CHAPTER 19

Complex Surveys: Categorical Data Analysis

19.1 INTRODUCTION

Chi-square test statistics are extensively used in goodness of fit, tests of homogeneity, and independency. For example, in the Botswana Aids Impact Survey-II, importance was given to determine if the new HIV/AIDS management program yielded any improvement on the HIV/AIDS prevalence rate. In "Household Income and Expenditure Surveys," one may be interested in finding out if there is an association between income and level of education, or to test whether the income distribution remains unchanged over the last census period or not. Researchers often answer these questions by computing chi-square test statistics using SPSS, SAS, or other statistical packages for data analysis. The chi-square test statistics are valid if the sample size is large and the sample is selected by simple random sampling with replacement (SRSWR) method. But in practice complex survey designs involving stratification, clustering, and unequal probability sampling are used for collection of data. The effect of stratification, clustering, and unequal methods of selection of sample invalidates chi-square tests. So, the standard statistical packages report wrong p-values. Rao and Scott (1981) reported that the chi-square tests based on complex survey designs produced actual level of type I error as high as 40% when the target type 1 error is at the 5% level. In this chapter, we will consider the effect of unequal probability sampling, stratification, and clustering on the usual chi-square tests for goodness of fit, independency, and homogeneity. We also propose some modification of the existing chi-square tests, which yield reasonably valid tests.

19.2 PEARSONIAN CHI-SQUARE TEST FOR GOODNESS OF FIT

Suppose that a population $U = (U_1,\ldots, U_i,\ldots, U_N)$ of N units is divided into k mutually exclusive and exhaustive classes $\Omega_1,\ldots, \Omega_i,\ldots, \Omega_k$ according

Survey Sampling Theory and Applications
ISBN 978-0-12-811848-1
http://dx.doi.org/10.1016/B978-0-12-811848-1.00019-4

to some criteria. Let N_i be the number of units that belong to the ith class and the respective proportion be $p_i = N_i/N \left(0 \le p_i \le 1, \sum_{i=1}^{k} p_i = 1 \right)$. The values of p_i's are not known and we are interested to test the following hypothesis of goodness of fit:

$$H_0 : p_i = p_{i0} \quad \text{for } i = 1, \ldots, k \tag{19.2.1}$$

against the alternative H_1:$p_i \neq p_{i0}$ for at least one $i(=1,\ldots, k)$ where p_{i0}'s are known nonnegative constants with $\sum_{i=1}^{k} p_{i0} = 1$.

Suppose a sample s of size n units is selected by SRSWR method and n_i is the number of units falling in the ith class. Then under H_0, the Pearson chi-square test statistic

$$X_P^2 = n \sum_{i=1}^{k} \frac{\left(\widetilde{p}_i - p_{i0}\right)^2}{p_{i0}} \tag{19.2.2}$$

follows χ^2_{k-1}, chi-square distribution with $k-1$ degrees of freedom (df) with $\widetilde{p}_i = n_i/n$. We reject the hypothesis H_0 at 100 α% level of significance if the computed value of X_P^2 exceeds $\chi^2_{\alpha,k-1}$, the upper 100 α% point of chi-square distribution with $k-1$ df. The test is valid if the sample size n is large and the expected cell frequency ($n_{io} = np_{i0}$) for each cell exceeds 5.

The likelihood-ratio chi-squared statistic

$$G^2 = 2 \sum_{i=1}^{k} n_i \log(n_i/n_{i0}) \tag{19.2.3}$$

may be used as an alternative to Pearsonian chi-square statistic X_P^2. Under the null hypothesis H_0, G^2 follows a chi-square distribution with $k-1$ df when the sample size n is large. Convergence of chi-squared is quicker than G^2 and the approximation of G^2 to chi-square is poor if $n/k < 5$ (Agresti, 2002).

19.3 GOODNESS OF FIT FOR A GENERAL SAMPLING DESIGN

Suppose a sample s of size n is selected from the population $U(=\Omega_1 \cup \ldots \cup \Omega_k)$ by some sampling design with probability $p(s)$. Let $\pi_i(>0)$ and $\pi_{ij}(i \neq j)$ be the inclusion probabilities of the ith, ith and jth units of the population. Let s_i be the set of units that belong to the ith class

Ω_i and y_j be the value of a certain character y for the jth ($j = 1,\ldots, N$) unit. Consider an unbiased estimator for the population total Y as

$$\widehat{Y} = \sum_{j\in s} b_{sj} y_j \tag{19.3.1}$$

where b_{sj}'s are constants that satisfy the unbiasedness condition

$$\sum_{s\supset j} b_{sj} p(s) = 1 \quad \text{for } j = 1, \ldots, N$$

Using Särndal et al. (1992) notation, an unbiased estimator of $Y_i\left(= \sum_{j\in \Omega_i} y_i\right)$, the total of the y-values of the units belonging to Ω_i is given by

$$\widehat{Y}_i = \sum_{j\in s} b_{sj} y_j I_j(i) \tag{19.3.2}$$

where $I_j(i) = \begin{cases} 1 & \text{if the } j\text{th unit belong to } \Omega_i \\ 0 & \text{otherwise} \end{cases}$.

Substituting $y_j = 1$ for $j = 1,\ldots, N$ in Eq. (19.3.2) we find an unbiased estimator of $p_i = N_i/N$ as

$$\widehat{p}_i = \frac{\sum_{j\in s} b_{sj} I_j(i)}{N} \tag{19.3.3}$$

The variance of $\widehat{p}_i$ and covariance of $\widehat{p}_i$ and $\widehat{p}_{i'}$ are, respectively, given by

$$\begin{aligned} Var(\widehat{p}_i) &= \sum_{j\in U} I_j(i)\alpha_{jj} + \sum_{j\neq} \sum_{k\in U} I_j(i) I_k(i) \alpha_{jk} \\ &= \sum_{j\in \Omega_i} \alpha_{jj} + \sum_{j\neq} \sum_{k\in \Omega_i} \alpha_{jk} \\ &= V_{ii}/n \end{aligned} \tag{19.3.4}$$

and

$$\begin{aligned} Cov(\widehat{p}_i, \widehat{p}_{i'}) &= \sum_{j\in \Omega_i} \sum_{k\in \Omega_{i'}} \alpha_{jk} \\ &= V_{ii'}/n \end{aligned} \tag{19.3.5}$$

where $\alpha_{jj} = \left(\sum_{s\supset j} b_{sj}^2 p(s) - 1\right) \Big/ N^2$ and $\alpha_{jk} = \left(\sum_{s\supset j,k} b_{sj} b_{sk} p(s) - 1\right) \Big/ N^2$.

Unbiased estimators of V_{ii} and $V_{ii'}$ are given by

$$\widehat{V}_{ii}/n = \left(\sum_{j\in s_i} \frac{\alpha_{jj}}{\pi_j} + \sum_{j\neq} \sum_{k\in s_i} \frac{\alpha_{jk}}{\pi_{jk}} \right) \text{ and } \widehat{V}_{ii'}/n = \left(\sum_{j\neq s_i} \sum_{k\in s_{i'}} \frac{\alpha_{jk}}{\pi_{jk}} \right). \tag{19.3.6}$$

