic conditions, X^2/δ. rejects null hypothesis at a rate significantly higher than the nominal level. The Jackknifed test incorporates an additional adjustment to avoid this property. The Jackknified statistic presented here includes a quantity K^+ based on variability in X^2 or G^2 over a set of replications. Under the null hypothesis, K^+/k. is a consistent estimator of δ, where k is the number of degrees of freedom. So we can use $X^2/(K^+/k)$ as an estimate of

X^2/δ. The quantity K^+ is readily computed even when no-closed form expression is available for estimates under log-linear models. One of the disadvantages of X^2/δ., however, is its rejection of the null hypothesis at a rate much greater than the nominal level when the corresponding δ_j's vary substantially from one another. The Jackknifed test incorporate a correction to avoid this problem.

Let us suppose that Y represents an observed cross-classification, possibly in the form of estimated population totals for a finite population derived from a complex surveys. Let $Y + W^{(i,j)}$ represents pseudoreplicates i=1,2, ---, I and j=1,2, --- Ji. Now the asymptotic theory for Jackknifed test requires that

$$\Sigma_j W^{(i,j)} = 0 \qquad - (4.48)$$

for each i. let $Cov^*(Y)$ is an estimate of covariance of Y and given as below

$$Cov^*(Y) = \Sigma_i b_i \Sigma_j W^{(i,j)} \oplus W^{(i,j)} \qquad - (4.49)$$

where $W^{(i,j)} \oplus W^{(i,j)}$ denotes the outer product of $W^{(i,j)}$ and bi is fixed set of constants appropriate for the problem under consideration. Assume that $W^{(i,j)}$ is uniformly small relative to the sampling variance of Y, this assumption is made to apply the asymptotic theory.

Let Y can be represented as the sum of n, i.i.d random variable $Z^{(j)}$ which is required for the application of the standard Jackknifing technique. The standard leave-one-out replicates. $Y^{(-j)} = Y - Z^{(j)}$ may be reweighted to the same expected total by the factor $\frac{n}{n-1}$ and written as

$$\left[\frac{n}{n-1}\right] Y^{-(j)} = \frac{n}{n-1}\,(Y - Z^{(j)}) = \frac{n\,Y}{n-1} - \frac{n}{n-1}\,Z^{(j)}$$

$$= \frac{(n-1)+1}{n-1} Y - \frac{n}{n-1}\,Z^{(j)}$$

$$= Y - \frac{(Y - n\,Z^{(j)})}{n-1} \qquad - (4.50)$$

Here we can see that $W^{(i,j)} = \dfrac{Y - n\,Z^{(j)}}{n-1}$, the subscript i is fixed at 1, $b_i = \dfrac{n-1}{n}$

Also

$$\sum_j W^{(i,j)} = \frac{nY - nY}{n-1} = 0$$ so it is also satisfying the condition required for this theory. Now assume that the universe may be considered to be divided in to L strata. The Jackknife technique may be adopted to this problem if the samples are selected independently from each stratum and if Y may be represented as

$$Y = \sum_h \sum_j Z^{(h,j)}, \quad \text{where } h = 1,2,--,L \text{ and } Z^{(h,j)}$$

is the n_h i.i.d. random variable within the stratum h. For each stratum h we have

$$Y + W^{(h,j)} = Y + \left[(\Sigma_{j'} Z^{(h,j')}) - n_h\, Z^{(h,j)} \right] \Big/ (n_h - 1) \qquad - (4.50)$$

This has the same expected value as Y

Here $W^{(h,j)} = \left[\Sigma_{j'} Z^{(h,j')} - n_h Z^{(h,j)} \right] / (n_h - 1)$

and $b_i = \dfrac{n_h - 1}{n_h}$

4.11.1 BALANCE HALF SAMPLES :

As it is well known that alternative approach to the variance estimation in the case of complex survey design is balanced half sample, selected to represent all source of variability with in the sampling design. Let in a design with two selection in each of the strata, half samples may be formed by picking one from each pairs of selections. Let $Z^{(h,1)}$, h=1,2,--,Z, represent the estimates of the totals, each based on half samples where each may be represented as

$$Z^{(h,1)} = Y + (Z^{(h,1)} - Y) \qquad - (4.51)$$

Here, $W^{(h,2)} = Y - Z^{(h,1)}$, $W^{(h,1)} = Z^{(h,1)} - Y$

An additional modification is necessary, however, since the asymptotic theory as discussed in the case of Jackknifing requires $W^{(h,j)}$ to be uniformaly small relative to the variation of Y and the half - sample estimate lacks this property.

So

$$W^{(h,1)} = d\,(Z^{(h\ 1)} - Y)$$

$$W^{(h,2)} = -\,W^{(h,1)}$$

and $b_i = \dfrac{1}{(2d^2L)^{-1}}$

generalizes the notion of half- sample replication. The asymptotic conditions may be met by allowing d to converge to 0 in a suitable manner. The sequence of half-samples may be based on either independent selection or balanced repeated replication.

4.11.2 STATISTICS :

The Jackknifed values of test statistics require refitting the given log-linear model to the replicate $Y + W^{(h,j)}$ and recomputing test statistics, $X^2 (Y + W^{(h,j)})$, $G^2(Y \quad W^{(h,j)})$ for the new tables proposed by Fay(1985). The standard formula for X^2 and G^2 are given below

$$X^2(p^*) = n \sum_i \frac{(p_i^* - r_i^*(p^*))^2}{r_i^*(p^*)}$$

$$G^2(p^*) = 2n \sum_i p_i^* \log_e \left[\frac{p_i^*}{r_i^*(p^*)} \right]$$

Now for weighted Y (or $Y + W^{(h,j)}$) the sum of the cell estimates of Y (or $Y + \omega^{(h,j)}$) replaces n in the sual X^2 or G^2. Using the b_i Jackknifed statistics X_J is defined as

$$X_J = \left[\left\{ X^2(Y) \right\}^{1/2} - (K^+)^{1/2} \right] \Big/ \left[V \Big/ \left\{ 8\, X^2(Y) \right\} \right]^{1/2}$$

- (4..52)

where $P_{hj} = X^2(Y + W^{(h,j)}) - X^2(Y)$ - (4.53)

$$K = \sum_h b_h \sum_j P_{hj} \qquad \text{- (4.54)}$$

$$V = \sum_h b_h \sum_j P_{hj}^2 \qquad \text{- (4.55)}$$

and K^+ takes the value for positive K and Zero otherwise. A test of difference of two chi-square test under nested model M_1 and M_2 is

given by

$$G_J = \frac{\left[G^2_{(2)}(Y) - G^2_{(1)}(Y) \right]^{1/2} - (K^+)^{1/2}}{\left\{ V/\ 8G^2_2\ (Y) - 8G^2_1\ (Y) \right\}^{1/2}} \qquad - (4.56)$$

where ,

$$P_{hj} = G^2_{(2)}\ (Y + W^{(h,j)}) - G^2_1\ (Y + W^{(h,j)}) - (G^2_{(2)}(Y) - G^2_{(1)}(Y)) \qquad - (4.57)$$

It is clear when M_1 is a saturated model then (4.57) becomes the direct analogue of (4.52). The only difference is instead of G^2 we use chi-square in the equation (4.52). We will now discuss the properties of this limiting distribution under the null hypothesis.

4.11.3 PROPERTIES OF THE LIMITING DISTRIBUTION :

We have the following limiting or asymptotic properties as

$n \longrightarrow \infty$

(1) Population proportion r are fixed and satisfy the linear model in question.

(2) The sample proportion estimated by Y_n are consistent for the population proportion r, and there are constants g and h depending on n such that $h_n\ (Y_n - g_n\ r) \xrightarrow{L} N(0, V)$ with non-zero covariance V .

(3) The ${}_nW^{(i,j)}$s introduced above are such that as $n \longrightarrow \infty$

$$\max_{(i,j)} h_n \parallel {}_nW^{(i,j)} \parallel \xrightarrow{P} 0$$

(4) Covariance estimator when multiplied by h^2_n converges in probability to V.

(5) The condition (4.48) is strictly satisfied for each n.

Under these conditions, X_J and G_J are shown to have as there limiting distribution of

$$X_J = \frac{\left[\sum_{i=1}^{k} \delta_i \chi^2_{(i)}\right]^{1/2} - \left[\sum_{i=1}^{k} \delta_i\right]^{1/2}}{\left\{\left[\sum_{i=1}^{k} \delta_i \chi_i^2\right] \Big/ \left[2\sum_{i=1}^{k} \delta_i \chi_i^2\right]\right\}^{1/2}} \qquad - (4.58)$$

Where $\chi^2_{i,}$ i = 1,2,--,k are set of independent Chi-square variate, each on single degree of freedom. δ_i's are set of non-negative weights depending on V and model in the question. k is the degrees of freedom of the test for multinomial sampling. In the case of multinomial sampling $\delta_i = 1$ $\forall$ i so the above equation (4.58) reduces to simple monotonic transformation of Chi-square distribution

$$X_J = 2^{1/2} \left\{(\chi^2_k)^{1/2} - (k)^{1/2}\right\} \qquad - (4.59)$$

The more general relationship between X^2/δ and X_J maybe seen by expressing

$$X_J = A^* \left[2^{1/2} \left\{\chi^2 (Y) \Big/ (k^+/k)^{1/2} - (k)^{1/2}\right\}\right] \qquad - (4.60)$$

where $A^* = \left[\left\{4\, X^2(Y)\, k^+\right\} \Big/ (k\, v)\right]^{1/2.}$

For equal δ_i's, A^* converges in probability to one. Comparison of X_J / A^* to the distribution of (4.59) gives a test asymptotically equivalent to the comparison of X^2/δ.

ESTIMATION OF PARAMETERS

ESTIMATION OF PARAMETERS

5.1 INTRDUCTION :

Sample surveys are generally designed to produce reliable estimates of simple descriptive parameters such as population totals, means and proportions. The sampling design that best meets desired objectives iß often complex and clustered because of cost and operational constraints in designing and implementing survey. As discussed previously, there has been growing trend to use survey data for statistical analysis, beyond the estimation of simple descriptive parameters. In case of categorical data analysis in survey data, the modification to the ordinary chi-square or likelihood-ratio test statistics is necessary to draw the required inferences correctly. Further, these modifications can be improved by estimating the various parameters of interest by taking care of the survey design applied for collecting sampling units. Hence, there is a need to adopt suitable estimators for the parameters, namely cell proportions and their variance-covariances. In this chapter some methods of estimation of these parameters are discussed. The various techeniques of variance estimations for complex surveys are discussed and compared in the case of categorical data analysis in survey sampling.

5.2 THE HORVITZ-THOMPSON ESTIMATOR :

Assume that N is the total number of sampling units in the population and n, is the number of units selected with the help of sampling design p(s). Let (i.j.)-th cell of the contingency table have N_{ij} element in the population, and let r_k denote the inclusion probability of the k-th element where i=1,2,---,I; j=1,2,---,J. Define, $Y_{ijk} = 1$, if k-th unit of the population falls in (i,j)-th cell and $Y_{ijk} = 0$, otherwise. The Horvitz-Thompson estimator for the proportion of (i,j)-th cell is given as

$$\hat{P}_{ij_{HT}} = \sum_{k=1}^{N} \frac{Y_{ijk}\ \delta_{ijk}}{N\ \Pi_k} \qquad -(5.1)$$

where,

$\delta_{ijk} = 1$, if $k \in s$ and $\delta_{ijk} = 0$, otherwise It's variance can be given by following equation

$$V(\hat{p}_{ij})_{HT} = \frac{1}{2N^2} \sum_{k=1}^{N} (\Pi_k \Pi_l - \Pi_{kl}) \left(\frac{Y_{ijk}}{\Pi_k} - \frac{Y_{ijl}}{\Pi_l} \right)^2 \qquad (5.2)$$

and its estimate of variance as proposed by yates and Grundy (1953) is given by

$$\hat{V}(\hat{p}_{ij})_{HT} = \frac{1}{N^2} \sum_{k<l}^{n} \frac{(\Pi_k \Pi_l - \Pi_{kl})}{\Pi_{kl}} \left(\frac{Y_{ijk}}{\Pi_k} - \frac{Y_{ijl}}{\Pi_l} \right)^2 \qquad -(5.3)$$

5.3 COMBINED RATIO ESTIMATOR OF PROPORTION :

Most common sampling design in survey sampling is stratified multi-stage sampling because of its advantages over the other sampling techniques. Let us assume that the population is divided into L strata, N_h is the total number of primary

sampling units (PSU's) in the h-th stratum; M_{ht} is the total number secondary sampling units in the t-th PSU of the h-th stratum and let $M_h = \sum_{t=1}^{N_h} M_{ht}$:Denote total number of SSU's in the population by M and weight for h-th stratum by M_h / M. Let the PSU's are selected with simple random sampling with replacement, and let Π_{ht} denote the probability of selecting t-th PSU from the h-th stratum. Similarly, let Π_{htk} be the probability of selecting k-th SSU from t-th PSU of h-th stratum.

Define

Y_{ihtk} = 1, if, k-th SSU from t-th PSU of h-th stratum falls in i-th category.

= 0, otherwise.

x_{htk} = 1, if, k-th SSU from t-th PSU falls in the h-th stratum.

= 0 otherwise.

Again, let m_{ht} denote the number of SSU's selected from t-th PSU of h-th stratum and n_h be the number of PSU's selected from the h-th stratum. The estimate of total units in the t-th PSU of h-th stratum in the i-th category can be given as

$$\hat{Z}_{iht} = \frac{1}{m_{ht}} \sum_{k=1}^{m_{ht}} \frac{Y_{ihtk}}{\Pi_{htk}}$$

where $k = 1,2,\text{----}, m_{ht}$

$h = 1,2,\text{----}, L$

$t = 1,2,\text{----}, n_h$

The estimate of total number of units in the t-th PSU of h-th stratum is given as

$$\hat{Z}_{ht} = \frac{1}{m_{ht}} \sum_{k=1}^{m_{ht}} \frac{x_{htk}}{\Pi_{htk}}$$

let $$z_{iht} = \frac{\hat{Z}_{iht}}{M_h \Pi_{ht}}$$

$$z_{ht} = \frac{\hat{Z}_{ht}}{M_h \Pi_{ht}}$$

$$\bar{z}_{ih} = \frac{1}{n_h} \sum_{t=1}^{n_h} z_{iht}$$

$$\bar{z}_i = \sum_{h=1}^{L} W_h z_{ih}$$

$$\bar{z}_h = \frac{1}{n_h} \sum_{t=1}^{n_h} z_{ht}$$

$$\bar{z} = \sum_{h=1}^{L} W_h \bar{z}_h$$

So, the combined ratio estimator is given by

$$\hat{P}_{ic} = \frac{\bar{z}_i}{\bar{z}} = \hat{\theta} \text{ (say)} \qquad - (5.4)$$

Now, as we know its mean square error can be given as

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

$$= E\left[\frac{\bar{z}_i}{\bar{z}} - \theta\right]^2$$

$$= E\left[\frac{\bar{z}_i - \bar{z}\theta}{\bar{Z}}\right]^2 \qquad \text{where } E(\bar{z}) = \bar{Z}$$

$$= E\frac{\left[\bar{z}_i - \bar{Z}_i - \theta(\bar{z} - \bar{Z})\right]^2}{\bar{Z}^2} \qquad [\because \theta \bar{Z} = \bar{Z}_i]$$

$$= \frac{1}{\bar{z}^2}\left[V(\bar{z}_i) + \theta^2 V(\bar{z}) - 2\,\theta\, Cov(\bar{z}_i, \bar{z}) \right] \qquad - (5.5)$$

So, the estimate of mean square error is written as

$$\hat{MSE}\ (\hat{\theta}) = \frac{1}{\bar{z}^2}\left[\sum_{h=1}^{L} \frac{W_h^2}{n_h}\, s_{hzi}^2 + \frac{\bar{z}_i^2}{\bar{z}^2} \sum_{h=1}^{L} \frac{W_h^2}{n_h}\, s_{hz}^2 - 2\frac{\bar{z}_i}{\bar{z}} \sum_{h=1}^{L} \frac{W_h^2}{n_h}\, s_{hzzi} \right] \qquad - (5.6)$$

Where,

$$s_{hz}^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (z_{hi} - \bar{z}_h)^2$$

$$s_{hzi}^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (z_{ihi} - \bar{z}_{ih})^2$$

$$s_{hzzi} = \frac{1}{n_h - 1} \sum_{h=1}^{n_h} (z_{ihi} - \bar{z}_{ih})\,(z_{hi} - \bar{z}_h)$$