Example 19.3.1

In case $b_{sj} = 1/\pi_j$, we get $\widehat{p}_i = \dfrac{\sum_{j\in s} I_j(i)/\pi_j}{N}$,

$$V_{ii}/n = \left[\sum_{j\in\Omega_i} \left(\frac{1}{\pi_j} - 1 \right) + \sum_{j\neq} \sum_{k\in\Omega_i} \left(\frac{\pi_{jk}}{\pi_j\pi_k} - 1 \right) \right] \Big/ N^2,$$

$$V_{ii'}/n = \sum_{j\in\Omega_i} \sum_{j'\in\Omega_{i'}} \left(\frac{\pi_{jj'}}{\pi_j\pi_{j'}} - 1 \right) \Big/ N^2,$$

$$\widehat{V}_{ii}/n = \left[\sum_{j\in s_i} \frac{1}{\pi_j} \left(\frac{1}{\pi_j} - 1 \right) + \sum_{j\neq} \sum_{k\in s_i} \frac{1}{\pi_{jk}} \left(\frac{\pi_{jk}}{\pi_j\pi_k} - 1 \right) \right] \Big/ N^2, \text{ and}$$

$$\widehat{V}_{ii'}/n = \sum_{j\in s_i} \sum_{k\in s_{i'}} \frac{1}{\pi_{jk}} \left(\frac{\pi_{jk}}{\pi_j\pi_k} - 1 \right) \Big/ N^2.$$

Example 19.3.2

For simple random sampling without replacement (SRSWOR), $\pi_j = n/N$ and $\pi_{jk} = n(n-1)/\{N(N-1)\}$. In this case, we have from Example 19.3.1

$$\widehat{p}_i = \sum_{j\in s} I_j(i)/n = n_i/n = \widetilde{p}_i,\ V_{ii} = \frac{N-n}{(N-1)} p_i(1-p_i),\ V_{ii'} = -\frac{N-n}{(N-1)} p_i p_{i'},$$

$$\widehat{V}_{ii}/n = \frac{N-n}{N(n-1)} \widetilde{p}_i(1-\widetilde{p}_i) \text{ and } \widehat{V}_{ii'}/n = -\frac{N-n}{N(n-1)} \widetilde{p}_i\widetilde{p}_{i'}.$$

Example 19.3.3

If we choose $\widehat{p}_i = \dfrac{\sum_{j\in s} I_j(i)/\pi_j}{\sum_{j\in s} 1/\pi_j}$ and $\widehat{p}_{i'} = \dfrac{\sum_{j\in s} I_j(i')/\pi_j}{\sum_{j\in s} 1/\pi_j}$, as ratio estimators for p_i and $p_{i'}$, respectively, then the approximate consistent estimators of the variance of $\widehat{p}_i$ and covariance of $\widehat{p}_i$ and $\widehat{p}_{i'}$ are obtained from the Theorem 8.2.2.

19.3.1 Wald Statistic for Goodness of Fit

Let $\widehat{p}_i \geq 0$ for $i = 1,\ldots, k-1$, $\widehat{p}_k = 1 - \sum_{i=1}^{k-1} \widehat{p}_i \geq 0$, $\widehat{\mathbf{p}} = \left(\widehat{p}_1, \ldots, \widehat{p}_{k-1}\right)'$, $\mathbf{p} = (p_1, \ldots, p_{k-1})'$, $\mathbf{p}_0 = \left(p_{10}, \ldots, p_{k-1,0}\right)'$, $\mathbf{V}/n = (V_{ij})/n =$ variance–covariance matrix of $\widehat{\mathbf{p}}$ of rank $k-1$, and $\widehat{\mathbf{V}}/n = \left(\widehat{V}_{ij}\right)/n$ be unbiased or consistent estimator of $\mathbf{V}/n$ of rank $k-1$. The Wald statistic for goodness of fit is defined by

$$\begin{aligned} X_W^2 &= n\left(\widehat{\mathbf{p}} - \mathbf{p}_0\right)' \mathbf{V}^{-1} \left(\widehat{\mathbf{p}} - \mathbf{p}_0\right) \\ &\cong n\left(\widehat{\mathbf{p}} - \mathbf{p}_0\right)' \widehat{\mathbf{V}}^{-1} \left(\widehat{\mathbf{p}} - \mathbf{p}_0\right) \end{aligned} \tag{19.3.7}$$

For a large n, X_W^2 is distributed as a chi-square distribution with $k-1$ df $\left(\chi_{k-1}^2\right)$ when H_0 is true. To use X_W^2 in practice, one needs a consistent estimator $\widehat{\mathbf{V}}$ of $\mathbf{V}$ in addition to the requirement of a large sample size. In case an unbiased estimator of $\mathbf{V}$ is not available because of complexity of the survey design, one may estimate variance by using methods such as linearization (LR), jackknife, balanced repeated replication (BRR), and bootstrap (BT) discussed in Chapter 18. The main drawback of the Wald Statistic is that if the number of classes is large then the sample sizes for some of the classes may become too small to provide consistent estimators for variances of the respective classes (see Fay, 1985). Furthermore, survey reports very often give estimates with their standard error but rarely publish estimates of the covariances.

19.3.1.1 Simple Random Sampling With Replacement

For SRSWR sampling, $\widehat{p}_i = \widetilde{p}_i = n_i/n$ where $n_i =$ number of unit following the ith class and $\widehat{\mathbf{p}}$ follows multinomial distribution with variance–covariance matrix $\mathbf{V}/n = (V_{ij})/n$, where $V_{ii} = p_i(1-p_i)$ and $V_{ij} = -p_i p_j$. Writing $\mathbf{V}^{-1}\big|_{\mathbf{p}=\mathbf{p}_0} = \mathbf{P}_0^{-1} = \left(\delta_0^{ij}\right)$ with $\delta_0^{ii} = \left(1/p_{i,0} + 1/p_{k,0}\right)$ and $\delta_0^{ij} = 1/p_{k,0}$, we find under H_0,

$$\begin{aligned} X_W^2 &= n\left(\widehat{\mathbf{p}} - \mathbf{p}_0\right)' \mathbf{P}_0^{-1} \left(\widehat{\mathbf{p}} - \mathbf{p}_0\right) \\ &= n \sum_{i=1}^{k} \frac{\left(\widetilde{p}_i - p_{i0}\right)^2}{p_{i0}} \\ &= X_P^2 \end{aligned} \tag{19.3.8}$$

Alternatively, writing $\widehat{V}_{ii} = \widetilde{p}_i\left(1 - \widetilde{p}_i\right)$ and $\widehat{V}_{ij} = -\widetilde{p}_i \widetilde{p}_j$, we find

$$\begin{aligned} X_W^2 &= n\left(\widehat{\mathbf{p}} - \mathbf{p}_0\right)' \widehat{\mathbf{V}}^{-1} \left(\widehat{\mathbf{p}} - \mathbf{p}_0\right) \\ &= n \sum_{i=1}^{k} \frac{\left(\widetilde{p}_i - p_{i0}\right)^2}{\widetilde{p}_i} \\ &= X_N^2 \end{aligned} \tag{19.3.9}$$

The statistic X_N^2 is known as Neyman statistic.

Replacing $\widetilde{p}_i$ by a consistent estimator $\widehat{p}_i$, the expression (Eq. 19.3.9) becomes

$$X_M^2 = n \sum_{i=1}^{k} \frac{\left(\widehat{p}_i - p_{i0}\right)^2}{\widehat{p}_i} \tag{19.3.10}$$

The statistic X_M^2 is known as modified chi-square statistic. Bhapkar (1966) established equivalency of the modified chi-square and Wald statistic for linear hypothesis.

19.3.2 Generalized Pearsonian Chi-Square Statistic

The estimator $\widetilde{p}_i = n_i/n$ is not a consistent estimator of p_i unless the sampling design is self-weighing. So, replacing $\widetilde{p}_i$ by $\widehat{p}_i$, an unbiased or consistent estimator of p_i in the expression of X_P^2 given in Eq. (19.2.2), we derive the generalized Pearsonian chi-square statistic for a complex survey design as follows:

$$X_G^2 = n \sum_{i=1}^{k} \frac{\left(\widehat{p}_i - p_{i0}\right)^2}{p_{i0}} \tag{19.3.11}$$

Rao and Scott (1981) provided the asymptotic distribution of X_G^2, which is stated in the following theorem without derivation.