5.4 ESTIMATOR OF MEAN SQUARE DUE TO POST-STRATIFIED WEIGHTING :

We can get an estimator of mean square error due to post-stratification weighting by considering the stratification according to various catagories and replacing y_{ihik} in equation (5.6) by $(y_{inik} - \bar{z}_i)$ (William, 1960). We get

$$MSE_p\ (\hat{\theta}) = \frac{1}{\bar{z}^2}\left[\sum_{h=1}^{L} \frac{W_h^2}{n_h}\, s_{hbi}^2 + \frac{\bar{b}_i^2}{\bar{z}^2} \sum_{h=1}^{L} \frac{W_h^2}{n_h}\, s_{hz}^2 - 2\frac{\bar{b}_i}{\bar{z}} \sum_{h=1}^{L} \frac{W_h^2}{n_n}\, s_{hzbi} \right] - (5.7)$$

where,

$$\hat{B}_{iht} = \frac{1}{m_{ht}} \sum_{k=1}^{m_{ht}} \frac{y_{ihik} - \bar{z}_i}{\Pi_{htk}}$$

$$b_{iht} = \frac{\hat{B}_{iht}}{M_h \Pi_{ht}}$$

$$\tilde{b}_{ih} = \frac{1}{n_h} \sum_{t=1}^{n_h} b_{iht}$$

$$\bar{b}_i = \sum_{h=}^{L} W_h \bar{b}_{ih}$$

$$s^2_{hbi} = \frac{1}{n_{n-1}} \sum_{t=1}^{n} (b_{iht} - \bar{b}_{ih})^2$$

$$s_{hzbi} = \frac{1}{n_h - 1} \sum_{t=1}^{n_h} (b_{iht} - \bar{b}_{ih})\ (z_{iht} - \bar{z}_{ih})$$

5.5 POST-STRATIFIED ESTIMATOR OF PARAMETERS :

Consider the case of stratified two-stage sampling. Let θ_1 and θ_2 be two independent sets of subscripts. For example, if in a survey four characteristics of a sampling unit are measured and their levels are denoted by i.j,m,l then θ_1 = {i,j}; and θ_2 = {m,l}. Suppose we have information for the combination of variables associated with the subscript in θ_2 by their projection from a previous survey. Here, each subscript is varying between one to the number of categories of the corresponding variables. Define the variable ${}_{\theta 2}y_{\theta 1(htk)}$ for the k-th sample element in the t-th first stage unit of the h-th stratum as one if element belongs to θ_1-th category and θ_2-th group and is zero, otherwise. Similary ${}_{\theta 2}x_{htk}$ is equal to one if (htk)-th element falls in θ_2-category and is Zero, otherwise, where h=1,2,--L; t=1,2,---, n_h; k = 1,2,--, m_{ht}.The sampling weight attached to (htk)th element is denoted by W_{htk}. The estimate of total count in θ_1-th category is given by

$$\hat{N}_{\theta_1} = \sum_{\theta_2} (\theta_2 \hat{N}_{\theta_1} / \hat{N}_{\theta_2}) \theta_2^N$$

where

$$\theta_2 \hat{N}_{\theta_1} = \sum_h \sum_t \left[\sum_k W_{htk} \theta_2^{y_{\theta_1(htk)}} \right] \sum_h \sum_t \theta_2^{B_{\theta_1(ht)}} \text{ (say)}$$

$$\theta_2 \hat{N} = \sum_h \sum_t \left[\sum_k W_{htk} \theta_2^{x_{htk}} \right] = \sum_h \sum_t \theta_2^{B_{ht}} \text{ (say)}$$

Now, the proportion for the θ_1-th category is given by

$$\hat{p}_{\theta_1} = \frac{\hat{N}_{\theta_1}}{\hat{N}} \qquad - (5.8)$$

where $\hat{N}_{\theta_1} = \sum_{\theta_2} \theta_2 \hat{N}_{\theta_1}$ and $\hat{N} = \sum_{\theta_1} \hat{N}_{\theta_1}$

similarly. Unadjusted estimate of proportion is given by

$$\hat{p}^*_{\theta_1} = \frac{\hat{N}^*_{\theta_1}}{\hat{N}^*} \qquad - (5.9)$$

where $\hat{N}^*_{\theta_1} = \sum_{\theta_2} \theta_2 \hat{N}^*_{\theta_1}$ and $\hat{N}^* = \sum_{\theta_1} \hat{N}^*_{\theta_1}$

where $\hat{N}^*$ is Horvitz-Thompson estimator $\sum_h \sum_t \sum_k W_{htk} y_{\theta_1(htk)}$.

The estimated variance-Covariance of the above estimator was given by the help of the following equations. Consider that the first stage units are assumed to be sampled with replacement within strata, then it is well known that

$$\hat{Cov}(\hat{N}^*_{\theta_1}, \hat{N}^*_{\theta_1'}) = \sum_{h=1}^{L} \frac{n_h}{n_h - 1} \sum_{t=1}^{n_h} \left(B_{\theta_1(ht)} - \bar{B}_{\theta_1(h)} \right) \left(B_{\theta_1(ht)} - \bar{B}_{\theta_1'(h)} \right) \qquad -(5.10)$$

where $\bar{B}_{\theta_1(h)} = \sum_t B_{\theta_1(ht)} / n_h$

Now, with the help of the paper by William (1960). The estimated covariance of post-stratified estimator $\hat{N}_{\theta_1}$, and $\hat{N}_{\theta_1'}$ is simply obtained by changing $Y_{\theta_1(htk)}$ to $Y_{\theta_1(htk)} - \Sigma_{\theta_2}(_{\theta_2}\hat{N}_{\theta_1}/_{\theta_2}\hat{N})$

$$_{\theta_2}x_{htk}) = B_{\theta_1(ht)} - \Sigma_{\theta_2}\ (_{\theta_2}\hat{N}_{\theta_1}/\ _{\theta_2}\hat{N})\ _{\theta_2}B_{ht}$$

$$\text{so}\quad \hat{Cov}\ (\hat{N}_{\theta_1}, \hat{N}_{\theta_1'}) = \sum_{h=1}^{L} \frac{n_h}{n_h - 1} \sum_{i} (Z_{\theta_1 ht} - \bar{Z}_{\theta_1 h})$$

$$* (Z_{\theta_1' ht} - \bar{Z}_{\theta_1' h})$$

where $Z_{\theta_1 ht} = B_{\theta_1(ht)} - \Sigma_{\theta_2}(_{\theta_2}\hat{N}_{\theta_1}/\ _{\theta_2}\hat{N})\ _{\theta_2}B_{ht}$

And

$$B_{\theta_1(ht)} = \sum_{\theta_2} {}_{\theta_2}B_{\theta_1(ht),}$$

$$\bar{Z}_{\theta_1 h} = \sum_{i} Z_{\theta_1 ht}/n_h$$

The estimated variance is obtained by putting $\theta_1 = \theta_1'$. The estimate $\hat{p}_i$ is the ratio of two post-stratified estimates hence estimated covariance of $\hat{p}_{\theta_1}$ and $\hat{p}_{\theta_1'}$ is obtained by changing

$$Y_{\theta_1(htk)} \text{ to } \bar{Y}_{\theta_1(htk)} = \left[Y_{\theta_1(htk)} - \sum_{\theta_2} (_{\theta_2}\hat{N}_{\theta_1}/_{\theta_2}\hat{N}) \right.$$

$$\left. _{\theta_2}x_{htk} \right] - \hat{p}_i \left[Y_{+(ht)} - \Sigma\ (_{\theta_2}N_{\theta_1}/_{\theta_2}\hat{N})\ _{\theta_2}x_{htk} \right]$$

where $Y_{+(htk)} = \sum_{i} Y_{\theta_1(htk),}$ i.e. $B_{\theta_1(ht)}$ is replaced by

$$\sum_{k} W_{htk}\ \bar{Y}_{\theta_1(htk)} = \left[B_{\theta_1(ht)} - \hat{p}_i\ B_{(ht)} \right]$$

$$- \Sigma_{\theta_2}\ (_{\theta_2}B_{ht}\ /_{\theta_2}\hat{N})\ (_{\theta_2}\hat{N}_{\theta_1} - \hat{p}_i\,_{\theta_2}\hat{N}) \text{ which leads to}$$

following equation

$$\hat{Cov}\ (\hat{P}_{\theta_1},\ \hat{P}_{\theta'_1}) = N^{-2}\ \hat{\sigma}_{\theta_1\theta'_1}$$

where $\hat{\sigma}_{\theta_1,\theta_1'}$ is the equation obtained through the above replacements.

5.6 COMPARISON OF VARIANCE ESTIMATION TECHNIQUES FOR COMPLEX SURVEYS FOR COMBINED RATIO ESTIMATOR :

Many large scale surveys now involve large number of strata with relatively few primary sampling units (PSU's) selected with in each stratum. In recent years, problems of statistical inference based on the data from such stratified cluster samples have received considerable attention. In particular, four general methods of estimating the variance of non-linear statistics, such as ratio or proportions have been advanced. These are : Taylor expansion or linearization, the Jackknife procedure, balanced repeated replication (BRR) and bootstrap technique. Now, the asymptotic theory for all the above methods will be discussed under the following regularity conditions.

(1) Max $(nW_h)/n_h)=0$ (1)as $n\longrightarrow\infty$, which allows L to be either bounded or unbounded. If the stratum sizes are bounded i.e. max $n_h=0$ (1) which reflects our intention to focus on surveys with large number of strata with relatively few PSU's selected with in each stratum, the above condition is equivalent to $\max_h W_h=$

$O(n^{-1})$ These two conditions roughly say that the allocation of sample across strata should not be disproportionately small relative to the stratum weights. Second condition is satistifed only if $\max_h W_h = O(L^{-1})$ and $n/L = O(1)$, that is, if no stratum is of disproportionate size and average sample size per stratum is bounded above.

(2) $\sum W_h S_{hjk} = O(1)$, where $S_{hjk} = E(Y_{hij} - \bar{Y}_{hj})(Y_{hik} - \bar{Y}_{hk})$. This condition means that the weighted avarage of within stratum covariances is bounded. This, together with the regularity condition (1) implies that $Var(\hat{y}_j - \hat{Y}_j) = O(n^{-1})$, elements $\bar{y}_j - \bar{Y}_j$ are of order not greater than $n^{-0.5}$ in probability which is denoted by $O_p(n^{-0.5})$

(3) $E(\hat{\theta}) = O(1)$

(4) $E(\bar{y}_j - \bar{Y}_j)^6 = O(n^{-3})$,

(5) There exists a closed bounded neighborhood B of $\bar{Y}_j$ such that all third derivatives are continuous and bounded in B. The condition (4) is satisfied if, (1) holds and weighted average of within-stratum moments (sixth order and lower) is bounded

consider

$$\hat{p}_{ic} = \frac{\bar{z}_i}{\bar{z}} = \hat{\theta} \text{ (say)}$$

where $\bar{z}_i = \sum_{h=1}^{L} \frac{m_h}{M} \bar{z}_{ih}$

$$= \sum_{h=1}^{L} \frac{m_h}{M} \frac{1}{n_h} \sum_{t=1}^{n_h} z_{iht}$$

$$= \sum_{h=1}^{L} \frac{m_h}{M} \frac{1}{n_h} \sum_{t=1}^{n_h} \frac{\hat{z}_{iht}}{M_h \Pi_{ht}}$$

$$= \sum_{h=1}^{L} \frac{1}{M} \frac{1}{n_h} \sum_{t=1}^{n_h} \frac{\hat{z}_{ihk}}{\Pi_{ht}}$$

So,

$$E(\bar{z}_i) = E_1 \left[E_2 (\bar{z}_i / t) \right]$$

$$= \sum_{h=1}^{L} \frac{1}{m} E_1 \left\{ \frac{1}{n_h} \sum_{t=1}^{n_h} \frac{z_{iht}}{\Pi_{ht}} \right\}$$

$$= \sum_{h=1}^{L} \frac{z_{ih}}{M} = \frac{z_i}{M} = P_i = \bar{Z}_i \text{ (say)}$$

Similarly,

$$\bar{z} = \sum_{h=1}^{L} W_h \bar{z}_h$$

$$= \sum_{h=1}^{L} \frac{m_h}{M} \frac{1}{n} \sum_{t=1}^{n_h} z_{ht}$$

$$= \sum_{h=1}^{L} \frac{M_h}{M} \frac{1}{n_h} \sum_{t=1}^{n_h} \frac{\hat{z}_{ht}}{M_h \Pi_{ht}}$$

so $E(\bar{z}) = E_1 \left[E_2 (\bar{z}/t) \right] = E_1 \left\{ \sum_{h=1}^{L} \frac{1}{M} \frac{1}{n_h} \sum_{t=1}^{n_h} \frac{z_{ht}}{\Pi_{ht}} \right\}$

$$= \sum_{h=1}^{L} \frac{M_h}{M} = 1 = \bar{Z} \text{ (say)}$$

Now various variance-covariance estimators of complex survey were developed and compared for the categorical data analysis.

Let $\theta = \frac{\bar{Z}_i}{\bar{Z}}$ or $\theta = \bar{Z}_i$

(1) Jackknifed Estimator : Let $\hat{\theta}^{ht} = \frac{\bar{z}_i^{ht}}{\bar{z}^{ht}} = \frac{\hat{Z}_i^{ht}}{\hat{Z}^{ht}}$

where, $\hat{\theta}^{h} = \frac{1}{n_h}\sum_{ht=1}^{n_h}\hat{\theta}^{ht}$ and $\hat{\theta}^{ht} = n_h\hat{\theta} - (n_h - 1)\hat{\theta}^{ht}$ where, $\bar{z}_i^{ht}$ is the unbiased estimator of $\bar{Z}_i$ from the sample after omitting t-th selected PSU from the h-th stratum. Similarly, $\bar{z}^{ht}$ is an unbiased estimator of $\bar{Z}$ after omitting t-th PSU from the h-th stratum. t = 1,2,--,N_h; h=1,2,---.L;i=1,2,---, I. We can also express

$$\hat{Z}_i^{ht} = M\, \bar{z}_i^{ht}$$

$$\hat{Z}^{ht} = M\, \bar{z}^{ht}$$

where,

$$\bar{z}_i^{ht} = \sum_{h' \neq h} W_h \cdot \bar{z}_{ih} + W_h (n_h \bar{z}_{ih} - z_{iht})/(n_h - 1)$$

$$= \sum_h W_h \bar{z}_{ih} - W_h \bar{z}_{ih} + W_h (n_h \bar{z}_{ih} - \bar{z}_{iht})/(n_h - 1)$$

$$= \sum_h W_h \bar{z}_{ih} + \left[-1 + \frac{n_h}{n_{h-1}}\right] W_h \bar{z}_{ih} - \frac{W_h z_{iht}}{n_{h-1}}$$

$$= \sum_h W_h \bar{z}_{ih} + W_h (\bar{z}_{ih} - \bar{z}_{iht})/(n_h - 1)$$

$$= \bar{z}_i + W_h (\bar{z}_{ih} - z_{iht})/(n_h - 1) \qquad -(5.11)$$

Similarly we can write

$$\bar{z}^{ht} = \bar{z} + W_h (\bar{z}_h - z_{ht})/(n_h - 1) \qquad (5.12)$$

Jone (1974) proposed the following estimator

$$\tilde{\theta}_j^{(1)} = (n+1-L)\,\hat{\theta} - \sum (n_h - 1)\hat{\theta}^h \qquad -(5.13)$$

Two other Jackknife estimators can be constructed with the help of the "pseudo-Valus" $\tilde{\theta}^{hi}$ as

$$\theta_j^{(2)} = \frac{1}{n} \sum_h \sum_i \tilde{\theta}^{hi} \qquad -(5.14)$$

$$\tilde{\theta}_j^{(3)} = \frac{1}{L} \sum_{h=1}^{L} \frac{1}{n_h} \sum_{i=1}^{n_h} \tilde{\theta}^{hi} \qquad -(5.15)$$

Assume that $L \geq 2$

$$\text{Bias}\,(\bar{\theta}_j) = \text{Bias}\,(\hat{\theta}) - E\,(\hat{B}_j)$$

where $\hat{B}_j = \hat{\theta} - \tilde{\theta}_j$

Now $\tilde{\theta}_j^{(i)} = \hat{\theta} - \hat{B}_j^{(i)}$

so

$$\hat{B}_j^{(1)} = \hat{\theta} - (n+1+L)\,\hat{\theta} + \sum_h (n_h - 1)\,\hat{\theta}^h$$

$$= \sum_{h=1}^{L} (1-n_h)\,\hat{\theta} + \sum_h (n_h - 1)\hat{\theta}^h$$

$$= \sum_{h=1}^{L} (n_h - 1)\,(\hat{\theta}^h - \theta) \qquad -(5.16)$$

Similarly, we can find

$$\hat{B}_j^{(2)} = \sum_{h=1}^{L} n^{-1} n_h\,(n_n - 1)\,(\hat{\theta}_h - \hat{\theta}) \qquad -(5.17)$$

$$\hat{B}_j^{(3)} = L^{-1}\,\hat{B}_j^{(1)} \qquad -(5.18)$$

Now under the asymptotic frame work, by linearizing $\hat{\theta}$ we get

$$\hat{\theta} = \theta - (\bar{z} - \bar{Z}) \frac{\bar{Z}_i}{\bar{Z}^2} + \frac{1}{\bar{Z}} (\bar{z}_i - \bar{Z}_i) + \frac{\bar{Z}_i}{\bar{Z}^3} (\bar{z} - \bar{Z})^2$$

$$- \frac{1}{\bar{Z}^2} (\bar{z}_i - \bar{Z}_i) (\bar{z} - \bar{Z}) \qquad -(5.19)$$

Now, by neglecting higher order terms we get

$$(\hat{\theta} - \theta)^2 = (\bar{z} - \bar{Z}) \frac{\bar{z}_i^2}{\bar{Z}^4} + \frac{1}{Z^2} (\bar{z}_i - \bar{Z}_i)^2 - \frac{2Z_i}{\bar{Z}^3} (\bar{z}_i - \bar{Z}_i) (\bar{z} - \bar{Z})$$