Theorem 19.3.1

Let $\lambda_{10},\ldots,\lambda_{k-1,0}$ be the eigenvalues of $\mathbf{D}_0 = \mathbf{P}_0^{-1}\mathbf{V}_0$, where $\mathbf{V}_0$ and $\mathbf{P}_0$ are the values of $\mathbf{V}$, and $\mathbf{P} = diag(\mathbf{p}) - \mathbf{p}\mathbf{p}'$ for $\mathbf{p} = \mathbf{p}_0$, then under the null hypothesis H_0:$\mathbf{p} = \mathbf{p}_0$, the asymptotic distribution of X_G^2 is

$$X_G^2 \sim \sum_{i=1}^{k-1} \lambda_{i0} Z_i^2 \tag{19.3.12}$$

where $Z_1,\ldots, Z_{k-1}$ are independent normal variates with mean zero and variance unity.

Hence X_G^2 is asymptotically distributed as a weighted sum of independent χ_1^2 (chi-square with 1 df) variables.

19.3.2.1 Design Effect

The design effect (deff) of a sampling design d with respect to an estimator T is defined as $V_d(T)/V_{srs}(t)$, where $V_d(T)$ is the variance of T with respect to the design d and $V_{srs}(t)$ is the variance of a comparable estimator based on an SRSWR sampling design.

Let $\lambda_1,\ldots, \lambda_{k-1}$ be the eigenvalues of $\mathbf{D} = \mathbf{P}^{-1}\mathbf{V}$, then following Rao and Scott (1981), we find for any arbitrary vector $\mathbf{c}' = (c_1, \ldots, c_{k-1})$,

$$\lambda_{\max} = \sup_c \frac{\mathbf{c}'\mathbf{V}\mathbf{c}}{\mathbf{c}'\mathbf{P}\mathbf{c}} = \sup_c \frac{V_d\left(\sum_{i=1}^{k-1} c_i \widehat{p}_i\right)}{V_{srs}\left(\sum_{i=1}^{k-1} c_i n_i/n\right)} \tag{19.3.13}$$

and

$$\lambda_{\min} = \inf_c \frac{\mathbf{c}'\mathbf{V}\mathbf{c}}{\mathbf{c}'\mathbf{P}\mathbf{c}} = \inf_c \frac{V_d\left(\sum_{i=1}^{k-1} c_i \widehat{p}_i\right)}{V_{srs}\left(\sum_{i=1}^{k-1} c_i n_i/n\right)} \tag{19.3.14}$$

Thus $\lambda_{\max}(\lambda_{\min})$ is the largest (smallest) deffs over all possible linear combination of the $\widehat{p}_i$'s. Rao and Scott (1981) termed λ_i's as generalized deffs.

In case λ_i's or their consistent estimates $\widehat{\lambda}_i$ are known, one can attain good approximations for the percentage points of the asymptotic distribution of X_G^2 using Solomon and Stephens (1977). But knowledge of λ_i's (or $\widehat{\lambda}_i$'s) require the knowledge of $\mathbf{V}$ or $\widehat{\mathbf{V}}$ and if $\mathbf{V}$ or $\widehat{\mathbf{V}}$ is known, one could use the Wald Statistic.

19.3.3 Modifications to X_G^2

19.3.3.1 Use of Maximum or Minimum Eigenvalues

From the expression (Eq. 19.3.12) we note

$$\frac{X_G^2}{\lambda_{\max,0}} \le \sum_{i=1}^{k-1} Z_i^2 = \chi_{k-1}^2 \tag{19.3.15}$$

where χ^2_{k-1} follows chi-square distribution with $k-1$ df and $\lambda_{\max,0} = \max\{\lambda_{i0}\}$.

If $\lambda_{\max,0}$ or its reliable estimate is known and one treats $\dfrac{X_G^2}{\lambda_{\max,0}}$ as a chi-square variable with $k-1$ df for testing the hypothesis $H_0: \mathbf{p} = \mathbf{p}_0$, then the test would result in a conservative test, which produces a lower significance level. In other words, suppose we set a level of significance α and reject H_0 if $\dfrac{X_G^2}{\lambda_{\max,0}} > \chi^2_{\alpha,k-1}$, where $\chi^2_{\alpha,k-1}$ is the upper 100 α% point of χ^2_{k-1}, then the true level of significance (type I error probability) of $\dfrac{X_G^2}{\lambda_{\max,0}}$ is lower than α. Similarly, treating $\dfrac{X_G^2}{\lambda_{\min,0}}$ with $\lambda_{\min,0} = \min\{\lambda_{i0}\}$ as a chi-square variable with $k-1$ df will produce a test with much higher significance level than the desired significance level.

19.3.3.2 Rao–Scott First-Order Corrections

Consider the test statistic

$$X_A^2 = X_G^2 / \overline{\lambda}_0 \tag{19.3.16}$$

where

$$\overline{\lambda}_0 = \sum_{i=1}^{k-1} \lambda_{i0}/(k-1) \tag{19.3.17}$$

Noting

$$E\left(X_A^2\right) = E\left(X_G^2\right)/\overline{\lambda}_0 = \sum_{i=}^{k-1} \lambda_{i0} E\left(Z_i^2\right)/\overline{\lambda}_0 = (k-1)$$

one may treat $\widehat{X}_A^2 = X_G^2 \Big/ \widehat{\overline{\lambda}}_0$ as χ^2_{k-1} with $\widehat{\overline{\lambda}}_0$ as an estimated value of $\overline{\lambda}_0$ and reject H_0 if $\widehat{X}_A^2 > \chi^2_{\alpha;k-1}$. In this case, the desired level of significance may be achieved approximately provided $\widehat{\lambda}_{i0}$'s do not vary much among themselves. One advantage of using $\widehat{\overline{\lambda}}_0$ is that it requires only estimation of V_{ii}'s but does not require estimation of V_{ij}'s because

$$\widehat{\overline{\lambda}}_0 = \frac{1}{(k-1)} \sum_{i=1}^{k} \frac{\widehat{V}_{ii}}{p_{i0}} = \frac{1}{(k-1)} \sum_{i=1}^{k} \frac{\widehat{p}_i\left(1-\widehat{p}_i\right)\widehat{d}_i}{p_{i0}}$$

with $\widehat{d}_i = \widehat{V}_{ii} \Big/ \left[\widehat{p}_i\left(1-\widehat{p}_i\right)\right]$ = estimated deff of the ith cell proportion.

19.3.3.3 Rao–Scott Second-Order Corrections

Consider

$$X_B^2 = \frac{X_A^2}{(1+a^2)} = \frac{X_G^2}{\overline{\lambda}_0(1+a^2)} \tag{19.3.18}$$

where a^2 is a constant, which makes

$$E\left(X_B^2\right) = E\left(\chi_\nu^2\right) = \nu \tag{19.3.19}$$

and

$$\begin{aligned} V\left(X_B^2\right) &= \frac{2}{(1+a^2)^2}\left[(k-1) + \sum_{i=1}^{k-1}\frac{\left(\lambda_{i0}-\overline{\lambda}_0\right)^2}{\overline{\lambda}_0^2}\right] \\ &= V\left(\chi_\nu^2\right) = 2\nu \end{aligned} \tag{19.3.20}$$

Eqs. (19.3.19) and (19.3.20) yield

$$\nu = \frac{(k-1)}{1+a^2} \quad \text{and} \quad a^2 = \frac{\sum_{i=1}^{k-1}\left(\lambda_{i0}-\overline{\lambda}_0\right)^2}{(k-1)\overline{\lambda}_0^2} \tag{19.3.21}$$

Thus under H_0, the statistic $\widehat{X}_B^2 = \dfrac{\hat{X}_A^2}{(1+\hat{a}^2)}$ with $\hat{a}^2 = \dfrac{\sum_{i=1}^{k-1}\left(\hat{\lambda}_{i0}-\hat{\overline{\lambda}}_0\right)^2}{(k-1)\hat{\overline{\lambda}}_0^2}$ is a good approximation of χ_ν^2 in the sense that it has first two order moments exactly equal to χ_ν^2. Thus the test statistic $\widehat{X}_B^2$ provides the desired level of significance. The main disadvantage of using $\widehat{X}_B^2$ is that we need to estimate a^2, which in turn requires estimation of $\sum_{i=1}^{k-1}\lambda_{i0}^2 = \sum_{i=1}^{k}\sum_{j=1}^{k} V_{ij}^2 \Big/ \left(p_{i0}p_{j0}\right)$, i.e., estimates of covariances V_{ij}/n's are needed in the calculation of $\widehat{X}_B^2$.