So, the estimated linearized variance can be written as

$$\nu_L(\hat{\theta}) = \frac{1}{\bar{z}^2} \left[\frac{\bar{z}_i^2}{\bar{z}^2} \sum_{h=1}^{L} \frac{W_h^2}{n_h} s_{hz}^2 + \sum_{h=1}^{L} \frac{W_h^2}{n_h} s_{hzi}^2 \right.$$

$$\left. - 2 \frac{\bar{z}_i}{\bar{z}} \sum_{h=1}^{L} \frac{W_h^2}{n_h} s_{hzzi} \right] \qquad -(5.20)$$

where

$$s_{hz}^2 = \frac{1}{n_h - 1} \sum_{t=1}^{n_h} (\bar{z}_{ht} - \bar{z}_h)^2$$

$$s_{hzi}^2 = \frac{1}{n_h - 1} \sum_{t=1}^{n_h} (z_{iht} - z_{ih})^2$$

$$s_{hzzi} = \frac{1}{n_h - 1} \sum_{t=1}^{n_h} (z_{iht} - \bar{z}_{ih}) (z_{ht} - \bar{z}_h)$$

Let $$\hat{e}_{ht} = z_{iht} - \bar{z}_{ih} - \frac{\bar{z}_i}{\bar{z}} (z_{ht} - \bar{z}_h) \qquad -(5.21)$$

so from (5.21) we have

$$\frac{1}{n_h - 1} \sum_{t=1}^{n_h} \hat{e}_{ht}^2 = s_{hzi}^2 + \frac{\bar{z}_i^2}{\bar{z}^2} s_{hz}^2 - 2 \frac{\bar{z}_i}{\bar{z}} s_{hzzi}$$

so, equation (5.20) reduces to

$$\nu_L(\theta) = \frac{1}{\bar{z}^2} \sum_{h=1}^{L} \frac{W_h^2}{n_h} s^2_{\hat{e}h} \qquad -(5.22)$$

where

$$(n_h - 1)\, s^2_{\hat{e}h} = \sum_{t=1}^{n_h} \hat{e}^2_{ht}$$

Again by expanding $\hat{\theta}^{ht}$ to get the variances of Jackknife estimator we get

$$\hat{\theta}^{ht} = \frac{\bar{z}_i^{ht}}{\bar{z}^{ht}} = \frac{\bar{z}_i}{\bar{z}} - \frac{\bar{z}_i}{\bar{z}^2}(\bar{z}^{ht} - \bar{z}) + \frac{1}{\bar{z}}(\bar{z}_i^{ht} - \bar{z}_i) + \frac{\bar{z}_i}{\bar{z}^3}(\bar{z}^{ht} - \bar{z}) - \frac{1}{\bar{z}^{(2)}}(\bar{z}^{ht} - \bar{z}_i)(\bar{z}^{ht} - \bar{z})$$

From equation (5.11) we get

$$\hat{\theta}^{ht} = \hat{\theta} + \frac{W_h}{n_h - 1}\frac{\bar{z}_i}{\bar{z}^2}(\bar{z}_h - z_{ht}) + \frac{W_h}{n_h - 1}\frac{1}{\bar{z}}(\bar{z}_{ih} - z_{iht}) + \frac{W_h^2}{(n_h-1)^2}\frac{\bar{z}_i}{\bar{z}^3}(\bar{z}_h - z_{ht})^2 - \frac{W_h^2}{(n_h-1)^2}\frac{1}{\bar{z}^2}(\bar{z}_{ih} - z_{iht})(\bar{z}_h - z_{ht}) + Op(n^{-3}) \qquad -(5.23)$$

Now, we have the following Jackknife variance estimators.

$$v_j^{(1)}(\hat{\theta}) = \sum_{h=1}^{L} \frac{n_h - 1}{n_h} \sum_{t=1}^{n_h} (\hat{\theta}^{ht} - \hat{\theta}^h)^2 \qquad -(5.24)$$

$$v_j^{(2)}(\hat{\theta}) = \sum_{h=1}^{L} \frac{n_h - 1}{n_h} \sum_{t=1}^{n_h} (\hat{\theta}^{ht} - \hat{\theta})^2 \qquad -(5.25)$$

$$v_j^{(3)}(\hat{\theta}) = \sum_{h=1}^{L} \frac{n_h - 1}{n_h} \sum_{t=1}^{n_h} \left(\hat{\theta}^{ht} - \frac{1}{n} \sum_{h=1}^{L} \sum_{t=1}^{n_h} \hat{\theta}^{ht} \right)^2 \qquad -(5.26)$$

$$v_j^{(4)}(\hat{\theta}) = \sum_{h=1}^{L} \frac{n_h - 1}{n_h} \sum_{t=1}^{n_h} \left(\hat{\theta}^{ht} - \frac{1}{L} \sum_{h=1}^{L} \hat{\theta}^h \right)^2 \qquad -(5.27)$$

$$v_j^{(5)}(\hat{\theta}) = \sum_{h=1}^{L} \frac{1}{n_h(n_h - 1)} \sum_{t=1}^{n_h} (\hat{\theta}^{ht} - \tilde{\theta}^{(2)})^2 \qquad -(5.28)$$

$$v_j^{(6)}(\hat{\theta}) = \sum_{h=1}^{L} \frac{1}{n_h(n_h - 1)} \sum_{t=1}^{n_h} (\tilde{\theta}^{ht} - \tilde{\theta}^{(3)})^2 \qquad -(5.29)$$

The most commonly used estimator is $v_j^{(2)}(\hat{\theta})$ which can also be

written as

$$v_j^{(2)}(\hat{\theta}) = \sum_{h=1}^{L} \frac{1}{n_h(n_h-1)} \sum_{t=1}^{n_h} (\tilde{\theta}^{ht} - \hat{\theta})^2 \qquad -(5.30)$$

Since we know that

$$\hat{\theta}^{ht} = n_h \hat{\theta} - (n_h - 1)\hat{\theta}^{ht}$$

or $$\hat{\theta}^{ht} = \frac{n_h \hat{\theta} - \tilde{\theta}^{ht}}{n_h - 1}$$

or $$(\hat{\theta}^{ht} - \hat{\theta})^2 = \frac{1}{(n_h - 1)^2} (\hat{\theta} - \hat{\theta}^{ht})^2$$

Now, calculating the variance of Jackknife estimator we have from (5.23) as

$$\begin{aligned}
(\hat{\theta}^{ht} - \hat{\theta})^2 = {} & \left(\frac{W_h}{n_h - 1}\right)^2 \frac{\bar{z}_i^2}{\bar{z}^4} (\bar{z}_h - z_{ht})^2 + \left(\frac{W_h}{n_h - 1}\right)^2 \frac{1}{\bar{z}^2} (\bar{z}_{ih} - z_{iht})^2 \\
& - 2\left(\frac{W_h}{n_h - 1}\right)^2 \frac{\bar{z}_i}{\bar{z}^3} (\bar{z}_h - z_{ht})(\bar{z}_{ih} - z_{iht}) \\
& - 2\left(\frac{W_h}{n_h - 1}\right)^3 \frac{\bar{z}_i^2}{\bar{z}^5} (\bar{z}_h - \bar{z}_{ht})^3 \\
& + 2\left(\frac{W_h}{n_h - 1}\right)^3 \frac{\bar{z}_i}{\bar{z}^4} (\bar{z}_{ih} - z_{iht})(\bar{z}_h - z_{ht})^2 \\
& + \left(\frac{W_h}{n_h - 1}\right)^3 \frac{2\bar{z}_i}{\bar{z}^4} (\bar{z}_h - z_{ht})^2 (\bar{z}_{ih} - z_{iht}) \\
& - \left(\frac{W_h}{n_h - 1}\right)^3 \frac{2}{\bar{z}^3} (\bar{z}_{ih} - z_{iht})^2 (\bar{z}_h - z_{ht}) + Op(n^{-4})
\end{aligned}$$

So, the variance of the Jackknife estimator is written as

$$\nu_j^{(2)}(\hat{\theta}) = \sum_{h=1}^{L} \frac{n_h - 1}{n_h} \sum_{t=1}^{n_h} (\hat{\theta}^{ht} - \hat{\theta})^2$$

$$= \frac{\bar{z}_i^2}{\bar{z}^4} \sum_{h=1}^{L} \frac{W_h^2}{n_h} s_{hz}^2 + \frac{1}{\bar{z}^2} \sum_{h=1}^{L} \frac{W_h^2}{n_h} s_{hzi}^2 - 2 \frac{\bar{z}_i}{\bar{z}^3} \sum_{h=1}^{L} \frac{W_h^2}{n_h} s_{hzzi}$$

$$- \sum_{h=1}^{L} \frac{W_h^2}{n_h(n_h - 1)} \frac{1}{n_h - 1} \sum_{t=1}^{n_h} \left[\frac{2z_i}{\bar{z}^4} (z_{ht} - \bar{z}_h)^2 (z_{iht} - \bar{z}_{ih}) \right.$$

$$- \frac{2\bar{z}_i^2}{\bar{z}^4} (z_{ht} - \bar{z}_h)^3 - \frac{2}{\bar{z}^3} (z_{ht} - \bar{z}_h) (z_{iht} - \bar{z}_{ih})^2$$

$$\left. + \frac{2\bar{z}_i}{\bar{z}^4} (z_{ht} - \bar{z}_h)^2 (z_{iht} - \bar{z}_{ih}) \right]$$

$$\nu_L(\hat{\theta}) - \sum_{h=1}^{L} \frac{W_h^2}{n_h(n_h - 1)} \frac{1}{(n_h - 1)} \sum_{t=1}^{n_h} l_{ht} q_{ht} + Op(n^{-3}) \qquad -(5.31)$$

where

$$l_{ht} = \frac{1}{\bar{z}} \left[z_{iht} - \bar{z}_{ih} - \frac{\bar{z}_i}{\bar{z}} (z_{ht} - \bar{z}_h) \right] = \frac{\hat{e}_{ht}}{\bar{z}} \qquad -(5.32)$$

$$q_{ht} = \frac{2}{\bar{z}^2} \left[\frac{\bar{z}_i}{\bar{z}} (z_{ht} - \bar{z}_h)^2 - (z_{ht} - \bar{z}_h) (z_{iht} - \bar{z}_{ih}) \right]$$

$$q_{ht} = \frac{2}{\bar{z}^2} \left[(z_{ht} - \bar{z}_h) \hat{e}_{ht} \right] \qquad -(5.33)$$

The equation (5.31) can be further written as

$$\nu_j^{(2)}(\hat{\theta}) = \nu_L(\hat{\theta}) - \sum_{h=1}^{L} \frac{W_h^2}{n_h(n_h - 1)} \frac{1}{(n_h - 1)} \sum_{t=1}^{n_h} \frac{\hat{e}_{ht}}{\bar{z}} * \left[-\frac{2}{\bar{z}^2} (z_{ht} - \bar{z}_h) \hat{e}_{ht} \right]$$

$$= \nu_L(\hat{\theta}) + \sum_{h=1}^{L} \frac{W_h^2}{n_h(n_h - 1)} \frac{1}{(n_h - 1)} \frac{2}{\bar{z}^2} \sum_{t=1}^{n_h} \hat{e}_{ht}^2 (z_{ht} - \bar{z}_h)$$

$$= \nu_L(\hat{\theta}) + \frac{2}{\bar{z}^3} \sum_{h=1}^{L} \frac{W_h^2}{n_h(n_h-1)} s^2_{\hat{e}zh} \qquad -(5.34)$$

where $(n_h-1)\, s^2_{\hat{e}zh} = \sum_{t=1}^{n_h} \hat{e}^2_{ht} (z_{ht} - \bar{z}_h)$

Further,

$$\nu_j^2(\hat{\theta}) = \nu_L(\hat{\theta}) + \frac{2}{\bar{z}^3} \sum_{h=1}^{L} \frac{W_h^2}{n_h(n_h-1)} s_{e^2zh} + Op\,(n^{-2.5}) \qquad -(5.35)$$

where

$$(n_h-1)\, s^2_{ezh} = \sum_{t=1}^{n_h} (e_{ht} - \bar{e})^2 (z_{ht} - \bar{z}_h)$$

Now an attempt will be made to study the asymptotic relations of $\nu_j^{(i)}$. First consider $\nu_j^{(1)}$ and $\nu_j^{(2)}$

From $\sum_{t=1}^{n_h} (\hat{\theta}^{ht} - \hat{\theta})^2 = \sum_{t=1}^{n_h} (\hat{\theta}^{ht} - \hat{\theta}^h)^2 + n_h(\hat{\theta}^h - \hat{\theta})^2 \qquad -(5.36)$

We have from (5.23) as

$$\hat{\theta}^h = \frac{1}{n_h} \sum_{t=1}^{n_h} \hat{\theta}^{ht} = \hat{\theta} + \frac{W_h^2}{n_h(n_h-1)} \frac{\bar{z}_i}{\bar{z}^3} s^2_{hz} - \frac{W_h^2}{n_h(n_h-1)} \frac{1}{\bar{z}^2} s_{hzzi} + Op\,(n^{-3})$$

$$= \hat{\theta} + Op\,(n^{-2}) \qquad -(5.37)$$

Hence,

$$(\hat{\theta}^h - \hat{\theta})^2 = Op\,(n^{-4})$$

So $\nu_j^{(2)}(\hat{\theta}) = \nu_j^{(1)} + Op\,(n^{-3}) \qquad -(5.38)$

Similarly

$$\nu_j^{(3)} = \nu_j^{(1)} + Op\,(n^{-3}) \qquad -(5.39)$$

$$\nu_j^{(4)} = \nu_j^{(1)} + Op\,(n^{-3}) \qquad -(5.40)$$

Now for any

$$\tilde{\theta} = \hat{\theta} - \hat{B}$$

Define $\tilde{\nu}_j$ as an estimator of $\nu_j^{(2)}$ in equation (5.30) with $\hat{\theta}$ is replaced by $\tilde{\theta}$ we get

$$\tilde{\nu}_j = \nu_j^{(2)} + 2\hat{B} \sum_h (\hat{\theta} - \hat{\theta}^h) + \hat{B}^2 \sum_h (n_h - 1)^{-1}$$

because

$$\tilde{\nu}_j = \sum_{h=1}^{L} \frac{1}{n_h (n_h-1)} \sum_{i=1}^{n_h} \left\{ (\tilde{\theta}^{hi} - \hat{\theta})^2 + \hat{B}_L^2 + 2\hat{B} (\tilde{\theta}^{hi} - \hat{\theta}) \right\} \sum (n_h - 1)^{-1}$$

$$= \sum_{h=1}^{L} \frac{1}{n_h (n_h-1)} \sum_{i=1}^{n_h} (\hat{\theta}^{hi} - \hat{\theta})^2 + \hat{B}^2 \sum_h (n_h - 1)^{-1}$$

$$+ 2\hat{B} \sum_{h=1}^{L} \frac{1}{n_h (n_h-1)} \sum_{i=1}^{n_h} (\hat{\theta}^{hi} - \hat{\theta})$$

$$= \nu_j^{(2)} + 2\hat{B} \sum_{h=1}^{L} (\hat{\theta} - \tilde{\theta}^h) + \hat{B}^2 \sum_h (n_h - 1)^{-1}$$

From (5.37) we have

$$\tilde{\nu}_j = \nu_j^{(2)} + Op\ (n^{-3}) \qquad -(5.41)$$

This imply that

$$\nu_j^{(5)} = \nu_j^{(2)} + Op\ (n^{-3}) \qquad -(5.42)$$

$$\nu_j^{(6)} = \nu_j^{(2)} + Op\ (n^{-3}) \qquad -(5.43)$$

(2) B R R Estimators of θ :

Mc carthy (1966, 1969) proposed the B R R method when $n_h = 2$ for all h, based on a number of half-samples formed by deleting one P S U from the sample in each stratum. The set of R- balanced half -samples used may be defined by an R x L design matrix $(\delta h)1 \le r \le R$ and $1 \le h \le L$, where δh^r is + 1 or -1 depending upon whether the first or second sample P S U in

the h-th stratum is in the r-th half-sample and $\sum_r \delta_h^r = 0$ for all h≠h'. Define

$$z_{ih}^{(r)} = z_{ih1} \quad \text{if } \delta_h^r = 1$$
$$= z_{ih2} \quad \text{if } \delta_h^r = -1$$

Similarly,

$$z_h^{(r)} = z_{h1} \quad \text{if } \delta_h^r = 1$$
$$= z_{h2} \quad \text{if } \delta_h^r = -1$$
$$z_{ih}^{(r)} = \bar{z}_{ih} + \delta_h^r (\Delta z_{ih})$$

where $\Delta z_{ih} = \frac{1}{2}(z_{ih1} - z_{ih2})$

So,

$$z_{ih}^{(r)} = \frac{z_{ih1} + z_{ih2}}{2} + \delta_h^r (\Delta z_{ih})$$

so if $\delta_h^r = 1$, $z_{ih}^{(r)} = z_{ih1}$ and if $\delta_h^r = -1$

then $z_{ih}^{(r)} = z_{ih2}$. Similarly we can define for $z_h^{(r)}$ as

$$z_h^{(r)} = \bar{z}_h + \delta_h^r (\Delta z_h)$$

where $\Delta z_h = \frac{1}{2}(z_{h1} - z_{h2})$

Let $\hat{\bar{Z}}_i^{(r)} = M \bar{z}_i^{(r)}$

where
$$\bar{z}_i^{(r)} = \sum_{h=1}^{L} W_h z_{ih}^{(r)} = \sum_{h=1}^{L} W_h \left\{ \bar{z}_{ih} + \delta_h^r (\Delta z_{ih}) \right\}$$
$$= \sum_{h=1}^{L} W_h \bar{z}_{ih} + \sum_{h=1}^{L} W_h \delta_h^r \Delta z_{ih}$$
$$= \bar{z}_i + \sum_{h=1}^{L} W_h \delta_h^r \Delta z_{ih} \qquad -(5.44)$$