19.3.3.4 Fellegi Correction

In case estimates of V_{ij}'s are either not available or reliable, the following mean deff adjustment was proposed by Fellegi (1980)

$$X_F^2 = X_G^2 \Big/ \widehat{\overline{d}} \tag{19.3.22}$$

where $\widehat{\overline{d}} = \sum_{i=1}^{k}\widehat{d}_i \Big/ k$ and $\widehat{d}_i = \widehat{V}_{ii} \Big/ \left[\widehat{p}_i\left(1-\widehat{p}_i\right)\right]$.

19.3.4 Simple Random Sampling Without Replacement

For SRSWOR sampling, $\mathbf{V} = (1 - n/N)\mathbf{P}$ and hence $\mathbf{D}_0 = \mathbf{P}_0^{-1}$ $\mathbf{V}_0 = (1 - n/N)\mathbf{I}_{\mathbf{k}-1}$. In this case $\lambda_{10} = \ldots = \lambda_{k-10} = 1 - n/N$ and $X_G^2/(1 - n/N)$ is asymptotically distributed as χ_{k-1}^2. Thus treating X_G^2 as χ_{k-1}^2, one achieves the desired level of significance provided the sampling fraction n/N is negligible. But if n/N is not negligible, the test statistic X_G^2 will provide a lower significance level.

19.3.5 Stratified Sampling

Consider a population stratified into L strata. Let W_h and p_{hi}, respectively, denote the population proportion of units from the stratum h and proportion of elements from stratum h that belong to the ith category. Let a sample s_h of size m_h be selected from the hth stratum by SRSWR and m_{hi} be the number of units that belong to ith category. Then $p_i = \sum_h W_h p_{hi}$, the population proportion of units that belong to the ith category can be estimated by

$$\widehat{p}_i = \sum_h W_h \widehat{p}_{hi} \quad \text{where } \widehat{p}_{hi} = m_{hi}/m_h \tag{19.3.23}$$

For proportional allocation $m_h = nW_h$ with $n = \sum_{h=1}^{L} m_h$. In this case $\widehat{p}_i$ reduces to n_i/n, where n_i is the total number of units in the sample $s(= s_1 \cup \ldots \cup s_h)$ that fall in the ith category. The variance of $\widehat{p}_i$ and the covariance of $\widehat{p}_i$ and $\widehat{p}_j$ under proportional allocation are, respectively, given by

$$Var(\widehat{p}_i) = \frac{1}{n}\sum_{h=1}^{L} W_h p_{hi}(1 - p_{hi}) = \frac{p_i}{n} - \frac{1}{n}\sum_{h=1}^{L} W_h p_{hi}^2 \text{ and}$$

$$Cov\left(\widehat{p}_i, \widehat{p}_j\right) = -\frac{1}{n}\sum_{h=1}^{L} W_h p_{hi} p_{hj}.$$

The variance–covariance matrix of $\widehat{\mathbf{p}} = \left(\widehat{p}_1, \ldots, \widehat{p}_{k-1}\right)'$ is $\mathbf{V}(st)/n$ with

$$\begin{aligned} \mathbf{V}(st) &= diag(\mathbf{p}) - \sum_{h=1}^{L} W_h \mathbf{p}_h \mathbf{p}_h' \\ &= \mathbf{P} - \sum_{h=1}^{L} W_h (\mathbf{p}_h - \mathbf{p})(\mathbf{p}_h - \mathbf{p})' \end{aligned} \tag{19.3.24}$$

where $\mathbf{p}_h = (p_{h,1}, \ldots, p_{h,k-1})'$ and $\mathbf{P}$ is as in Theorem 19.3.1.

Noting $0 \leq \mathbf{c}'\mathbf{V}(st)\mathbf{c}/\mathbf{c}'\mathbf{P}\mathbf{c} \leq 1$, we find that all the eigenvalues of $\mathbf{D}_0 = \mathbf{P}^{-1}\mathbf{V}(st)\big|_{\mathbf{p}=\mathbf{p}_0}$ are less than 1. Hence under $\mathbf{H}_0$,

$$0 \leq X_G^2 \leq \sum_{i=1}^{k-1} Z_i^2 \approx \chi_{k-1}^2 \tag{19.3.25}$$

Thus for stratified sampling, Pearsonian chi-square X_G^2 in Eq. (19.3.11) always becomes an asymptotically conservative test. Rao and Scott (1981) showed that for two strata ($L = 2$)

$$X_G^2 \cong \chi_{k-2}^2 + (1 - \delta_0)\chi_1^2 \tag{19.3.26}$$

where $\delta_0 = W_1 W_2 \sum_{i=1}^{k} (p_{1i} - p_{2i})^2 / p_{i0}$ is the minimum eigenvalue of $\mathbf{D}_0$.

Hence, X_G^2 can be asymptotically well approximated by χ_{k-1}^2 unless k is small.

19.3.6 Two-Stage Sampling

Consider a two-stage sampling where a population consists of R first-stage units (fsu's) and hth fsu consists of M_h second-stage units (ssu's). A sample s of r fsu's is selected from R fsu's by probability proportional to size with replacement sampling using $W_h = M_h/M \left(M = \sum_{h=1}^{R} M_h\right)$ as a normed size measure for the hth fsu. If the hth fsu is selected in the sample s, a subsample s_h of m ssu's is selected from it by SRSWR method. So, the total number of the selected ssu's in the sample s is $n = rm$. Let us denote $\gamma_{hj}(i)$ as 1 if the jth ssu of the hth fsu belong to the ith category and $\gamma_{hj}(i)$ is 0 otherwise. Then the proportion of units that belong to the ith category in s_h is $\hat{p}_{hi} = m_{hi}/m = \overline{y}_h(i) = \sum_{j \in s_h} \gamma_{hj}(i)/m$ where m_{hi} is the total number of ssu's that belong to the ith category and $\sum_{j \in s_h}$ denotes the sum over fsu's in the sample s_h including repetition. The proportion of units belonging to the ith category in the population is $p_i = \sum_{h=1}^{R} W_h \overline{Y}_h(i) = \sum_{h=1}^{R} W_h p_{hi}$, where $\overline{Y}_h(i) = \sum_{j=1}^{M_h} \gamma_{hj}(i)/M_h = p_{hi}$.