Similarly we can find out

$$\bar{z}^{(r)} = \bar{z} + \sum_{h=1}^{L} W_h \delta_h^r \Delta z_h \qquad -(5.45)$$

The one B R R estimator of θ is given as

$$\hat{\theta}_B = \frac{1}{R}\sum_{j=1}^{R} \hat{\theta}^{(r)} \qquad -(5.46)$$

where $\hat{\theta}^{(r)} = \dfrac{\bar{z}_i^{(r)}}{\bar{z}^{(r)}}$

By expanding $\hat{\theta}^{(r)}$ we have

$$\hat{\theta}^{(r)} = \theta - \frac{\bar{Z}_i}{\bar{Z}^2}(\bar{z}^{(r)} - \bar{Z}) + \frac{1}{\bar{Z}}(\bar{z}^{(r)} - \bar{Z}_i) + \frac{\bar{Z}_i}{\bar{Z}^3}(\bar{z}^{(r)} - \bar{Z})^2 - \frac{1}{\bar{Z}^2}(\bar{z}_i^{(r)} - \bar{Z}_i)(\bar{z}^{(r)} - \bar{Z}) \qquad -(5.47)$$

Assuming the regularity conditions

$$\text{Cov}\left[\sum_{h=1}^{L} W_h \delta_h^r \Delta z_{ih}, \sum_{h=1}^{L} W_h \delta_h^r \Delta z_h\right]$$

$$= \frac{1}{2}\sum_{h=1}^{L} W_h^2 S_{hzzi} = O(n^{-1}) \qquad -(5.48)$$

where $S_{hzzi} = E\left[(z_{ihl} - \bar{Z}_{ih})(z_{hl} - \bar{Z}_h)\right]$

Firstly with the help of (5.19) we can write

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta$$

$$= \frac{\bar{Z}_i}{\bar{Z}^3} E(\bar{z} - \bar{Z})^2 - \frac{1}{\bar{Z}^2} E(\bar{z}_i - \bar{Z}_i)(\bar{z} - \bar{Z})$$

$$= \frac{\bar{Z}_i}{\bar{Z}^3} E\sum_{h=1}^{L} W_h^2 (\bar{z}_h - \bar{Z}_h)^2 - \frac{1}{\bar{Z}^2}\sum_{h=1}^{L} W_h^2 E(\bar{z}_{ih} - \bar{Z}_{ih})(\bar{z}_h - \bar{Z}_h)$$

$$= \frac{\bar{Z}_i}{\bar{Z}^3}\sum_{h=1}^{L}\frac{W_h^2}{n_h} S_{hz}^2 - \frac{1}{\bar{Z}^2}\sum_{h=1}^{L}\frac{W_h^2}{n_h} S_{hzzi} \qquad -(5.49)$$

Now, with the help of (35) and (37) we under assumption (4) we get

$$B(\hat{\theta}^{(r)}) = E(\hat{\theta}^{(r)}) - \theta$$

$$= 2B(\hat{\theta}) + O(n^{-2}) \qquad -(5.50)$$

This immediately suggests a bias reducing BRR estimator

$$\hat{\theta}_B = \hat{\theta} - (\hat{\theta}_B - \hat{\theta}) = 2\hat{\theta} - \hat{\theta}_B \qquad -(5.51)$$

With B $(\hat{\theta}_B) = O(n^{-2})$

The stability of $\tilde{\theta}_B$ can be further improved by using

$$\hat{\theta}^*_B = \frac{1}{2R} \sum_{r=1}^{R} (\hat{\theta}^{(r)} - \hat{\theta}^{(r)}_c) \qquad -(5.52)$$

in the place of $\hat{\theta}_B$ in (5.51), where $\hat{\theta}^{(r)}_c$ is based on complement of r-th sample.

For $n_h = 2$ the BRR variance estimators are given by

$$\nu^{(1)}_B(\hat{\theta}) = \frac{1}{R} \sum_{r} (\hat{\theta}^{(r)} - \hat{\theta})^2 = \nu_{BRR-H} \qquad -(5.53)$$

$$\nu^{(2)}_B(\hat{\theta}) = \frac{1}{2R} \sum_{r=1}^{R} (\hat{\theta}^{(r)} - \hat{\theta}^{(r)}_c)^2 = \nu_{BRR-D} \qquad -(5.54)$$

$$\nu^{(3)}_B(\hat{\theta}) = \frac{1}{2R} \sum_{r=1}^{R} \left[(\hat{\theta}^{(r)} - \hat{\theta})^2 + (\hat{\theta}^{(r)}_c - \hat{\theta})^2 \right] \qquad -(5.55)$$

$$= \nu_{BRR-S} = \frac{1}{2} (\nu_{BRR-H} + \nu_{BRR-C})$$

where, $\hat{\theta}^{(r)}_c$ is the estimator based on the complement of the r-th half-sample and $\nu_{BRR-C} = \frac{1}{R} \sum_{r=1}^{R} (\hat{\theta}^{(r)}_c - \hat{\theta})^2$

To compare BRR-Variance estimators we get general smooth functions g(.) than (5.47) by noting from (5.44) and (5.45) that

$$\bar{z}^{(r)}_i = \bar{z}_i + \sum_{h=1}^{L} W_h \delta^r_h \Delta z_{ih}$$

and

$$\bar{z}^{(r)} = \bar{z} + \sum_{h=1}^{L} W_h \delta^r_h \Delta z_h$$

Also, let $(\bar{z}^{(r)}_i - \bar{z}_i) = \Delta \bar{z}^{(r)}_i$

and $(\bar{z}^{(r)} - \bar{z}) = \Delta \bar{z}$

With the help of above equations we get

$$\hat{\theta}^{(r)} = \frac{\bar{z}^{(r)}_i}{\bar{z}} = \hat{\theta} - \frac{\bar{z}_i}{\bar{z}^2} \Delta \bar{z}^{(r)} + \frac{1}{\bar{z}} \Delta \bar{z}^{(r)}_i$$

$$+\frac{\bar{z}_i}{\bar{z}^3}\left[\Delta\,\bar{z}^{(r)}\right]^2-\frac{1}{\bar{z}^2}\Delta\,\bar{z}^{(r)}\,\Delta\,\bar{z}_i^{(r)}$$

$$+\frac{1}{31}\left[-\frac{6\bar{z}_i}{\bar{z}^4}(\Delta\,\bar{z}^{(r)})^3+\frac{2}{\bar{z}^3}\left\{(\Delta\,\bar{z}^{(r)})^2(\Delta\bar{z}_i^{(r)})\right.\right.$$

$$\left.\left.+(\Delta\,\bar{z}^{(r)})^2\,(\Delta\,\bar{z}_i^{(r)})+(\Delta\,\bar{z}^{(r)})^2\,(\Delta\,\bar{z}^{(r)})\right\}\right]$$

$$\hat{\theta}^{(r)}=\hat{\theta}-\frac{\bar{z}_i}{\bar{z}^2}\Delta\,\bar{z}^{(r)}+\frac{1}{\bar{z}}\Delta\,\bar{z}_i^{(r)}+\frac{\bar{z}_i}{\bar{z}^3}\left[\Delta\,\bar{z}^{(r)}\right]^2-\frac{1}{\bar{z}^2}\Delta\,\bar{z}^{(r)}\Delta\bar{z}_i^{(r)}$$

$$-\frac{\bar{z}_i}{\bar{z}^4}(\Delta\,\bar{z}^{(r)})^3+\frac{1}{\bar{z}^3}(\Delta\,\bar{z}^{(r)})^2\,(\Delta\,\bar{z}_i^{(r)}) \qquad -(5.57)$$

Hence, noting that $\Delta\,\bar{z}^{(r)}=-\Delta\,\bar{z}_c^{(r)}$ and $\Delta\,\bar{z}_i^{(r)}=-\Delta\,\bar{z}_{ic}^{(r)}$ we can write from equation (5.57) as

$$\frac{1}{2}(\hat{\theta}^{(r)}-\hat{\theta}_c^{(r)})=\frac{-\bar{z}_i}{\bar{z}^2}\Delta\,\bar{z}^{(r)}+\frac{1}{\bar{z}}\Delta\,\bar{z}_i^{(r)}-\frac{\bar{z}_i}{\bar{z}^4}(\Delta\,\bar{z}^{(r)})^3$$

$$+\frac{1}{\bar{z}^3}(\Delta\,\bar{z}^{(r)})^2\,(\Delta\,\bar{z}_i^{(r)})$$

But from equation (5.54) we know that

$$\nu_{BRR\text{-}D}=\frac{1}{4R}\sum_{r=1}^{R}(\hat{\theta}^{(r)}-\hat{\theta}_c^{(r)})$$

$$=\frac{1}{R}\sum_{r=1}^{R}\left[\frac{z_i^{-2}}{\bar{z}^4}(\Delta\,\bar{z}^{(r)})^2+\frac{1}{\bar{z}^2}(\Delta\,\bar{z}_i^{(r)})^2\right.$$

$$-\frac{2\bar{z}_i}{\bar{z}^3}\Delta\,\bar{z}^{(r)}\,\Delta\,\bar{z}_i^{(r)}+\frac{\bar{z}_i^2}{\bar{z}^6}(\Delta\,\bar{z}^{(r)})^4$$

$$-\frac{z_i}{\bar{z}^5}(\Delta\,\bar{z}^{(r)})^3\,(\Delta\,\bar{z}_i^{(r)})$$

$$-\frac{z_i}{\bar{z}^4}(\Delta\,\bar{z}^{(r)})^3\,(\Delta\,\bar{z}_i^{(r)})$$

$$\left.+\frac{1}{\bar{z}^4}(\Delta\,\bar{z}^{(r)})^2\,(\Delta\,\bar{z}_i^{(r)})^2\right]$$

$$+Op\,(n^{-2.5}) \qquad -(5.58)$$

With the help of equation (5.20) we can rewrite this equation as follows

$$\nu_{BRR\text{-}D} = \nu_L(\hat{\theta}) + B + Op\,(n^{-2.5})$$

$$= \nu_L(\hat{\theta}) + Op\,(n^{-2}) \qquad -(5.59)$$

where

$$B = \frac{1}{R}\sum_{r=1}^{R}\left[\frac{z_i^{-2}}{\bar{z}^6}(\Delta\bar{z}^{(r)})^4 - \frac{\bar{z}_i}{\bar{z}^5}(\Delta\bar{z}^{(r)})^3(\Delta\bar{z}_i^{(r)}) - \frac{\bar{z}_i}{\bar{z}^4}(\Delta\bar{z}^{(r)})^3 * (\Delta\bar{z}_i^{(r)}) + \frac{1}{\bar{z}^4}(\Delta\bar{z}^{(r)})^2(\Delta\bar{z}^{(r)})^2\right]$$

$$= Op\,(n^{-2}) \qquad -(5.60)$$

Similarly we can find

$$\nu_B^{(1)}(\hat{\theta}) = \nu_{BRR\text{-}H} = \frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}^{(r)} - \hat{\theta})^2$$

with the help of equation (5.47) we have

$$\nu_B^{(1)}(\hat{\theta}) = \nu_L(\hat{\theta}) + A + B + C + Op\,(n^{-2.5})$$

$$\nu_L(\hat{\theta}) + Op\,(n^{-1.5}) \qquad -(5.61)$$

where

$$A = \frac{1}{R}\sum_{r=1}^{R}\left[-\frac{2\bar{z}_i^2}{\bar{z}^5}(\Delta\bar{z}^{(r)})^3 + \frac{2\bar{z}_i}{\bar{z}^4}(\Delta\bar{z}^{(r)})^2(\Delta\bar{z}_i^{(r)}) + 2\frac{\bar{z}_i}{\bar{z}^4}(\Delta\bar{z}^{(r)})^2(\Delta\bar{z}_i^{(r)}) - \frac{2}{\bar{z}^3}(\Delta\bar{z}^{(r)})(\Delta\bar{z}_i^{(r)})^2\right]$$

$$C = \frac{1}{4R}\sum_{r=1}^{R}\left[\frac{4\bar{z}_i^2}{\bar{z}^6}(\Delta\bar{z}^{(r)})^4 - \frac{8\bar{z}_i}{\bar{z}^5}(\Delta\bar{z}^{(r)})^3(\Delta\bar{z}_i^{(r)}) + \frac{4}{\bar{z}^4}(\Delta\bar{z}^{(r)})^2(\Delta\bar{z}_i^{(r)})^2\right] \qquad -(5.63)$$

Finally,

$$\nu_{BRR\text{-}S} = \frac{1}{2}\left(\nu_{BRR\text{-}H} + \nu_{BRR\text{-}C}\right)$$

$$= \nu_L + B + C + Op\,(n^{-2.5})$$

$$= \nu_{BRR\text{-}D} + C + Op\,(n^{-2.5})$$

$$= \nu_L + Op\,(n^{-2}) \qquad -(5.64)$$

$$\left[\because \nu_{BRR\text{-}C} = \nu_L(\hat{\theta}) - A + B + C + Op\,(n^{-2})\right]$$

To study the bias behaviour of BRR Variance estimators, we need the following results, the proofs of which are given below :

$$(I)\quad E(A) = O\,(n^{-3}) \qquad -(5.65)$$

Proof : As we know that

$$A = \frac{1}{R}\sum_{r=1}^{R}\left[\frac{-2\bar{z}_i^2}{\bar{z}^5}\,(\Delta\bar{z}^{(r)})^3 + \frac{2\bar{z}_i}{\bar{z}^4}\,(\Delta\bar{z}^{(r)})^2\,(\Delta\bar{z}_i^{(r)}) + 2\,\frac{\bar{z}_i}{\bar{z}^4}\,(\Delta\bar{z}^{(r)})^2\,(\Delta\bar{z}_i^{(r)}) - \frac{2}{\bar{z}^3}\,(\Delta\bar{z}^{(r)})\,(\Delta\bar{z}_i^{(r)})^2\right]$$

Now, as we know that if we expand $\frac{\bar{z}_i^2}{\bar{z}^5}$, $\frac{\bar{z}_i}{\bar{z}^4}$, and $\frac{1}{\bar{z}^3}$ about $\frac{\bar{Z}_i^2}{\bar{Z}^5}$, $\frac{\bar{Z}_i}{\bar{Z}^4}$ and $\frac{1}{\bar{Z}^4}$ respectively and take expectation, then rest of the terms reduce to zero and A becomes

$$E(A) = \frac{1}{R}\sum_{r=1}^{R}\left[-\frac{2\bar{Z}_i^2}{\bar{Z}^5}\,E\,(\Delta\bar{z}^{(r)})^3 + \frac{4\bar{Z}_i}{\bar{Z}^4}\,E\,(\Delta\bar{z}^{(r)})^2\,(\Delta\bar{z}_i^{(r)}) - \frac{2}{\bar{Z}^4}\,E\,(\Delta\bar{z}^{(r)})\,(\Delta\bar{z}_i^{(r)})^2\right]$$

$$- \frac{\bar{Z}_i}{\bar{Z}^2}\left\{-\frac{6\bar{Z}_i}{\bar{Z}^4}\,(\bar{z}-\bar{Z}) + \frac{2}{\bar{Z}^3}\,(\bar{z}_i-\bar{Z}_i)\right.$$

$$+\frac{2}{\bar{Z}^3}(\bar{z}-\bar{Z})+\frac{2}{\bar{Z}^3}(\bar{z}-\bar{Z})\Big\}+\frac{1}{Z}\Big\{\frac{-6\bar{Z}_i}{\bar{Z}^4}$$

$$*(\bar{z}-\bar{Z})+\frac{2}{\bar{Z}^3}(\bar{z}_i-\bar{Z}_i)+\frac{2}{\bar{Z}^3}(\bar{z}-\bar{Z})+\frac{2}{\bar{Z}^3}(\bar{z}-\bar{Z})\Big\}\Big]$$

$$*\Big[\;E(\Delta\bar{z}^{(r)})^3+E(\Delta\bar{z}^{(r)})^2(\Delta\bar{z}_i^{(r)})+(\Delta\bar{z}^{(r)})(\Delta\bar{z}_i^{(r)})^2\;\Big]$$

And as we known that $\Delta\bar{z}_i^{(r)}=\Delta\bar{z}^{(r)}=Op\ (n^{-1/2})$ so we only need to prove that

$$E(\Delta\bar{z}^{(r)})^3=E(\Delta\bar{z}^{(r)})^2(\Delta\bar{z}_i^{(r)})=E(\Delta\bar{z}^{(r)})(\Delta\bar{z}_i^{(r)})^2=0 \qquad -(IA)$$

and

$$E\Big[(\bar{z}-\bar{Z})+(\bar{z}_i-\bar{Z}_i)\Big]\Big[(\Delta\bar{z}^{(r)})^3+(\Delta\bar{z}^{(r)})^2(\Delta\bar{z}_i^{(r)})+(\Delta\bar{z}^{(r)})(\Delta\bar{z}_i^{(r)})^2\Big]=0 \qquad -(IB)$$

Therefore,

$$\sum_r(\Delta\bar{z}^{(r)})^3=\Big[\sum_{h=1}^{L}W_h^3\,\delta_h^+(\Delta z_h)^3+\sum_h\sum_{h'}W_h^2W_{h'}(\Delta z_h)^2(\Delta z_{h'})$$

$$+\sum_h\sum_{h'}\sum_{h''}W_hW_{h'}W_{h''}\Big(\sum_r\delta_h^r\delta_{h'}^r\delta_{h''}^r\Big)(\Delta z_h)(\Delta z_{h'})(\Delta z_{h''})$$

where $\delta_h^+=\sum_{r=1}\delta_h^r$. The expectation of second and third terms are zero because of independence in stratas. Now we have to prove only that

$$E(\Delta z_h)^3=0$$

Now as we know that

$$\Delta z_h=\frac{1}{2}(z_{h1}-z_{h2})$$

Let $d_{ht}=(z_{ht}-\bar{z}_h)$

and $d_{iht}=(z_{iht}-\bar{z}_{ih})$

so $\Delta z_h=\frac{1}{2}(d_{h1}-\bar{Z}_h-d_{h2}-\bar{Z}_h)$

$$=\frac{1}{2}(d_{h1}-d_{h2})$$

Similarly, we get $\Delta\ z_{ih} = \frac{1}{2}(d_{ih1} - d_{ih2})$

It follows from the above

$$E\ (\Delta\ z_h)^3 = \frac{1}{8} E\ (d_{h1} - d_{h2})^3$$

so $E\ (d_{h1}^3 - d_{h2}^3) = 0$

since different strata are independent of each other.

Similarly we can prove that

$E\ (\Delta\ \bar{z}^{(r)})^2\ (\Delta\ z_i^{(r)}) = E\ (\Delta\ \bar{z}^{(r)})\ (\Delta\ \bar{z}_i^{(r)})^2 = 0$

For the second term (IB) consider

$$E\ (\bar{z}_h - \bar{Z}_h)\ (\Delta\ \bar{z}^{(r)})^3 = \frac{1}{16} E\left[(d_{h1} + d_{h2})(d_{h1} - d_{h2})^3\right] = 0$$

and $E\left[(z_{h'} - \bar{Z}_{h'})\ \Delta z_{h'1} = \frac{1}{4} E\left[(d_{h'1} + d_{h'2})\ (d_{h'1} - d_{h2})\right] = 0\right.$

Similarly because of independence of defferent strata other terms are equal to zero. Putting all the results in E(A) we get

$$E\ (A) = O(n^{-3})$$

(II) Now, we have to prove that

$$E\left[\frac{1}{R}\sum_{r=1}^{R}(\Delta\ \bar{z}^{(r)})^4\right] = 3\left[\sum_{h=1}^{L}\frac{W_h^2}{2}S_{hz}^2\right]^2 + O\ (n^{-3})$$

$$= D_{zzzz} + O(n^{-3})\ \text{(say)} \qquad -(5.66)$$

Proof : We can prove it from the independence in different strata and $E\ (\Delta\ z_{ih}) = E(\Delta\ z_h) = 0$ the left hand side of the equation (5.66)is equal to

$$\sum_h W_h^4\ (\Delta\ z_h)^4 + 3\sum_{h}\sum_{h'} W_h^2\ W_{h'}^2\left\{E(\Delta z_h)^2\ E(\Delta\ z_{h'})^2\right\}$$

$$= O(n^{-3}) + 3\left[\sum_{h=1}^{L}\frac{W_h^2}{2}S_{hz}^2\right]^2$$

$$= D_{zzzz} + O(n^{-3})$$

Similarly, we can prove the following terms

$$E\left[\frac{1}{R}\sum_{r=1}^{R}(\Delta\bar{z}^{(r)})^3(\Delta\bar{z}_i^{(r)})\right] = D_{zzzz_i} + O(n^{-3})$$

$$E\left[\frac{1}{R}\sum_{r=1}^{R}(\Delta\bar{z}^{(r)})^2(\Delta\bar{z}_i^{(r)})^2\right] = D_{zzz_iz_i} + O(n^{-3})$$

and

$$E\left[\frac{1}{R}\sum_{r=1}^{R}(\Delta\bar{z}^{(r)})(\Delta\bar{z}_i^{(r)})^3\right] = D_{zz_iz_iz_i} + O(n^{-3})$$

From the equation (5.65)and (5.66) and under regularity conditions from (5.58) we have the bias of $\nu_{BRR\text{-}D}$ as

$$B(\nu_{BRR\text{-}D}) = B(\nu_L) + \frac{\bar{Z}_i^2}{\bar{Z}_1^6} D_{zzzz} - \frac{2\bar{Z}_i}{\bar{Z}^5} D_{zzzz_i} - \frac{}{\bar{Z}^4} D_{zzz_iz_i} + O(n^{-3}) \qquad -(5.67)$$

Where

$$D_{zzzz_i} = 3\left[\sum_{h=1}^{L}\frac{W_h^2}{1}S_{hz}^2\right]\left[\sum_{h=1}^{L}\frac{W_h^2}{2}S_{hzz_i}\right]$$

$$D_{zzz_iz_i} = \left[\sum_{h=1}^{L}\frac{W_h^2}{2}S_{hz}^2\right]\left[\sum_{h=1}^{L}\frac{W_h^2}{2}S_{hz_i}^2\right] + 2\left[\sum_{h=1}^{L}\frac{W_h^2}{2}\delta_{hzz_i}\right]^2$$

$$D_{zz_iz_iz_i} = 3\left[\sum_{h=1}^{L}\frac{W_h^2}{2}S_{hzz_i}\right]\left[\sum_{h=1}^{L}\frac{W_h^2}{2}S_{hz_i}^2\right]$$

Again from (5.61) we have

$$B(\nu_{BRR\text{-}H}) = B(\nu_{BRR\text{-}D}) + \frac{1}{R}\sum_{r=1}^{R}\frac{\bar{Z}_i^2}{\bar{Z}^6}(\Delta\bar{z}^{(r)})^4 - \frac{2\bar{Z}_i}{\bar{Z}^5}(\Delta\bar{z}^{(r)})^3(\Delta\bar{z}_i^{(r)}) + \frac{1}{\bar{Z}^4}(\Delta\bar{z}^{(r)})^2(\Delta\bar{z}_i^{(r)})^2 + O(n^{-3}) \qquad -(5.68)$$

In addition from 5.61 we have

$$B(\nu_{BRR\text{-}S}) = B(\nu_{BRR\text{-}H}) + O(n^{-3}) \qquad -(5.69)$$

$$\therefore \nu_{BRR\text{-}C} = \nu_L(\hat{\theta}) - A + B + C + Op(n^{-2})$$

But we can write (5.68) as

$$B(\nu_{BRR\text{-}H}) = B(\nu_{BRR\text{-}D}) + \frac{1}{R}\sum_{r=1}^{R}\left[\frac{\bar{Z}_i}{\bar{Z}^3}(\Delta\bar{z}^{(r)})^2 - \frac{1}{\bar{Z}^2}(\Delta\bar{z}^{(r)})(\Delta\bar{z}_i^{(r)})\right]^2 + O(n^{-3})$$

which imply

$$B(\nu_{BRR\text{-}S}) = (\nu_{BRR\text{-}H}) > B(\nu_{BRR\text{-}D}).$$

Bias of BRR Variance Estimators :

The bias of BRR variance estimators was found for $n_h = 2$.

From equation (5.58) $\nu_{BRR\text{-}D}$ can be written as

$$\nu_{BRR\text{-}D} = \frac{\bar{Z}_i^2}{\bar{Z}^4}\sum_{h=1}^{L}\frac{W_h^2}{2}s_{hz}^2 + \frac{1}{\bar{Z}^2}\sum_{h=1}^{L}\frac{W_h^2}{2}s_{hz_i}^2 - \frac{\bar{Z}\,z_i}{\bar{Z}^3}\sum_{h=1}^{L}\frac{W_h}{2}s_{hzz_i} + \frac{2}{\bar{Z}^4}\frac{1}{R}\sum_{r=1}^{R}(\Delta\bar{e}^{(r)})^2(\Delta\bar{z}^{(r)})^2 + Op(n^{-2.5})$$

$$= \nu_L(\hat{\theta}) + \frac{2}{\bar{Z}^4}\frac{1}{R}\sum_{r=1}^{R}(\Delta\bar{e}^{(r)})^2(\Delta\bar{z}^{(r)})^2 \qquad (5.70)$$

where

$$\Delta\bar{e}^{(r)} = \sum_{h=1}^{L} W_h\,\delta_h^r\,\Delta e_h$$

$$\Delta\bar{z}^{(r)} = \sum_{h=1}^{L} W_h\,\delta_h^{(r)}\,\Delta z_h$$

$$e_h = e_{h1} - e_{h2}$$

$$\Delta z_h = z_{h1} - z_{h2}.$$

Similarly, from (5.61) we have

$$\nu_{BRR\text{-}H} = \nu_L(\hat{\theta}) - \frac{2}{\bar{Z}^3}\frac{1}{R}\sum_{r=1}^{R}(\Delta\bar{e}^{(r)})^2(\Delta\bar{z}^{(r)})$$

$$+ \frac{3}{\bar{Z}^4} \cdot \frac{1}{R} \sum_{r=1}^{R} (\Delta \bar{e}^{(r)})^2 (\Delta \bar{z}^{(r)})^2 + Op(n^{-2.5}) \qquad -(5.71)$$

$$\nu_{BRR-S} = \nu_L(\hat{\theta}) + \frac{3}{\bar{Z}^4} \frac{1}{R} \sum_{r=1}^{R} (\Delta \bar{e}^{(r)})^2 (\Delta \bar{z}^{(r)})^2 + Op(n^{-2.5}) \qquad -(5.72)$$

Let

$$a = \frac{1}{\bar{Z}^2} \sum_{h=1}^{L} \frac{W_h^2}{n_h^2} S^2_{ezh} \qquad -(5.73)$$

$$b = \frac{1}{\bar{Z}^4} \left[\sum_{h=1}^{L} \frac{W_h^2}{n_h} S^2_{xh} \right] \left[\sum_{h=1}^{L} \frac{W_h^2}{n_h} S^2_{eh} \right] \geq 0 \qquad -(5.74)$$

$$c = \frac{1}{\bar{Z}^4} \left[\sum_{h=1}^{L} \frac{W_h^2}{n_h} S_{ezh} \right]^2 \geq 0 \qquad -(5.75)$$

Hence,

$$S^2_{zh} = E(z_{ht} - \bar{Z}_h)$$

$$S^2_{eh} = E(e_{ht} - \bar{E}_h)^2$$

$$S_{ezh} = E(z_{ht} - \bar{Z}_h)(e_{ht} - \bar{E}_h)$$

$$S_{e^2 zh} = E(e_{ht} - \bar{E}_h)^2 (z_{ht} - \bar{Z}_h)$$

$$\bar{E}_h = E(e_{ht}) = 0$$

Now, we try to prove that

$$\frac{1}{\bar{Z}^4} E \left[\frac{1}{R} \sum_{r=1}^{R} (\Delta \bar{e}^{(r)})^2 (\Delta \bar{z}^{(r)})^2 \right] = b - 2c + O(n^{-3}) \qquad -(5.76)$$

Proof : From the R.H.S. of (5.76) we have

$$\frac{1}{\bar{Z}^4} \left[\sum_{h=1}^{L} \frac{W_h^2}{n_h} S^2_{zh} \right] \left[\sum_{h} \frac{W_h^2}{n_h} S^2_{eh} \right] - \frac{2}{\bar{Z}^4} \left[\sum_{h=1}^{L} \frac{W_h^2}{n_h} S_{ezh} \right]^2 + O(n^{-3})$$

$$= b - 2c + O(n^{-3})$$

since we have

$$\bar{e}_h = \bar{z}_{ih} - \bar{Z}_{ih} \left(\frac{\bar{Z}_i}{\bar{Z}} \right) (\bar{z}_h - \bar{Z}_h)$$

$$\Delta \bar{e}^{(r)} = \sum^{L} W_h \delta_h^r \Delta e_h$$

$$\Delta \bar{z}^{(r)} = \sum_{h=1}^{L} W_h \delta_h^r \Delta z_h$$

so L H S of (5.76) multiplied by $\bar{Z}^4$ gives

$$\frac{1}{R} \sum_{r=1}^{R} E \left[\sum_{h=1}^{L} W_h^2 (\Delta e_h)^2 + \sum_{h \neq h'} W_h W_{h'} \delta_h^r \delta_{h'}^r \Delta e_h \Delta e_{h'} \right] * \left[\sum_{h=1}^{L} W_h^2 (\Delta z_h)^2 + \sum_{h \neq h'} W_h W_{h'} \delta_h^r \delta_{h'}^r \Delta z_h \Delta z_{h'} \right]$$

$$= E \left[\sum_{h=1}^{L} W_h^2 (\Delta e_h)^2 \right] \left[W_h^2 (\Delta z_h)^2 \right] + 2E \left[\sum_{h \neq h'} W_h^2 W_{h'}^2 \Delta e_h \Delta z_h \Delta e_{h'} \Delta z_{h'} \right]$$

$$= \left(\sum_h \frac{W_h^2}{2} S_{eh}^2 \right) \left(\sum_h \frac{W_h^2}{2} S_{zh}^2 \right) + 2 \left(\sum \frac{W_h^2}{2} S_{ezh} \right)^2 + O(n^{-3})$$

The above derivation was obtained because of independence in different strata.

From the equation (5.22) we get

$$B [\nu_L(\hat{\theta})] = -2a + b + O(n^{-3}) \qquad -(5.77)$$

Now from (5.70), (5.71), (5.72) and (5.77) we have

$$B(\nu_{BRR-D}) = B(\nu_L) + 2(b + 2c) + O(n^{-3})$$

$$= -2a + 3b + 4c + O(n^{-3}) \qquad -(5.78)$$

and

$$B(\nu_{BRR-H}) = B(\nu_{BRR-S}) = B(\nu_L) + 3(b + 2c) + O(n^{-3})$$

$$= -2a + 4b + 6c + O(n^{-3}) \qquad -(5.79)$$

The preceeding results show that

$$B(\nu_{BRR-H}) = B(\nu_{BRR-S}) > B(\nu_{BRR-D}) > B(\nu_L)$$

Bias of Jackknife Variance Estimator :

Now, an attempt will be made to find the bias of Jackknife variance estimator. From (5.34) and (5.77) we have

$$B\left[\nu_j^{(2)}(\hat{\theta})\right] = B\left[\nu_L(\hat{\theta})\right] + \frac{2}{\bar{Y}^2}\sum_{h=1}^{L}\frac{W_h^2}{n_h^2}\cdot\frac{n_h-2}{n_h-1}S_{e^2zh} + O(n^{-2})$$

because $E(s_{e^2zh}) = \left(\frac{n_h-2}{n_h}\right)S_{e^2zh}$ for $n \geq {}_h2$

so

$$B[\nu_j^{(2)}(\hat{\theta})] = B[\nu_L(\hat{\theta})] + 2a - 2a' + O(n^{-2})$$

$$= b - 2a' + O(n^{-2}) \qquad -(5.80)$$

where

$$a' = \frac{1}{\bar{Y}^2}\sum_{h=1}^{L}\frac{W_h^2}{n_h^2}\cdot\frac{\delta_{e^2zh}}{n_h-1}$$

On the basis of this asymptotic method of comparison we can draw the following conclusions. The Jackknife estimation $\tilde{\theta}_j^{(2)}$ and $\tilde{\theta}_j^{(3)}$ of θ do not in general achieve bias reduction where as the Jackknife.Estimator $\tilde{\theta}_j^{(1)}$accomplishes bias reduction. The BRR estimation $\hat{\theta}_B$ of θ has bias twice as large as that of $\hat{\theta}$, but bias -corrected BRR estimator $\hat{\theta}_D$ accomplishes bias reduction. For stratum PSU sizes n_h=2 for all h, the BRR variance estimator ν_{BRR-D} and linearization variance estimator $\nu_L(\hat{\theta})$areidentical. The six Jackknife variance estimator $\nu_j^{(i)}(\hat{\theta})$, i=1, ----, 6, are asymptotically equal to higher order terms. Further it can be shown that for the common two PSU's per stratum design $\nu_L(\hat{\theta})$and $\nu_j^{(i)}(\hat{\theta})$are asymptotically equal to higher order terms. This result suggests that choice between ν_L and $\nu_j^{(i)}$should depend more on non-statistical considerations such as computational costs. Unlike Jackknife variance estimators, BRR variance estimators and $\nu_L(\hat{\theta})$are closer to each

other. ν_{BRR-D} and $\nu_{BRR-S,}$ however are closer to $\nu_L(\hat{\theta})$than ν_{BRR-H}. As for as biases of the variances are concerned biases of BRR Variance estimators ν_{BRR-H} and ν_{BRR-S} are equal and are greater than the bias of BRR Variance estimator ν_{BRR-D} which in turn is greater than the bias of linearization variance estimator $\nu_L(\hat{\theta})$.