Theorem 19.3.2

(i) $\widehat{p}_i = \frac{1}{r}\sum_{h\in s}\overline{y}_h(i) = \frac{1}{r}\sum_{h\in s}\widehat{p}_{hi}$ is unbiased for p_i

(ii) $V(\widehat{p}_i) = \frac{1}{n}\left[p_i(1-p_i) + (m-1)\sum_{h=1}^{R} W_h(p_{hi}-p_i)^2\right]$

(iii) $Cov\left(\widehat{p}_i, \widehat{p}_j\right) = \frac{1}{n}\left[-p_ip_j + (m-1)\sum_{h=1}^{R} W_h\{p_h(i)-p_i\}\{p_h(j)-p_j\}\right]$

Proof

(i) $E(\widehat{p}_i) = E\left[\frac{1}{Mr}\sum_{h\in s}\frac{1}{W_h}M_hE\{\overline{y}_h(i)|s\}\right]$

$= \frac{1}{M}\sum_{h=1}^{R} M_h\overline{Y}_h(i) = p_i.$

(ii) $V(\widehat{p}_i) = E[V(\widehat{p}_i|s)] + V[E(\widehat{p}_i|s)]$

$= E\left[\frac{1}{M^2r^2}\sum_{h\in s}\frac{1}{W_h^2}M_h^2\frac{p_{hi}(1-p_{hi})}{m}\right] + V\left(\frac{1}{Mr}\sum_{h\in s}\frac{1}{W_h}M_hp_{hi}\right)$

$= \frac{1}{n}\left[\sum_{h=1}^{R} W_hp_{hi}(1-p_{hi}) + m\left(\sum_{h=1}^{R} W_hp_{hi}^2 - p_i^2\right)\right]$

$= \frac{1}{n}\left[p_i(1-p_i) + (m-1)\sum_{h=1}^{R} W_h(p_{hi}-p_i)^2\right]$

$$
\begin{aligned}
\text{(iii) } Cov\left(\widehat{p}_i, \widehat{p}_j\right) &= Cov\left[E\left(\widehat{p}_i, \widehat{p}_j \middle| s\right)\right] + E\left[Cov\left(\widehat{p}_i, \widehat{p}_j \middle| s\right)\right] \\
&= \frac{1}{M^2}\left[Cov\left\{\frac{1}{r}\sum_{h\in s}\frac{1}{W_h}M_h\overline{Y}_h(i), \frac{1}{r}\sum_{h\in s}\frac{1}{W_h}M_h\overline{Y}_h(j)\right\}\right] \\
&\quad + \frac{1}{M^2}\left[E\left\{Cov\left(\frac{1}{r}\sum_{h\in s}\frac{1}{W_h}M_h\overline{y}_h(i), \frac{1}{r}\sum_{h\in s}\frac{1}{W_h}M_h\overline{y}_h(j)\middle| s\right)\right\}\right] \\
&= \frac{1}{M^2}\left[\frac{1}{r^2}\sum_{h\in s}Cov\left(\frac{M_h}{W_h}\overline{Y}_h(i), \frac{M_h}{W_h}\overline{Y}_h(j)\right)\right. \\
&\quad \left. - E\left\{\frac{1}{r^2 m}\sum_{h\in s}\left(\frac{M_h}{W_h}\right)^2\overline{Y}_h(i)\overline{Y}_h(j)\right\}\right] \\
&= \frac{1}{M^2}\frac{1}{r}\left[\left\{E\left(\frac{M_h}{W_h}p_{hi}\frac{M_h}{W_h}p_{hj}\right) - M^2 p_i p_j\right\} - \frac{1}{m}\sum_{l=1}^{R}\frac{M_h^2}{W_h}p_{hi}p_{hj}\right] \\
&= \frac{1}{r}\left[\sum_{h=1}^{R}W_h p_{hi}p_{hj} - p_i p_j - \frac{1}{m}\sum_{l=1}^{R}W_h p_{hi}p_{hj}\right] \\
&= \frac{1}{n}\left[-p_i p_j + (m-1)\sum_{h=1}^{R}W_h(p_{hi}-p_i)(p_{hj}-p_j)\right]
\end{aligned}
$$

For large r, the distribution of $\widehat{\mathbf{p}} = \left(\widehat{p}_1, \ldots, \widehat{p}_{k-1}\right)'$ is approximately normal with mean $\mathbf{p} = \left(p_1, \ldots, p_{k-1}\right)'$ and variance–covariance matrix $\mathbf{V}(t)/n$, where

$$
\mathbf{V}(t) = \mathbf{P} + (m-1)\mathbf{A}, \tag{19.3.27}
$$

$\mathbf{A} = \sum_{h=1}^{R} W_h(\mathbf{p_h} - \mathbf{p})(\mathbf{p_h} - \mathbf{p})'$, $\mathbf{p_h} = \left(p_{h1}, \ldots, p_{h,k-1}\right)'$, and $\mathbf{P} = diag(\mathbf{p}) - \mathbf{p}\mathbf{p}'$.

Hence under $\mathbf{H}_0$: $\mathbf{p} = \mathbf{p}_0$,

$$
\begin{aligned}
X_G^2 &= n\sum_{i=1}^{k-1}\frac{\left(\widehat{p}_i - p_{i0}\right)^2}{p_{i0}} \\
&= \sum_{i=1}^{k-1}\{1 + (m-1)\rho_{i0}\}Z_i^2
\end{aligned} \tag{19.3.28}
$$

where $\rho_{10},\ldots,\ \rho_{k-1,0}$ are the eigenvalues of $\mathbf{P}^{-1}\mathbf{A}$ for $\mathbf{p} = \mathbf{p}_0$. Noting $0 \le c'\mathbf{A}c/c'\mathbf{P}c \le 1$, we find that $\rho_1,\ldots,\ \rho_{k-1}$, the eigenvalues of $\mathbf{A}$, are all positive and less than unity. Hence we can write

$$\mathbf{X}_G^2 \le [1+(m-1)\rho_{\max 0}]\sum_{i=1}^{k-1} Z_i^2 \ \text{ with } \ \rho_{\max,0} = \max\{\rho_{10},\ldots,\rho_{k-1,0}\}$$

$$\text{i.e.,} \quad \frac{X_G^2}{m} \le \sum_{i=1}^{k-1} Z_i^2 \tag{19.3.29}$$

Thus $\dfrac{X_G^2}{m}$ provides asymptotically a conservative test. Rao and Scott (1981) called ρ_i's a generalized measure of homogeneity similar to intraclass correlation ρ. However, one can use the Wald statistic if $\widehat{\mathbf{p}}_h$ are available because $\widehat{\mathbf{V}}(t) = m\sum_{h=1}^{r}(\widehat{\mathbf{p}}_h - \widehat{\mathbf{p}})(\widehat{\mathbf{p}}_h - \widehat{\mathbf{p}})'\big/(r-1)$, the estimator of $\mathbf{V}(t)$ is easy to compute. Thomas and Rao (1987) showed that the Wald test provides poor control of type1 error if the degree of freedom f for estimating $\widehat{\mathbf{V}}$ is not more than the degrees of freedom for the hypothesis. So, the application of Wald statistics is limited because the degrees of freedom $\widehat{\mathbf{V}}$ of most survey designs are at most moderate. To overcome the instability problem of $\widehat{\mathbf{V}}$, the following F-corrected Wald statistics are proposed by Rao and Thomas (1988).

$$F_{1w} = \frac{f-k+2}{f(k-1)}X_W^2 \ \text{ and } \ F_{2w} = X_W^2/(k-1) \tag{19.3.30}$$

where $f =$ number of sampled clusters $-$ number of strata.

The statistics F_{1w} and F_{2w} are treated as F distribution with degrees of freedom $k-1$, $f-k+2$ and $k-1$, f, respectively. Here we note that for $k=2$ both the statistics F_{1w} and F_{2w} reproduce the original Wald statistic. For further details readers are referred to Lehtonen and Pahkinen (2004).

19.3.7 Residual Analysis

In case the hypothesis of goodness of fit $H_0: \mathbf{p} = \mathbf{p}_0$ is rejected, it is important to check which of the p_i's differ significantly from p_{i0}. In this situation we test the following k hypotheses separately. The hypothesis

$$H_{0i}: p_i = p_{i0} \ \text{ against alternative } \ H_{1i}: p_i \neq p_{i0} \quad \text{for } \ i = 1,\ldots,k$$

should be rejected at $\alpha \times 100\%$ level of significance if $|\widehat{e}_i| > z_{\alpha/2}$, where $\widehat{e}_i = \dfrac{\hat{p}_i - p_{i0}}{\sqrt{\hat{V}(\hat{p}_i)}} = \dfrac{(\hat{p}_i - p_{i0})}{\sqrt{\hat{d}_i \hat{p}_i (1-\hat{p}_i)/n}} = \dfrac{e_i}{\sqrt{\hat{d}_i}}$, $e_i = \dfrac{(\hat{p}_i - p_{i0})}{\sqrt{\hat{p}_i(1-\hat{p}_i)/n}} =$ standardized residual under SRSWR sampling, and z_α is the upper $\alpha \times 100\%$ point of the standardized normal distribution.

Example 19.3.4

Consider an artificial example where a population of 300 enumeration areas (EAs), which was stratified into 15 strata. From each of the strata two EAs were selected using a suitable inclusion probability proportional to size sampling scheme using the number of people in the EA as a measure of size variable. From each of the selected EAs a sample of 50 individuals was selected by SRSWOR method. The selected sample of 1500 individuals was classified into six categories according to their income level. The estimated and hypothesized proportions for the six categories and estimated variance–covariance matrix $\left(\widehat{V}/n\right)$ were computed as follows:

Income level	Estimated proportion ($\hat{p}_i$)	Hypothesized proportion (p_{i0})
A	0.120	0.120
B	0.205	0.190
C	0.210	0.230
D	0.214	0.250
E	0.126	0.120
F	0.125	0.090

$$\widehat{V}/n = 10^{-5} \times \begin{pmatrix} 20.875 & 1.024 & -0.624 & -5.070 & -3.258 & -12.947 \\ 1.024 & 10.258 & 3.646 & 0.177 & -4.816 & -10.289 \\ -0.624 & 3.646 & 20.745 & -6.317 & -4.586 & -12.864 \\ -5.070 & 0.177 & -6.317 & 30.785 & -5.718 & -13.857 \\ -3.258 & -4.816 & -4.586 & -5.718 & 20.650 & -2.272 \\ -12.947 & -10.289 & -12.864 & -13.857 & -2.272 & 52.229 \end{pmatrix}$$

Here, the generalized Pearsonian chi-square:

$X_G^2 = n\sum_{i=1}^{k} \dfrac{(\hat{p}_i - p_{i0})^2}{p_{i0}} = 1500 \sum_{i=1}^{6} \dfrac{(\hat{p}_i - p_{i0})^2}{p_{i0}} = 33.028$ with 5 $(=k-1)$ df and has a p-value 0.0000.

Modified chi-square:

$X_M^2 = n\sum_{i=1}^{k}\frac{\left(\hat{p}_i - p_{i0}\right)^2}{\hat{p}_i} = 1500\sum_{i=1}^{6}\frac{\left(\hat{p}_i - p_{i0}\right)^2}{\hat{p}_i} = 28.716$ with 5 df and has a p-value 0.0000.

Rao–Scott first-order correction:

The estimated deffs for the six categories that are obtained using the formula $\widehat{d}_i = \widehat{V}_{ii}/\left[\widehat{p}_i(1-\widehat{p})\right]$ are as follows:

$\widehat{d}_1 = 2.965, \widehat{d}_2 = 0.944, \widehat{d}_3 = 1.876, \widehat{d}_4 = 2.745, \widehat{d}_5 = 2.813,$ and $\widehat{d}_6 = 7.163.$

The estimated value of

$\overline{\lambda}_0 = \sum_{i=1}^{6}\widehat{V}_{ii}/(5p_{i0}) = \sum_{i=1}^{6}\widehat{p}_i(1-\widehat{p}_i)\ \widehat{d}_i\Big/(5p_{i0}) = 3.581.$

$\widehat{X}_A^2 = X_G^2\Big/\widehat{\overline{\lambda}}_0 = 30.028/3.581 = 9.222$ with 5 df and has a p-value 0.1005.

Rao–Scott second-order correction:

$1 + \widehat{a}^2 = \frac{1}{\left(5\widehat{\overline{\lambda}}^2\right)}\sum_{i=1}^{6}\sum_{j=1}^{6}\frac{\widehat{V}_{ij}^2}{p_{i0}p_{j0}} = 1.794$ and $\nu = 5/(1+\widehat{a}^2) = 2.787.$

$\widehat{X}_B^2 = \widehat{X}_A^2\Big/(1+\widehat{a}^2) = 9.222/1.794 = 5.140$ with 2.787 df and has a p-value 0.1618.

Fellegi correction:

$\widehat{X}_F^2 = \widehat{X}_G^2\Big/\widehat{\overline{d}} = 33.028/3.084 = 10.708$ with 5 df and has a p-value 0.0575.

Wald statistic:

$X_W^2 = \left(\widehat{\mathbf{p}} - \mathbf{p}_0\right)'\left(\frac{\mathbf{V}}{n}\right)^{-1}\left(\widehat{\mathbf{p}} - \mathbf{p}_0\right) = 13.121$ with 5 df and ha a p-value 0.0223.

F-corrections for Wald statistic:

Noting $f = m - h = 50 - 15 = 35$, we get

$F_1 = \frac{f-k+2}{f(k-1)}X_W^2 = 1.771 \times 13.121 = 2.324$ with 5 and 31 df and has a p-value 0.0665.

$F_2 = X_W^2/(k-1) = 13.121/5 = 2.624$ with 5 and 35 df and has a p-value 0.0407.

Residual analysis:

The standardized residuals are $e_1 = 0$, $e_2 = 1.481$, $e_3 = -1.388$, $e_4 = -2.051$, $e_5 = 0.417$, and $e_6 = 1.531$.

Here we note that the values of statistics X_G^2 and X_M^2 are highly significant and provide strong evidence of rejecting the hypothesis H_0 while the statistics $\widehat{X}_A^2$, $\widehat{X}_B^2$, F_1, and F_2 provide strong evidence of accepting the hypothesis. The Wald statistic X_W^2 favors acceptance of hypothesis H_0. The residual analysis reveals that the population proportions for the income groups do not deviate significantly from the respective hypothesized proportions.

19.4 TEST OF INDEPENDENCE

Consider a finite population of N identifiable units, which has been classified into "a" levels of factor A and "b" levels of the factor B. Let the proportion of individuals belonging to the ith level of factor A and jth level of the factor B in the population be p_{ij}, $i = 1,\ldots, a$; $j = 1,\ldots, b$. Here we are interested in testing the independence of the two factors A and B, i.e., we want to test

$$\mathbf{H}_0: \theta_{ij} = p_{ij} - p_{i\cdot}p_{\cdot j} = 0 \quad \text{for } i = 1, \ldots, a;\ j = 1, \ldots, b \tag{19.4.1}$$

against alternative $\mathbf{H}_1: \theta_{ij} \neq 0$ for at least one of the combination i, j where $p_{i\cdot} = \sum_{j=1}^{b} p_{ij}$ and $p_{\cdot j} = \sum_{i=1}^{a} p_{ij}$.

Let a sample s of size n be selected from the population with probability $p(s)$ using a complex survey design and $\widehat{p}_{ij}$ be an unbiased or consistent estimator of p_{ij} based on the sample s.

19.4.1 Wald Statistic

For large n, under $\mathbf{H}_0$ the Wald statistic

$$X_w^2 = n\widehat{\boldsymbol{\theta}}'\ \widehat{\mathbf{V}}_{\theta}^{-1}\ \widehat{\boldsymbol{\theta}} \tag{19.4.2}$$

follows a χ_t^2 (chi-square distribution with t df) with $t = (a-1)\ (b-1)$ where $\widehat{\boldsymbol{\theta}}' = \left(\widehat{\theta}_{11}, \widehat{\theta}_{12}, \ldots, \widehat{\theta}_{a-1,b-1}\right)$ and $\widehat{\mathbf{V}}_{\theta}/n$ is an estimated variance–covariance matrix of $\widehat{\boldsymbol{\theta}} = V\left(\widehat{\boldsymbol{\theta}}\right)\Big/ n = \mathbf{V}_{\theta}/n$. The estimator $\widehat{\mathbf{V}}_{\theta}$ can be obtained by applying any of the methods of variance estimation, e.g. LR, RG, BRR, or BT described in Chapter 18. The estimators of covariances become unstable, if the cell frequencies are small. Hence in practice, the performance of the Wald statistic for a large contingency table is poor (Lohr, 1999) because of small cell frequencies.