ILLUSTRATION

ILLUSTRATION

6.1 INTRODUCTION :

The data for this empirical investigation was taken from the project entitled *"Pilot sample survey to evolve a suitable sampling methodology for estimation of inland fishery resources and catch in a region of Orissa"*. The survey was undertaken jointly by the Indian Agricultural Statistics Research Institute, New Delhi and The Directorate of Fisheries, Orissa. For estimation of catch of fish from fresh water resources, the survey was conducted in the rounds. In the first round, all water units in the selectd gram panchayats were enumerated and data were collected on their characteristics such as seasonality, extent of utilization for fishing purpose, pisciculture techniques followed etc; data on catch of fish from selected ponds and tanks were obtained in the second round. The district selected for estimation of fish catch from fresh water units were Cuttack, Bolangir and Sambalpur. On the basis of the available frame of sampling units from the state Directorate of Fisheries namely, the distribution of Ponds and tanks in Gram Panchayats in each selected districts. It was decided to follow two stage stratified sampling design. Hence, the districts are treated as strata and blocks within the district, ponds and tanks as primary sampling units (PSU'S) and secondary sampling units (SSU'S) respectively. Because of variations in the number of blocks among three districts, two

blocks from Bolangir, three from Sambalpur and nine from Cuttack were selected by probability proportional to size with replacement (PPSWR) sampling using number of water units in each block as sizes. Then for each selected block water units i.e., ponds and tanks were selected by simple random sampling without replacement. The number of PSU's and the number of SSU's are given in the tables VII and VIII of the appendix, which are used to calculate weights attached with each of the SSU's. Six contingency tables are formed from the data of the project mentioned above, and these are given in the appendix from table I to table VI. In these tables the number of SSU's falling in cell of the contingency table are given according to the attributes it posses, where as various categories along with codes of the variables under study is given in the table IX of appendix.

6.2 METHODS OF COMPUTATION :

A special class of statistical technique called log-linear models has been formulated for the analysis of categorical data. (Haberman, 1978; Bishop, Feinberg and Holland, 1975). These models are useful for uncovering the potentially complex relationships among the variables in a multidimensional contingency table, In log-linear models, all variables that are used for classification are independent variable and dependent variable is the number of cases in a cell of cross-tabulation. In this chapter log-linear models for

two-dimensional tables are used for simplicity to describe the general procedure for multi- dimensional contingency table. Denote row variable by X and column variable by Y, then the log-linear model for two- dimensional contingency table is written as

$$\log m_{ij} = \mu + \mu_X(i) + \mu_Y(j) + \mu_{XY}(i,j) \qquad (6.1)$$

where, m_{ij} is observed frequency for $(i,j)^{th}$ cell of the table, $\mu_X(i)$ is denoting the effect of variable X on i^{th} row of the table, similarly, $\mu_Y(j)$ is the effect of variable y on j^{th} column, $\mu_{XY}(i,j)$ is the effect of association between variable X, Y and μ is general mean. The estimate of the various parameters of the above saturated model was obtained as

$$\mu_X(i) = \log \hat{r}_{i+} - (\sum_h \log \hat{r}_{h+})/I$$

$$\mu_Y(j) = \log \hat{r}_{+j} - (\sum_h \log \hat{r}_{+h})/J$$

$$\mu_{XY}(i,j) = \log m_{ij} - \sum_j \log m_{ij}/J - \sum_i \log m_{ij}/I + \sum_i \sum_j \log m_{ij}/IJ$$

$$\mu = \log n + (\sum_h \log \hat{r}_{h+})/I + (\sum_h \log \hat{r}_{+h})/J$$

where, $i = 1,2,\ldots,I$; $j = 1,2,\ldots,J$.

The parameters $\{\mu_X(i)\}$, $\{\mu_Y(j)\}$, $\mu_{XY}(i,j)$ satisfy

$$\sum_i \mu_X(i) = \sum_j \mu_X(j) = 0$$

$$\sum_i \mu_{XY}(i,j) = \sum_j \mu_{XY}(i,j) = 0$$

For main effects of log-linear models, the parameter estimates corresponding to the first k-1 categories of the variables are given, where k is the total number of categories,

where as for the association parameters, the number of estimates are (I-1)(J-1).Standard error for the parameter estimates are also obtained . Now, consider the model for the independence i.e that does not contain all possible parameters, so it is also called unsaturated model and given as

$$\log \hat{m}_{ij} = \mu + \mu_X (i) + \mu_Y (j) \qquad (6.2)$$

where , i=1,2,---,I; j=1,2,----,J.

Here, $\hat{m}_{ij}$ is no longer the observed frequency in the (i,j)-th cell,but is now the expected frequency based on the model of independence

Let,$\hat{p}_1 = \{p_{1ij}\}$ denote the maximum-likelihood estimates of the cell proportion under unsaturated model (6.2) based on the multinomial-likelihood and $\hat{p}_2 = \{p_{2ij}\}$ denote the corresponding estimates under saturated model (61) . the the various statistics are given as

1) ODINARY CHI-SQUARE STATISTICS :

$$X^2 = n \sum_i \sum_j (p_{1ij} - p_{2ij})^2 / p_{2ij}$$

2) ORDINARY LIKELIHOOD-RATIO STATISTICS :

$$G^2 = 2n \sum \sum p_{1ij} \log (p_{1ij} / p_{2ij})$$

3) MODIFIED CHI-SQUARE STATISTICS [*Felligi*, 1980)

$$X_F^2 = X^2 / \bar{b}_F$$

$$\text{Where, } \bar{b}_F = \frac{n}{IJ} \sum_i \sum_j \frac{V_{ij}}{p_{1ij}(1 - p_{1ij})}$$

where V_{ij} is the variance of corresponding proportion p_{1ij}

4) MODIFIED CHI-SQUARE STATISTICS [*Rao*, 1981)

$$X^2_R = X^2/\bar{b}_R$$

$$\text{Where, } \bar{b}_R = \frac{n}{IJ-1} \sum_i \sum_j \frac{V_{ij}}{p_{ij}}$$

5) MODIFIED CHI-SQUARE STATISTICS [*Rao*, 1981)

$$X^2_{RD} = X^2/\bar{b}_{RD}$$

$$\text{Where, } \bar{b}_{RD} = \frac{n}{(I-1)(J-1)} \left[\sum_i \sum_j \frac{V_{iij}}{p_{ii+}p_{i+j}} - \sum_j \frac{V_{ii+}}{p_{ii+}} - \sum_j \frac{V_{i+j}}{p_{i+j}} \right]$$

6) MODIFIED SATTERTHWAITE CHI-SQUARE

Let X is design matrix of saturated model and V is Variance-covariance matrix . Also, X_1 is design matrix for unsaturated log-linear model described above . Then, let

$$\tilde{X}_2 = (I - X_1(X_1'PX_1) X_1'P)X_2$$

$$\tilde{M} = (\tilde{X}_2'P\tilde{X}_2)^{-1}(\tilde{X}_2V\tilde{X}_2)$$

$$\text{Where, } X = (X_1 : X_2)'$$

then degrees of freedom for Satterthwaite approximation was calculated as

$$\nu = [Tr\ (M^*)]^2/Tr(M^*M^*)$$

and Satterthwaite Chi-square was given by

$$X^2_S = \frac{\nu}{Tr\ (M^*)} X^2 \sim \chi^2_\nu$$

7) JACKKNIFED CHI-SQUARE :

Let each replicate is of the form $Y + W^{(h,j)}$, h=1,---,L; j=1,---,n . Let $X^2_{(1)}(Y)$ denote the value of Pearson chi-square test for evaluating the fit of $\hat{P}_1$ and $X^2_{(2)}(Y)$ the test for evaluating the fit for $\hat{p}_2$.So, the jackknifed chi-square is calculated as

$$X_j = \frac{\{X^2_{(2)}(y)\}^{1/2} - (K^+)^{1/2}}{\{V^*/\ 8\ X^2_{(2)}(y)\}^{1/2}}$$

Where, $V^* = \sum_h b \sum_j R^2_{hj}$

$K^* = \sum_h b \sum_j R_{hj}$

$R^2_{hj} = X^2_{(2)}(y + W^{(h,j)}) - X^2_{(2)}(y)$

$b_h = (n_h - 1)/n_h$

n_h = number of PSU's in h-th stratum and k^+ is k^* when latter is positive, 0, otherwise. The tabulated Jackknifed chi-square is calculated with the help of following formula

$$X_{jt} = \sqrt{2}\ [\sqrt{\chi^2_T} - \sqrt{k}]$$

where χ^2_T is tabulated ordinary chi-square at k degrees of freedom.

6.3 CALCULATION FOR LEVEL OF SIGNIFICANCE :

The level of significance for each of the above mentioned statistics were obtained with the help of following relations. The Satterthwaite approxiation treats

$$X^2_s = \frac{X^2}{\theta.(1+C^2)}$$

As χ^2_ν, a χ^2 random variable with $\nu = k/(1+C^2)$ degrees of

freedom, where $\theta = \sum\theta_i/k$, is average of eigen values of design effect matrix and c is the coefficeint of variation of eigen values of design effect matrix. Hence, the asymptotic significance level of X^2 under Ho is obtained approximately as

$$SL(X^2) = Pr\ [X^2 \geq X^2_K(\alpha)]$$
$$\tilde{=} Pr\ [X^2_\nu \geq X^2_K(\alpha)/\{\theta.(1+C^2)\}]$$

where $X^2_k(\alpha)$ is the upper α-percentage point of a χ^2 random variable with k degrees of freedom. The modified statistics are asymptotically of the form $X^2(b.) = X^2/b.$, where b. is average of the estimate of some parametric values b_i, Hence using Satterthwaite's approximation, the asymptotic significance level of $X^2(b.)$ is obtained as

$$SL\ [X^2(b.)] = Pr\ [X^2(b.) \geq X^2_K(\alpha)]$$
$$\tilde{=} Pr\ [X^2_\nu \geq \{b.X^2_K(\alpha)\}/\{\theta.(1+C^2)\}]$$

The significance levels are estimated by replacing the estimated θ_i and $\hat{b}..$ If the coefficient of variation of θ_i's is large, the modified statistics $X^2(\hat{b}.)$ may not provide satisfactory approximations. In such cases if $\hat{\theta}.$ and $\hat{c}.$ are available, the Satterthwaites correction is

$$X^2(k,\alpha) = X^2\ [X^2_K(\alpha)\}/\{\theta.(1+C^2)\chi^2_\nu(\alpha)\}]$$

Would control the size of the test

$$Pr\ [X^2(k,\alpha) \geq X^2_K(\alpha)] \tilde{=} \alpha$$

Further, the level of significance is calculated by the help of the approximation given by Johnson and Kotz. From *Johnson and Kotz* (1970), we transform X^2 variate to N (0,1) variate using

$$t = A/B$$

Where $A = F^{1/3} - (1 - 2/9k)^{1/2}$

$$B = (2/9k)^{1/2}$$

$$\Pr[X_k^2 < F] = 1 - P(t) \quad \text{for } t \geq 0$$

$$= P(-t) \quad \text{for } t < 0$$

By Hasting approximation

$$P(t) = 0.5\ (1 + c_1 t + c_2 t^2 + c_3 t^3 + c_4 t^4)^{-4}$$

$c_1 = 0.196854$, $c_2 = 0.115194$,

$c_3 = 0.000344$, $c_4 = 0.019527$

6.4 RESULTS AND DISCUSSION

Results of this empirical investigation are presented with the help of eight tables. Estimates of the parameters along with its standard error of the saturated log-linear models forvarious contingency tables from the survey data were given in table number 1 to 6. With the help of this table our log-linear models can be completely specified and the effect of various categories of a variable can be studied. The log-linear models with these parameters completely specify the contingency table under consideration and hence difference between expected number of units falling in a particular cell and the observed number of units for that particular cell is equal and consequently leads to zero chi-square and likelihood values.

Table 7 show the calculated values of various test statistics under consideration for the same set of contingency tables. It was found that the values of ordinary chi-square statistics are very high and closely followed by G^2. There are two reasons for these high values, these are (1) non-independence of sample observations because of imposed

sampling design and (2) dependence of chi-square and likelihood ratio test statistics on the sample size, which is generally high for large scale surveys. Other statistics presented in table 7 are obtained by suitably modifying the calculated values of ordinary chi-square for the case of survey sampling, hence gives reasonably low values when compared to the ordinary chi-square statistics and likelihood chi-square statistics. The above fact is true for all the contingency tables used for this study.

Table 8 shows the size of the critical region for various test statistics considered in table 7, that is, their actual rejection rates when null hypothesis is true. The tests were evaluated at the nominal 0.05 level. This table clearly indicates the inflation produced to the size of actual critical region of the ordinary chi-square test statistics due to the sampling design (i.e. non-independence. of the sampling units) in to consideration. the size of the actual critical region is as high 97.8104% when evaluated at nominal 5% level. As expected the size of the critical regions of likelihood chi-square statistics are slightly higher than ordinary chi-square for all contingency tables. Further, the actual size of the critical regions for the modifications prosposed by Fellegi (1980) i.e. X_F^2 and Rao (1981) i.e. X_R^2 are found to be very conservative at nominal level of 0.05, except for the contingency table of seasonality vs depth at the time of monsoon. This exceptional behaviour for the contingency table in respect of seasonality vs depth at the time of monsoon may

be attributed to it less degrees of freedom associated with the chi-square i.e. are, for which the approximations used may not be very accurate.

Now observing the actual size of the critical region of the Rao's (delta) statistics i.e. X^2_{RD}, it was found tha its actual critical regions are slightly higher than nominal critical region except for the case of contingency table of seasonality vs depth at the time of monsoon, the resaon may be same as for X^2_F and X^2_R.

Comparing the second order correction to the ordinary chi-square i.e. Satterthwaite approximation (1946) with Jackknifed chi-square procedure proposed by Fay (1985), it was found that Satterthwaite chi-square i.e. X^2_D are having actual size of the critical region very near to the nominal critical region i.e. 0.05 except for the above discussed contingency table, where as Jackknifed chi-square was again found to be very conservative for the nominal 5% level except for the contingency tables for seasonality vs source of water unit and utilization of water unit vs soil type. The reason for the first case may be the same as discussed previously and for latter it may be due to its higher degrees of freedom. As discussed by Fay(1985), a heuristic argument offers an explanation for this observed conservation, when one of the δ_i's i.e. eigen values of design effect matrix, say δ_1 is substantially larger than the other, numerator and denominator

of the Jackknifed. Chi-square statistics will be positively correlated, which could have effect of reducing probability of extreme values. Although, Fay(1989) in his Monte carlo design shown that Jackknife chi-square perform slightly better than Satterthwaite approximation but as pointed out by him that this superiority may be due to the fact that he used multinomial distirbution for the comparis on of these statistics, which is rarely found in case of survey sampling. Hence, with help of the results obtained in this empᵩrical investiᵕgation, it can be concluded that Satterthwaite correction to the ordinary chi-square performs better for all practical situations in case of survey sampling. Although, it is too early to conclude about these methods in general, there is strong need to investigate the methods in detail with the help of simulation experiments and empᵩrical investigations.