19.4.2 Bonferroni Test

To avoid calculations of covariances, one may perform the following $(a-1)\,(b-1)$ separate tests

$$H_{011}: \theta_{11} = 0;\ H_{012}: \theta_{12} = 0; \ldots, H_{0(a-1)(b-1)}: \theta_{a-1,b-1} = 0$$

The hypothesis $\mathbf{H}_0$ stated in Eq. (19.4.1) is rejected at α 100% level of significance if any of the test H_{0ij} is rejected at $\dfrac{\alpha}{2t} \times 100\%$ level of significance, i.e., if any one

$$\frac{\left|\widehat{\theta}_{ij}\right|}{\sqrt{\widehat{V}\left(\widehat{\theta}_{ij}\right)}} > t_{\frac{\alpha}{2t}, q} \tag{19.4.3}$$

where $t_{\frac{\alpha}{2t}, q}$ is the upper $(\alpha/2t)$ 100% point of t distribution with q df and q is the degree of freedom carried by the estimator of the variance. More details are given by Lohr (1999) and Thomas (1989).

19.4.3 Modified Chi-Square

The modified chi-square statistic for complex survey design for testing $\mathbf{H}_0$ stated in Eq. (19.4.1) is given by

$$X_M^2 = n \sum_{i=1}^{a-1} \sum_{j=1}^{b-1} \frac{\left(\widehat{p}_{ij} - \widehat{p}_{i\cdot}\widehat{p}_{\cdot j}\right)^2}{\widehat{p}_{i\cdot}\widehat{p}_{\cdot j}} \tag{19.4.4}$$

Now, writing $\widehat{\boldsymbol{\theta}} = \left(\widehat{\theta}_{11}, \widehat{\theta}_{12}, \ldots, \widehat{\theta}_{a-1,b-1}\right)'$, $\mathbf{p} = \left(p_{11}, p_{12}, \ldots, p_{a-1,b-1}\right)'$, $\widehat{\mathbf{p}}_r = \left(\widehat{p}_{1\cdot}, \ldots, \widehat{p}_{a-1\cdot}\right)'$, $\widehat{\mathbf{p}}_c = \left(\widehat{p}_{\cdot 1}, \ldots, \widehat{p}_{\cdot c-1}\right)'$, $\widehat{\mathbf{P}}_r = diag\left(\widehat{\mathbf{p}}_r\right) - \widehat{\mathbf{p}}_r\widehat{\mathbf{p}}_r'$, and $\widehat{\mathbf{P}}_c = diag\left(\widehat{\mathbf{p}}_c\right) - \widehat{\mathbf{p}}_c\widehat{\mathbf{p}}_c'$, the expression (Eq. 19.4.4) can be written as

$$X_M^2 = n\widehat{\boldsymbol{\theta}}'\left(\widehat{\mathbf{P}}_r^{-1} \otimes \widehat{\mathbf{P}}_c^{-1}\right)\widehat{\boldsymbol{\theta}} \tag{19.4.5}$$

where $\otimes$ denotes the kronecker product.

Rao and Scott (1981) showed that under $\mathbf{H}_0$, X_M^2 is asymptotically distributed as

$$X_M^2 \approx \sum_{i=1}^{t} \delta_{i0} Z_i^2 \tag{19.4.6}$$

where Z_i's are iid $N(0, 1)$ and δ_{i0}'s are eigenvalues of $\mathbf{D}_{\boldsymbol{\theta}} = \mathbf{P}_r^{-1} \otimes \mathbf{P}_c^{-1}\mathbf{V}_{\boldsymbol{\theta}}$ under H_0. Let $\widehat{\delta}_{i0}$ be an estimator of δ_{i0}, $\widehat{\delta}_{\max,0} = \max\left\{\widehat{\delta}_{i0}\right\}$, and

$\widehat{\bar{\delta}}_0 = \sum_{i=1}^{t} \widehat{\delta}_{i0} \Big/ t$. Then, $X_M^2 \Big/ \widehat{\delta}_{\max,0}$ provides a conservative test, while $X_M^2 \Big/ \widehat{\bar{\delta}}_0 \approx \sum_{i=1}^{t} \left(\widehat{\delta}_{i0} \Big/ \widehat{\bar{\delta}}_0 \right) Z_i^2$ proves a better test.

19.5 TESTS OF HOMOGENEITY

Suppose we have g populations each of which is classified into k categories. Let the number of units belonging to the ith category of the jth population be N_{ij}, $i = 1,\ldots, k$; $j = 1,\ldots, g$ and their proportion be $p_{ij} = N_{ij}/N_i$, where $N_j = \sum_{i=1}^{k} N_{ij}$ is the size of the jth population and $\sum_{i=1}^{k} p_{ij} = 1$ for $j = 1,\ldots, g$. The problem of homogeneity consists of testing the hypothesis

$$H_0 : p_{i1} = p_{i2} = \cdots = p_{ig} \quad \text{for } i = 1, \ldots, k-1 \tag{19.5.1}$$

against alternative H_1:H_0 is false.

In matrix notation, H_0 can be written as

$$\mathbf{H}_0 : \mathbf{p}_1 = \mathbf{p}_2 = \cdots = \mathbf{p}_\mathbf{g} \text{ with } \mathbf{p}_\mathbf{j} = \left(p_{1j}, p_{2j}, \ldots, p_{k-1,j}\right)'$$

From each of the populations, samples are selected independently using some complex sampling design. Let s_j be a sample of size n_j selected from the jth population and let the number of units falling in the ith class be n_{ij}. For simplicity, let us consider $g = 2$ and consider the null hypothesis

$$\mathbf{H}_0 : \mathbf{p}_1 = \mathbf{p}_2 \text{ against } \mathbf{H}_1 : \mathbf{p}_1 \neq \mathbf{p}_2 \tag{19.5.2}$$

As usual, let $\widehat{p}_{ij}$ be an unbiased or a consistent estimator of p_{ij} and $\widehat{\mathbf{p}}_j = \left(\widehat{p}_{1j}, \widehat{p}_{2j}, \ldots, \widehat{p}_{k-1,j}\right)'$.

19.5.1 Wald Statistic

Let $\widehat{V}_i / n_i$ be a consistent estimator of the variance–covariance matrix of $\widehat{\mathbf{p}}_\mathbf{i}$ for $i = 1, 2$. Under the null hypothesis $\mathbf{H}_0$, the Wald statistic

$$X_{WH}^2 = \left(\widehat{\mathbf{p}}_1 - \widehat{\mathbf{p}}_2\right)' \left(\frac{\widehat{\mathbf{V}}_1}{n_1} + \frac{\widehat{\mathbf{V}}_2}{n_2} \right)^{-1} \left(\widehat{\mathbf{p}}_1 - \widehat{\mathbf{p}}_2\right) \tag{19.5.3}$$

follows χ_{k-1}^2, chi-square distribution with $k-1$ df when sample sizes n_1 and n_2 are so large that each of $\widehat{\mathbf{p}}_1$ and $\widehat{\mathbf{p}}_2$ are distributed independently as $k-1$ normal variables. The statistic X_{WH}^2 has limited application because of the problem of reliable estimation of covariances, especially when the number of classes is large.