TABLE-1

Estimates of Parameters of log-linear model for B*E contingecy table

Parameters	Cofficient	Standard Error
$\mu_B(1)$	0.22536	0.01354
$\mu_E(1)$	-0.04939	0.01354
$\mu_{BE}(1,1)$	-0.34692	0.01354

Note : B= Seasonality
E= Depth at monsoon

TABLE-2

Estimtes of Parameters of log-linear model for B*C conningency table

Parameters	Cofficient	Standard Error
μ_B (1)	0.22950	0.06719
μ_c (1)	3.45907	0.06798
μ_c (2)	0.18886	0.08551
μ_c (3)	-0.83919	0.13027
μ_c (4)	-2.55982	0.22833
μ_{BC} (1,1)	-0.19695	0.6798
μ (1,2)	-0.18295	0.08551
μ_{BC} (1,3)	0.82671	0.13027
μ_{BC} (1,4)	0.19415	0.22833

Note : B=Sesonality

C= Source of water unit

TABLE-3

Estimates of parameters of log-linear model for C*E contingency table

Parameter	Coefficient	Standard Error
$\mu_C(1)$	3.43259	0.07633
$\mu_C(2)$	0.14462	0.09285
$\mu_C(3)$	-0.41214	0.10427
$\mu_C(4)$	-2.79241	0.27878
$\mu_E(1)$	-0.30325	0.07563
$\mu_{CE}(1,1)$	0.25547	0.07633
$\mu_{CE}(2,1)$	0.50114	0.09285
$\mu_{CE}(3,1)$	0.04142	0.10420
$\mu_{CE}(4,1)$	-0.50147	0.27878

Note : C= Source of water unit

E= Depth at monsoon

TABLE-4

Estimates of parameters of log-linear model for C*H contingency table

Parameter	Coefficient	Standard Error
$\mu_C(1)$	3.34008	0.6676
$\mu_C(2)$	0.13309	0.08692
$\mu_C(3)$	-0.43800	0.10108
$\mu_C(4)$	-2.54216	0.22789
$\mu_H(1)$	0.53521	0.06570
$\mu_{CH}(1,1)$	0.00063	0.06676
$\mu_{CH}(2,1)$	-0.14436	0.08692
$\mu_{CH}(3,1)$	-0.07286	0.10108
$\mu_{CH}(4,1)$	-0.11156	0.22789

Note : C= Source of water unit

H = Extent of silting

TABLE-5

Estimates of parameters of log-linear model for B$M contingency table

Parameter	Coefficient	Standard Error
$\mu_B(1)$	0.09822	0.02140
$\mu_M(1)$	0.81274	0.02827
$\mu_M(2)$	1.19531	0.02638
$\mu_M(3)$	-0.24757	0.04036
$\mu_M(4)$	-0.28064	0.04035
$\mu_{BM}(1,1)$	0.06109	0.02827
$\mu_{BM}(1,2)$	-0.36352	0.02638
$\mu_{BM}(1,3)$	0.53054	0.04036
$\mu_{BM}(1,4)$	0.46753	0.04035

Note : B= Seasonality

M=Utilization of water unit

TABLE-6

Estimates of parameters of log-linear model for M*G contingency table

Parameter	Coefficient	Standard Error
$\mu_M(1)$	0.80174	0.03656
$\mu_M(2)$	1.08704	0.03584
$\mu_M(3)$	-0.32579	0.05549
$\mu_M(4)$	-0.26081	0.05216
$\mu_G(1)$	0.32545	0.03733
$\mu_G(2)$	1.27491	0.03189
$\mu_G(3)$	-0.39638	0.04685
$\mu_{MG}(1,1)$	0.19872	0.05061
$\mu_{MG}(1,2)$	-0.19307	0.04418
$\mu_{MG}(1,3)$	-0.44872	0.07105
$\mu_{MG}(2,1)$	-0.09571	0.05014
$\mu_{MG}(2,2)$	0.06921	0.04204
$\mu_{MG}(2,3)$	0.31463	0.05929
$\mu_{MG}(3,1)$	-0.03228	0.07667
$\mu_{MG}(3,2)$	0.29939	0.06301
$\mu_{MG}(3,3)$	-0.05063	0.09489
$\mu_{MG}(4,1)$	-0.27475	0.7702
$\mu_{MG}(4,2)$	0.10465	0.06078
$\mu_{MG}(4,3)$	0.27241	0.08495

Note: M= Utilization of water unit
G= Soil Type

TABLE - 7

Calculated Values of Various Test Statistics

Classifying Variables	k	v	χ^2	χ^2_G	χ^2_F	χ^2_R	χ^2_{RD}	χ^2_S	$\chi^2_{(k)}$	χ^2_j
B X E	1	1	765.5170	781.9877	24.9481	24.7080	137.6786	137.6500	137.6500	8.5152
B X C	4	2	89.9809	100.5402	3.5299	3.7160	5.3696	2.4992	3.9580	2.1520
C X E	4	3	64.5577	67.6347	2.2620	2.4657	6.7211	4.7241	5.7354	1.4012
C X H	4	3	16.4902	16.6140	0.7398	0.7384	5.4156	3.9882	4.8420	1.3168
B X M	4	3	774.9172	806.0244	15.3938	15.7924	33.0471	24.7431	30.0400	5.2157
M X G	12	9	264.5854	261.7435	6.1277	6.4620	13.8095	10.9440	13.6006	0.9956

Note :- B = Seasonality
C = Source of water Unit
E = Depth at Monsoon
G = Soil Type
H = Extent of Silting
M = Utilization of water unit

TABLE - 8

Percentage Level Of Significance of Various Test Statistics

Classifying Variables	k	v	χ^2	χ^2_G	χ^2_F	χ^2_R	χ^2_{RD}	χ^2_S	χ^2_j
B X E	1	1	77.3362	77.8454	18.8666	18.8003	38.3840	38.3804	3.8613
B X C	4	2	97.8184	98.5772	4.1037	4.7264	10.8189	4.5882	7.8607
C X E	4	3	77.2579	79.1602	2.8298	2.2221	7.3498	4.7611	2.2943
C X H	4	3	26.8641	26.8932	0.8885	0.7992	6.9355	4.7611	2.1463
B X M	4	3	98.8868	98.3963	2.7183	2.8888	6.7946	4.7611	0.5762
M X G	12	9	67.6431	66.8925	3.7758	4.1495	18.1338	5.9128	8.4495

Note :- B = Seasonality
C = Source of water Unit
E = Depth at Monsoon
G = Soil Type
H = Extent of Silting
M = Utilization of water unit

SUMMARY

Frankel and Kish have drawn attention to the problems that arise in the application of standard statistical method to survey data analysis which is based on the assumption that observations are independent.

In most of the situations of survey data analysis, the assumption of independence of observations is rarely met in actual practice. Any large scale survey will involve stratified multi-stage sampling and correlations between units in the same cluster (or stratum) can have a substantial impact on the statistic under consideration. The impact on linear statistics of the sample design used in obtaining survey data is subject of much of sampling literature. Recently, considerable attention has been paid to study the impact of sampling design on non-linear statistics and to the problem of estimating at least the first two moments of such statistics. In this thesis the problem of estimating the cell proportions and the variance-Covariance matrix for contingency tables has been investigated. Also, the behaviour of the chi-square statistic computed from a complex sample survey data to test the hypothesis of goodness-of-fit, independence of attributes and homogeneity of proportions is studied. Further, alternative tests for two-way as well as multi-way cross-classification have been discussed in detail.

Chapter one of this thesis introduces the subject of categorical data analysis in survey sampling whereas chapter Two provides a brief and critical review of this subject along with its background. Chapter three is confined to the categorical data analysis of survey data from two dimensional contingency tables. In this chapter the effect of stratification and clustering on the asymptotic distributions of Pearson criterion functions for goodness-of-fit (simple hypothesis), independence of attributes and homogeneity of proportions has been investigated. It has been shown that these statistics are asymptotically distributed as weighted sum of independent χ^2, random variables, and then relate weights to familiar "deffs". The "generalized deffs" has been discussed, which is useful for studying the effect of survey design, and also for getting corrections to chi-square statistics which leads to conservative tests. These simple corrections of the chi-square statistic require only the knowledge of variance estimates (or deffs) for individual cells. Also, the correction requiring estimate of full variance-covariance matrix which is also known as Satterthwaite correction or second order correction is discussed in detail . Effect of constant design effect has been considered. Moreover, the nature of chi-square statistics under constant design effect model has been observed and it was found that in both the

cases ordinary Chi-square statistic is multiple of a constant factor which can be calculated easily.

In chapter Four, the generalization of statistical methods discussed in chapter three has been discussed for multidimensional contingency tables. The asymptotic distribution of chi-square or equivalently likelihood ratio test statistics is obtained as a weighted sum of independent χ^2, random variables under nested log-linear models. An important special case of saturated model is also investigated. A simple correction to chi-square (or likelihood) statistics is obtained, which requires only the cell "deff" and the "deffs" of collapsed tables (marginals) whenever, the likelihood equation under the multinomial sampling admits explicit solutions. Further, similar type of correction which leads to more conservative test than above is considered for the models not admitting direct solutions to the multinomial likelihood equations. The effect of constant design effect is also observed for multidimensional contingency table. Moreover, the method of Jackknifed chi-square which is based upon recomputing the chi-square tests or differences of chi-square statistics of two nested models, each compared to the saturated model for a series of replicated samples based on the sample data is also covered in this chapter.

Chapter five of thesis is related to the estimation of the parameters involved in the categorical data analysis of the survey data like, cell proportion and its variance-covariance matrix. The methods of H-T estimator, combined ratio estimator and post-stratified estimators have been considered and their estimated variances obtained. The various methods of variance estimation for complex survey like linearization, Jackknifing and BRR have been compared for the case of combined ratio estimator under stratified two-stage sampling design in which units at the first stage were selected with PPSWR and in second stage with SRSWOR. It has been shown that the Jackknife estimator $\tilde{\theta}_J^{(1)}$ accomplishes bias reduction where as in case of BRR-technique this reduction is for bias-corrected BRR estimator. The results also show that choice between variance estimator due to linearization v_L, and any variance estimator due to Jackknifing $v_J^{(i)}$, i=1,2,----,6 depends more on non-statistical considerations, such as computational costs. BRR and linearized variance estimators are close to each other while Jackknife variance estimator are smaller than these two. Variances estimated through BRR-techniques have more bias than the variance obtained through linearization.

In chapter six the theory developed in the preceeding chapters has been applied to the data of a sample survey conducted at IASRI. The data have been drawn from the *Pilot*

Sample Survey to evolve a suitable sampling methodology for estimation of in-land fisheries resources and catch in a region of Orissa. Sampling design of the survey was stratified two stage sampling where first stage units were selected with PPSWR and second stage units were selected with SRSWOR. Six cross-tabulations were obtained with the purpose of demonstrating the effect of various modifications to the ordinary Chi-square and to study the relationships among various factors like, seasonality, depth at the time of monsoon, extent of silting etc. related to the water units (i.e. ponds, tanks etc). First (k-1) parameters (i.e. all independent parameters) of the main effect and interactions of log-linear model under consideration for a particular contingency tables are estimated along with their standard errors. The values of ordinary chi-square and likelihood ratio test statistics were exceptionally high and the corresponding size of the critical region was also as high as 97.8%. This may be due to the effect of non-independence of the primary sampling units and large sample size. First order correction like X^2_F, X^2_R were found to be very consarvative whereas the size of type I error for the second order correction i.e. Satterthwaite approximation was found to be very near to the nominal size i.e. 0.05. Further, Jackknifed chi-square was also found to be conservative for most of the contingency tables under consideration. This may be due to exceptionally higher eigen values of design effect matrix. Hence, from this empirical

investigation, it can be concluded that Satterthwaite correction is found to be superior than the Jackknifed chi-square for the survey data.

REFERENCES

Agresti Alan (1990). Categorical data analysis. John Wiley & Sons.

Altham, P.M.E. (1976). Discrete variable analysis for individual grouped into families. *Biometrika*. **63**: 263-269.

Bartlett, M.S. (1935). cotingency table interations. *J.R. Statist. soc. Suppl.* **2**: 248-252.

Berkson, J. (1944). Application of the logistic function to bioassay. *J. Am. Statist. Assoc.* **39**, 357-365.

Berkson, J. (1946). Approximation of Chi-square by 'Probits' and by 'Logits' *J. Am. Statist. Assoc.* **41**: 70-74.

Berkson, J. (1953). A statistical precise and relatively simple method of estimating the bioassay with quantal response based on the logistic function. *J. Am Statist. Assoc.* **48**: 565-599.

Bhapkar, V.P. (1961). Some tests for categorical data. *Ann. Math. Statist.* **32**: 72-83.

Bhapkar, V.P. (1966). A note on the equivalence of two test criteria for hypothesis in categorical data. *J. Am. Statist. Assoc.* **61**: 228-235.

Bhapkar, V.P. (1970). Categorical data analogies of some multivariate tests. In: Essays in probability and statistics. Eds. R.C. Bose, I.M. Chakrvarti, P.C. Mahalanobis, C.R. Rao, K.J.C. Smith. pp. 85-110. Chapel Hill: University of North Carolina Press.

Bhapkar, V.P. and Koch, G.G. (1968a). Hypothesis of `no interaction' in multidimensional contigency tables. *Technometrics* **10**: 107-123.

Bhapkar, V.P. and Koch, G.G. (1968b). On the hypothesis of 'no interation' in multidimensional contigency tables. *Biometrics* **24**: 567-594.

Birch, M.W. (1963). Maximum likelihood in three-way contigency tables. *J. Rh. Statist. Soc.* **B 25**: 220-233.

Birch, M.W. (1964). The detection of partial association, I. the 2 x 2 case. *J.R. Statist. Soc.* **B26**: 313-324.

Birch, M.W. (1965). The detection of partial association, II: The general case. *J.R. Statist Soc.* B **27**: 111-124.

Bishop, Y.M.M. (1967). Multidimensional contigency tables: Cell estimates. Ph. D. Dissertation. Department of Statistics. Harvard University.

Bishop, Y.M.M. (1969). Full contingency tables, logits and split contingency tables. *Biometrics* **25**: 383-400.

Bishop, Y.M.M. (1971). Effect of collapsing smultidimensional contingency tables. Biometrics **27**: 545-562.

Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975). Discrete Multivariate Analysis: Theory and Practice. Cambridge: Massachusetts Institute of Technology Press.

Bliss, C.I. (1934a). The method of probits. *Science* **79**: 38-39.

Bliss, C.I. (1934b). The method of probits. A Correlation. *Science* **79**: 409-411.

Bliss, C.I. (1935). The calculation of the dosage mortality curve (Appendix by R.A. Fisher). *Ann. Appl. Biol.* **22**: 134-167.

Bock, R.D. (1975). Multivariate Statistical Methods in Behavioral Research. New York: Mc Graw Hill.

Bock, R.D. and Jones, L.V. (1968). The measurement and prediction of judgement and choice. Holden-Day, San Francisco.

Bradley, R.A. (1976). Science, Statistcs and paired comparisons. *Biometrics* **32**: 213-233.

Bradley, R.A. and Terry, M.E. (1952). Rank analysis of incomplete block designs I. The method of paired comparisons. *Biometrika* **39**: 324-345.

Breslow, M.E. and Day, N.E. (1975). Indirect standardization and multiplicative models for rates with reference to the age adjustment of cancer incident and relative frequency data. *J. Chron. Dis.* **28**: 289-303.

Brier, S.E. (1980). Analysis of contingency tables under cluster sampling. *Biometrika* **67**: 591-596.

Bunker, J.P., Forrest, W.H., Jr. Mosteller, F. and Vandam, L. (1969). Eds. The National Halothane Study. Report of the subcommittee on the National Halothane Study of the committee on Anesthesia, Division of Medical Science,

National Academy of Sciences- National Research Council, National Institute of Health, National Institute of General Medical Sciences. Bethesda, Mol. Washington, D.C.

Chapman, D.W. (1966). An approximate test of independence based on replications of complex sample survey design. Unpublished master's thesis, Cornell University.

Cohen, J.E. (1976). The distribution of Chi-square statistics under cluster sampling from contingency tables. *J. Am. Statist. Assoc.* **71**: 665-669.

Cox, D.R. (1970). The Analysis of Binary Data London: Methuen.

Darroch, J.N. (1962). Interactions in multi-factor contingency tables. *J.R. Statist Soc.* **B 24**: 162-166.

Darroch, J.N. and Ratcliff, D. (1972). Generalized iterative scaling for log-linear models. *Ann. Math. Statist.* **43**: 1470-1480.

Darroch, J.N. and Speed, T.P. (1979). Multiplicative and additive models and interactions. Research Report No. 49, Department of Theoretical Statistics, Institute of Mathematics, University of Aarhus, Denmark.

David, H.A. (1963). The Methods of Paired Comparisons. New York: Hafner.

Deming, W.E. and Stephan, F.F. (1940). On a least squares adjustment of sampled frequency table when the expected marginal totals are known. *Ann. Math. Statist.* **11**: 427-444.

Dyke, G.V. and Patterson, H.D. (1952). Analysis of factorial arrangements when data are proportions. *Biometrics* **8**: 1-12.

Fay, R.E. (1983). CPLX-contingency table analysis for Complex Samples, Program Documentation. Unpublished report. U.S. Bureau of the Census.

Fay, R.E. (1984). Application of linear and log linear models to data from complex survey. *Survey Meth.* **10**: No.1, 82-96.

Fay, R.E. (1985). A jackknifed chi-square test for complex samples. *J. Am. Statist. Assoc.* **80**: 148-157.

Fay, R.E. (1989). Additional evaluation of chi-square methods for complex samples. *Proc. Am. Statist. Assoc. Sur. Meth. Sec.* 680-685.

Fencher, G.T. (1860). Elemente der Psychophysick. Leipzig and Hartel.

Fellegi, I.P. (1980). Approximate tests of independence and goodness of fit based on statified multi-stage samples. *J. Am. Statist. Assoc.* **75**: 261-268.

Fienberg, S.E. (1970). An iterative procedure for estimation in contingency tables. *Ann. Math. Statist.* **41**: 907-917.

Fienberg, S.E. (1978). The Analysis of Cross-Classified Categorical data Cambridge. Massachusetts. Institute of Technology Press.

Finney, D.J. (1971). Statistical Method in Biological Assay, 2nd edition, New York: Hafner.

Fisher, R.A. (1935). Appendix to Bliss (1935a). The case of zero survivers. *Ann. Appl. Biol.* **22**: 164-165.

Fleiss, J.L. (1973). Statistical Mehods for Rates and proportions: Wiley.

Forthofer, R.N. and Koch, G.G. (1973). An analysis for compounded functuions of categorical data. *Biometrics* **29**: 143-157.

Freeman, D.H. Jr. and Holford, T.R. (1980). Summary Rates. *Biometrics* **36**: 195-205.

Gaddum, J.H. (1933). Reports on Biological Standards III. Methods of biological assay depending on a quantal response. Medical Research Council Special Report Series No. 183.

Gokhale, D.V. (1971). An iterative procedure for analyzing log-linear models. *Biometrics* **27**: 681-687.

Gokhale, D.V. (1972). analysis of log-linear models. *J.R. Statist. Soc.* **B34**: 371-376.

Gokhale, D.V. and Vullback, S. (1978). The information in contingency tables. New York: Marcel Dekker.

Goldstein, M. and Dillon, W.R. (1978). Discrete Discriminent analysis. New York: Wiley.

Good, I.J. (1958). Significance tests in parallel and in series. *J. Am. Statist. Assoc.* **53**: 799-813.

Good, I.J. (1960). The interaction algorithm and practical fourier analysis. *J.R. Statist. Soc.* B22: 372-375.

Good, I.J. (1963). Maximum entripy for hypothesis fomulation especially for multi-dimensional contingency tables. *Ann. Math. Statist.* **34**: 911-934.

Good, I.J. (1965). The estimation of probabilities: An Essay on Modern Bayesian Methods. Research Monograph No. 30. Cambridge Mass, MIT Press.

Goodman, L.A. (1964). Interactions in multi-dimensional contingency tables. *Ann. Math. Statist.* **35**: 632-646.

Goodman, L.A. (1970). The multivariate analysis of quanlitative data interations among multiple classifactions. *J. Am. Statist. Assoc.* **65**: 226-256.

Goodman, L.A. and Kraskal, W.H. (1954). Measures of association for cross-classifications. *J. Am. Statist. Assoc.* **49**: 732-764.

Goodman, L.A. and Kruskal, W.H. (1959). Measures of association for cross-classifications, II: further discussion and references. *J. Am. Statist. Assoc.* **54**: 123-163.

Goodman, L.A. and Kruskal, W.H. (1963). Measures of association for cross-classifications, III: Approximate sampling theory. *J. Am. Statist. Assoc.* **58**: 310-364.

Goodman, L.A. and Kruskal, W.H. (1972). Measures of association for cross-classifications, IV: Simplification of asymptotic variances. *J. Am. Statist. Assoc.* **67**: 415-421.

Grizzle, J.E. and Allen, D.M. (1969). Analysis of growth and dose response curves. *Biometrics* **25**: 357-382.

Grizzle, J.E. and Starmer, C.F. and Koch, G.G. (1969). Analysis of categorical data by linear models. *Biometrics* **25**: 489-504.

Grizzle, J.E. and Williams, O.D. (1972). Log-linear models and tests of independence for contingency tables. *Biometrics* **28**: 137-156.

Gurland, J., Lee, I. and Dahm, P.A. (1960). Polychotomous quantal response in biological assay. *Biometrics* **16**: 382-398.

Haberman, S.J. (1972). Log-linear fit for contingency tables (Algorith As 51). *Applied Statist.* **21**: 218-225.

Haberman, S.J. (1973). Log linear models for frequency data: Sufficient Statistics and likelihood equation. *Ann. Statist.* **1**: 617-632.

Haberman, S.J. (1974a). The Analysis of Frequency Data. University of Chicago Press.

Haberman, S.J. (1974b). Log-linear models for frequency tables with ordered classifications. *Biometrics* **30**: 589-600.

Haberman, S.J. (1978). Analysis of Qualitative Data. vol.1, Introductory Topics and vol. 2, New Developments. New York. Academic Press.

Aegelmayer, F. (1852). Uber das gedachtnis fur linearan-;schauungen. Arch. Fur. Physiolo-gische Heilkunde. **11**: 844-853.

Hold, D.A., Scott, A.J. and Ewings, P.D. (1980). Chi-squared tests with survey data. *J.R. Statist. Soc.* **A143**:303-320.

Imrey, P.B., Koch, G.G. and Stokes, M.E. Categorical data analysis. Some reflection on the log linear model and logistic regression Part I. *Int. Statist. Rev.* **49**: 265-283.

Imrey, P.B., Koch, G.G. and Stokes, M.E. Categorical data analysis. Some reflection on the log-linear model and logistic regression Part II. *Int. Statist. Rev.* **50**: 35-63.

Johnson, N.L. and Kotz, S. (1970). Continuous Univariate Distributions-2: Wiley.

Killion, R.A. and Zahn, D.A. (1976). A bibliography of contingency table literature: 1900-1974. *Inst. Statist. Rev.* **44** 71-112.

Kish, L. and Frankel, M.R. (1974). Inference from complex samples. *J.R. Statist. Soc.* **36**: 1-37.

Koch, G.G., Freeman, D.H., Jr. and Freeman, J.L. (1975). Strategies in multivariate analysis of data from complex surveys. *Int. Statist. Rev.* **43**: 59-78.

Koch, G.G. and Greenberg, B.G. (1971). The growth curve model approach to the statistical analysis of large data files. Mimeo Series No. 786. capel Hill: Universtiy of North Carolina Institute of Statistics.

Koch, G.G., Imrey, P.B., Freeman, D.H., Jr. and Tolley, H.D. (1976). The asymptotic covariance structure of estimated parameters from contigency table log-linear models. In: Proc. 9th Int. Biometric Conf. Vol I pp. 317-336. Raleigh, N.C. Biometric Society.

Koch, G.G., Landis, J.R., Freeman, J.L., Freeman, D.H., Jr. and Lehnen, R.G. (1977). A general methodology for the analysis of experiments with repeated measurements of categorical data. *Biometrics* **33**: 133-158.

Koch, G.G., & Talley, H.D.(1975). A generalized modified X^2 analysis of categorical data from a complex dilution clustered attribute data *Biometrics*. **32**: 337-354.

Ku, H.H. and Kullback, S. (1968). Interaction in multidimensional contingency tables: an information theoretic approach. *J. Res. Nat. Bur. Stand.* **72B**: 159-199.

Ku, H.H., Varner, R.N. and Kullback, S. (1971). analysis of multidimensional contingency tables. *J. Am. Statist. Assoc.* **66**: 55-64.

Kumar, s. and Rao, J.N.K. (1984). Logistic regression analysis of labour force survey data. *Survey Meth.* **10**: No.1, 62-81.

Lachenbruch, P.A. (1975). Discriminant analysis. New York. Hafner.

Lancaster, H.O. (1951). complex contingency tables treated by the partition of chi-square. *J.R. Statist. Soc.* **B13**: 242-249.

Lancaster, H.O. (1957). Some properties of the bivariate normal distribution considered in the form of a contingency table. *Biometrika* **49**: 289-292.

Lancaster, H.O. (1969). The chi-squared Disttribution. New York: Wiley.

Lancaster H.O. and Hamdam, M.A. (1964). Estimation of the correlation coefficient in contingency tables with possible non-metrical character. *Psychometrika* **29**: 381-391.

Landis, J.R., Heyman, E.R. and Koch, G.G (1978).. Average partial association in three-way contingency tables. a review and discussion of alternative tests. *Int. Statist. Rev.* **46**: 237-254.

Landis, J.R. and Koch, G.G. (1977). A one-way component of variance model for categorical data. *Biometrics.* **33**: 671-679.

Lewis, B.N. (1962). On the analysis of interaction in multi-dimensional contingency tables. *J.R. Statist. Soc.* **A125**: 89-117.

Luce, R.D. (1959). Individual choice Behaviour. New York: Wiley.

Mantel, N. (1966). Models for complex contingency tables and polychtomous dose response curves. *Biometrics* **22**: 83-95.

McCarthy, P.J. (1966). Replication: An approach to the analysis of data from complex surveys. vital and Health Statistics Ser 2, No-14, Washington, D.C., U.S. Government Pringting Office.

McCarthy, P.J. (1969). Pseudoreplication. Half-samples. *Rev. Int. Statist. Inst.* **37**: 239-264.

Mosteller, F. (1968). Association and estimation in contingency tables. *N. Am Statist. Assoc.* **63**: 1-28.

Mathan, G. (1969). Tests of independence in contingency tables from stratified samples. In: M.L. Johnson and H. Smith, eds., New Developments in Survey Sampling: Wiley, New York. pp. 578-600.

Nathan, G. (1972). On the asymptotic power of tests for independence in contingency tables from stratified samples. *J. Am. Statist. Assoc.* **67**: 917-920.

Nathan, G. (1973). Approximate tests of independence in contingency tables from stratified samples. National Centre for Health Statistcs, vital and Health Statistics, Series, 2. No. 53, Washington. D.C.

Nathan, G. (1975). Test of independence in contingency tables from stratified proportional samples. Sankhya, **37**. Series c. Port I. 77-87.

Nelder, J.A. (1974). Log-linear models for contingency tables: A generalization of classical least squances. **Appl. Staitst. 23**: 323-329.

Nelder, J. and Wedderburn, R.W. (1972). Generalized linear models. *J.R. Statist. Soc. **A 135**:* 370-384.

Nerlove, M. and Press, S.J. (1973). Univariate and multivariate log-linear and logistic models. Tech. Rep. R-1306-EDA/NIH. Santa Monica, Calif: Rand Corporation.

Neyman, J. (1949). Contribution to the theory of the X^2 test. Proc. Ist Berkley Symp. 230-273.

Norton, H.W. (1945). Calculation of chi-square from complex contingency tables. *J. Am. Statist. Assoc.* **40**:251-258.

Pearson, K. (1904). Mathematical contributions to the theory of evolution XIII: On the theory of contingency and its relation to association and normal correlation. Draper's Co. Research Memoirs. Biometric Series, No. 1. (Reprinted 1948 in Karl Pearson's Early Papers, ed. by E.S. Pearson, Cambridge. Cambridge University Press).

Pearson, K. (1913). On the probable error of a correlation coefficient as found from four fold table. *Biometrika.* **9**: 22-27.

Plackett, R.L. (1974). The Analysis of categorical Data. London: Griffin.

Pothoff, R.F. and Roy, S.N. (1964). A generlized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika.* **51**: 122-127.

Rao, J.N.K. and Scott, A.J. (1979). Chi-squared tests for analysis for analysis of categorical data from complex surveys. *Proc. Am, Statist. Sec. Surr. Res. Meth.* 58-66.

Rao, J.N.K. and Scott, A.J. (1981). The analysis of categorical data from complex sample surveys. Chi-squared t3ests for goodness of fit and independenc of two-way tables. *J. Am. Statist. Assoc.* **76**: 221-230.

Rao, J.N.K. and Hidiroglou (1981). Chi-squre tests for the analysis of categorical data from the Canada Health Survey. *Bull. Int. Stist. Inst.***49**: 699-718.

Rao, J.N.K. and Scott, A.J. (1984). On, Chi-squred test for multi-way contingency tables with proportions estimated from survey data. *Ann. Statist.* **12**:46-60.

Rao, N.N.K. and Scott, A.J. (1987). On Simple adjustments to chi-square tests with sample survey data. *Ann. Staitst,* **15**: No.1: 385-397.

Roy, S.N. and Kastanbaum, M.A. (1956). On the hypothesis of no 'interaction' in multi-way contingency table. *Ann. Math. Statist.* **27**: 749-757.

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

Schmidt, P. and Strauss, R.P. (1975a). The prediction of occupation using multiple logit models. *Int. Econ. Rev.* **16**: 471-486.

Schmidt, P. and Strauss, R.P. (1975b). Estimation of models withjointly dependent qualitative variables. a simultaneous logit approach. *Econometrica* **43**: 745-756.

Scott. A.J. .and Styan, G.P.H. (1985). On generalized eigenvalues and a problem in sample survey analysis. Technical report, Carleton Univ. Ottawa. Canada.

Searle, S.R.(1971). Linear Models: Wiley.

Shryock, S.S. and Siegel, J.S. (1973). The methods and materials of demography. 2. Washington, D.C. U.S. Government Printing Office. 702-712.

Shuster, J.J. and Downing, D.J. (1976). Two-way contingency tables from complex sampling scheme. *Biometrika.* **63**: 271-276.

Singer, B. (1979). Distribution-free methods for non-parametric problems: A classified and selected bibliography. *Br. J. Math. Statist. Psychol.* **32**:.**1**.: 1-60.

Singh, A.C. and Kumar,s. (1986). Categorical data analysis for complex surveys. *Proc. Sec. Survey. Meth. Am. Statist. Assoc.* 88-96.

Solomen, H. and Stephens, M.A. (1977). Distribution of a sum of weighted chi-square variables. *J. Am,.Statist. Assoc.* **72**: 881-885.

Theil, H. (1969). A multinomial extension of the linear logit model. *Int. Econ. Rev.* **10**: 251-259.

Theil, H. (1970). On estimation of relationships involving qualitative variables. *Am. J. Sociol.* **76**. 103-154.

Thomas, D.R. and Rao, J.N.K. (1984). A Monte Carlo study of exact levels for chi-squred Goodness of fit statistcs under cluster sampling. Technical Report. 35, Carletton University Ottawa.

Thomas, D.R. and Rao, J.N.K. (1985). On the power of some Goodness-of fit tests under cluster sampling. Technical Report. 66, in analysis of categorical data from sample survey: A collection of 5 papers. Carlton University Ottawa.

Thurstone, L.L. (1927a). A law of comparitive judgment. *Psyco. Rev.* **34**: 278-286.

Thurstone, L.L. (1927b). Psychophysical analysis. *Am. J. Psychol.* **38**: 368-389.

Thurstone, L.L. (1927c). The method of paired comparisons for social values. *J. Abnormal Social Psychol.* **21**:384-400.

Thurstone, L.L. (1928). The phi-gamma hypothesis. *J. Exp. Psychol.* **11**: 293-305.

Thurstone, L.L. (1959). The measurement of values. University of Chicago Press, I 11.

Tolley, H.D. and Koch, G.G. (1974). A two-stage approach to the analysis of longitudinal type categorical data. Memeo Seris No. 962. Chapel Hill: University of North Carolina Institute of Stistics.

Truett, J. Cornfield, J. and Kannel, W. (1967). A multivariate analysis of risk of coronary heart disease in Framingham, *J. Chon. Dis.* **20**: 511-524.

Urban, F.M. (1908). The application of statistical Methods to the Problems of Psycophysics. Philadelphia. Pa: Psychological Clinic Press.

Wald, A. (1943). Tests of statistical hypothesis concerning general parameters when the number of observations is large. *Trans. Am. Math. Soc.* **54**:426-482.

Walker, S.H. and Duncan, D.B. (1976). Estimation of the probability of an event as a function of several independent variables. *Biometrika* **54**: 167-179.

Williams, W.H. (1960). The variance of an estimator with Post-stratified weighting. *J. Am. Statist Assoc.*. **57**: 622-627.

Yule, G.V. (1900). On the association of attributes in statistics: with illustration from the material of the childhood society. *Phil. Trans. R. Soc. A.* **194**: 257-319.

APPENDIX

TABLE-I

CONTIGENCY TABLE OF SEASONALITY VS DEPTH AT MONSOON (B*E)

E \ B	1	2	Total
1	996	1906	2902
2	2201	1051	3252
Total	3197	2957	6154

TABLE-II

CONTINGENCY TABLE OF SEASONALITY VS SOURCE OF WATER UNIT (B*C)

B / C	1	2	Total
1	2919	2735	5654
2	112	102	214
3	111	13	124
4	10	5	15
5	45	102	147
Total	3197	2957	6154

TABLE-III

CONTINGENCY TABLE OFDEPTH AT MONSOON VS SOURCEOF WATER UNIT (E¢C)

E / C	1	2	Total
1	2691	2963	5654
2	127	87	214
3	45	79	124
4	6	9	15
5	33	114	148
Total	2902	2352	6154

TABLE-IV

CONTINGENCY TABLE OF EXTENT OF SILTING VS SOURCE OF WATER UNIT (H‡C)

H / C	1	2	Total
1	4212	1442	5654
2	147	67	214
3	89	35	124
4	9	6	15
5	126	21	147
Total	4583	1571	6154

TABLE-V

CONTINGENCY TABLE OF SEASONALITY VS UTILIZATION OF WATER UNIT (B&M)

B / M	1	2	Total
1	1044	759	1803
2	1001	1702	2703
3	578	164	742
4	525	169	694
5	49	163	212
Total	3197	2957	6154

TABLE-VI

CONTINGENCY TABLE OF SOIL TYPE VS UTILIZATIO OF WATER UNIT (G¤M)

B / C	1	2	3	4	Total
1	550	961	134	158	1803
2	545	1662	399	97	2703
3	141	509	67	25	742
4	118	447	99	30	694
5	67	107	25	13	212
Total	1421	3686	724	323	6154

TABLE-VII

INFORMATION ABOUT SAMPLE SIZE

District	Total No. of SSU's	No. of selected SSU's
Bolangir	4733	433
Sambalpur	3593	445
Cuttack	26387	5276

TABLE-VIII

District	Block No.	No.Of SSU's Selected
1	01	152
1	02	129
2	01	178
2	02	111
2	03	158
3	01	469
3	02	554
3	03	120
3	04	799
3	05	347
3	06	460
3	07	1084
3	08	89
3	09	222

TABLE -IX

INFORMATION ABOUT VARIABLE VARIABLE UNDER STUDY

Variable (code)	Name of the categories-code
Seasonality (B)	(a) Perrenial-1 (b) Seasonal -2
Source of water unit(C)	(a) Rainfall -1 (b) River channels-2 (c) Sewage -3 (d) Coastal water-4 (e) Others - 5
Depth at monsson (E)	(a) 0-2 mets -1 (b) >2 mets -2
Soil type (G)	(a) Sandy -1 (b) Clay -2 (c) Loamy -3 (d) Others -4
Extent of silting (H)	(a) Patially silted-1 (b) Badly silted -2
Utilization of water unit (M)	(a) Fish cultivation -1 (b) Not fish cultivation -2 (c) Irrigation -3 (d) Multiple purpose -4 (e) Others-5