19.5.2 Modified Chi-Square Statistics

The Pearsonian chi-square statistic for testing $\mathbf{H}_0$ is given by

$$X_{PH}^2 = n\sum_{i=1}^{2}\sum_{j=1}^{k}\frac{\left(n_{ij} - n_{i\cdot}n_{\cdot j}/n\right)^2}{n_{i\cdot}n_{\cdot j}} \tag{19.5.4}$$

with $n_{i\cdot} = \sum_{j=1}^{k} n_{ij}$ and $n_{\cdot j} = \sum_{i=1}^{2} n_{ij}$.

The statistic X_{PH}^2 asymptotically follows χ_{k-1}^2, if the samples s_1 and s_2 are selected by SRSWR method. For a complex survey design, however, one should use the following modified chi-square statistic given by Scott and Rao (1981):

$$\begin{aligned} X_{MPH}^2 &= \sum_{i=1}^{2}\sum_{j=1}^{k}\frac{n_i\left(\widehat{p}_{ij} - \widehat{p}_{+j}\right)^2}{\widehat{p}_{+j}} \\ &= \frac{n_1 n_2}{n_1 + n_2}\left(\widehat{\mathbf{p}}_1 - \widehat{\mathbf{p}}_2\right)'\widehat{\mathbf{P}}^{-1}\left(\widehat{\mathbf{p}}_1 - \widehat{\mathbf{p}}_2\right) \end{aligned} \tag{19.5.5}$$

where $\widehat{p}_{+j} = \dfrac{n_1\hat{p}_{1j} + n_2\hat{p}_{2j}}{n_1 + n_2}$, $\widehat{\mathbf{P}} = diag\left(\widehat{\mathbf{p}}_+\right) - \widehat{\mathbf{p}}_+\widehat{\mathbf{p}}_+'$, and $\widehat{\mathbf{p}}_+ = \left(\widehat{p}_{+1}, \ldots, \widehat{p}_{+k-1}\right)'$.

So, the statistic X_{MPH}^2 is a particular case of Wald statistic X_{WH}^2 given in Eq. (19.5.3) when $\widehat{\mathbf{V}}_1 = \widehat{\mathbf{V}}_2 = \widehat{\mathbf{P}}$. The statistic X_{MPH}^2 does not asymptotically follow chi-square distribution. Scott and Rao (1981) showed that under the null hypothesis $\mathbf{H}_0$: $\mathbf{p}_1 = \mathbf{p}_2 = \mathbf{p}$,

$$X_{MPH}^2 = \sum_{j=1}^{k-1}\lambda_j Z_j^2 \tag{19.5.6}$$

where Z_i's are iid $N(0, 1)$ and $\lambda_1, \ldots, \lambda_{k-1}$ are eigenvalues of $\mathbf{D} = (n_2\mathbf{D}_1 + n_1\mathbf{D}_2)/(n_1 + n_2)$ with $\mathbf{D_i} = \mathbf{P}^{-1}\mathbf{V_i}$ and $\mathbf{P} = diag\ \mathbf{p} - \mathbf{pp}'$. The application of the statistics X_{MPH}^2 has already been discussed in Section 19.3.3. On the basis of extensive empirical studies based on two-stage sampling with varying probabilities, Scott and Rao (1981) concluded that treating X_{MPH}^2 as an ordinary chi-square results in severe distortion of significance level.

19.6 CHI-SQUARE TEST BASED ON SUPERPOPULATION MODEL

19.6.1 Altham's Model

Cohen (1976) proposed chi-square statistic based on a cluster sampling with a fixed cluster of size 2, using a superpopulation model. Altham (1976) extended

the theory to any fixed cluster size M. Rao and Scott (1981) extended this further to cover a general two-stage sampling. The Rao and Scott (1981) procedure is described as follows: Consider a two-stage sampling of R fsu's and the hth fsu consists of M_h ssu's. A sample s of r fsu's is selected from R fsu's by some suitable sampling scheme. If the hth fsu is selected in the sample s, a subsample s_h of size m_h ssu's is selected from the hth fsu by using a suitable sampling scheme. The total number of ssu's in the sample $s = (s_1 \cup \ldots \cup s_r)$ is $\sum_{h=1}^{r} m_h = n$. Let $\gamma_{hj}(i) = 1$ if the jth ssu of the hth fsu belong to the ith category and $\gamma_{hj}(i) = 0$ otherwise; $h = 1,\ldots, R$; $i = 1,\ldots, k$. Hence the total number of ssu's that belong to the ith category in the entire sample s is $n_i = \sum_{h\in s}\sum_{j\in s_h} \gamma_{hj}(i)$ and $n = \sum_{i=1}^{k} n_i$. Altham (1976) considered the following superpopulation model:

$$\text{(i) random variables } \gamma_{hj}(i) \text{ for different clusters (fsu's) are independent,}$$
$$\text{(ii) } E_m\{\gamma_{hj}(i)\} = \pi_i \text{ and (iii) } C_m\{\gamma_{hj}(i), \gamma_{hj'}(i')\} = \sigma_{ii'} \quad \text{for } j \neq j' \tag{19.6.1}$$

where E_m and C_m denote expectation and covariance operators with respect to the model.

From the model (Eq. 19.6.1), we have for a given s,

$$\text{(i) the model expectation of } n_i = E_m(n_i) = \sum_{h\in s}\sum_{j\in s_h} E_m\{\gamma_{hj}(i)\} = n\pi_i, \tag{19.6.2}$$

$$\begin{aligned}\text{(ii) the model variance of } n_i = V_m(n_i) &= \left[\sum_{h\in s} V_m\left\{\sum_{j\in s_h}\gamma_{hj}(i)\right\}\right. \\ &\quad \left. + \sum_{h\neq}\sum_{h'\in s} C_m\left\{\sum_{j\in s_h}\gamma_{hj}(i), \sum_{j\in s_{h'}}\gamma_{h'j}(i)\right\}\right] \\ &= \pi_i(1-\pi_i)n + \sigma_{ii}\left\{\sum_{h\in s} m_h(m_h-1)\right\} \\ &= n\pi_i(1-\pi_i) + \sigma_{ii}\left(\sum_{h\in s} m_h^2 - n\right) \\ &\quad (\text{noting } V_m\{\gamma_{hj}(i)\} = \pi_i(1-\pi_i))\end{aligned} \tag{19.6.3}$$

and
(iii) the model covariance of n_i and $n_{i'}$ is

$$\begin{aligned} C_m(n_i, n_{i'}) &= \sum_{h\in s} C_m\left(\sum_{j\in s_h}\gamma_{hj}(i), \sum_{j\in s_h}\gamma_{hj}(i')\right) \\ &= \sum_{h\in s}\left[\sum_{j\in s_h} C_m\{\gamma_{hj}(i), \gamma_{hj}(i')\}\right. \\ &\qquad \left. + \sum_{j\neq} \sum_{j'\in s_h} C_m\{\gamma_{hj}(i), \gamma_{hj'}(i')\}\right] \\ &= -n\pi_i\pi_{i'} + \sigma_{ii'}\left(\sum_{h\in s} m_h^2 - n\right) \\ &\qquad (\text{noting } C_m\{\gamma_{hj}(i), \gamma_{hj}(i')\} = -\pi_i\pi_{i'}) \end{aligned} \tag{19.6.4}$$

Let $\widehat{\boldsymbol{\pi}} = (\widehat{\pi}_1, \ldots, \widehat{\pi}_i, \ldots, \widehat{\pi}_{k-1})'$ with $\widehat{\pi}_i = n_i/n$. Then from Eqs. (19.6.2)–(19.6.4) we note that under the model (Eq. 19.6.1) and a given s, the expectation and variance–covariance matrix of $\widehat{\boldsymbol{\pi}}$ are, respectively,

$$E_m(\widehat{\boldsymbol{\pi}}) = \boldsymbol{\pi} = (\pi_1, \ldots, \pi_i, \ldots, \pi_{k-1})' \text{ and } V_m(\widehat{\boldsymbol{\pi}}) = \mathbf{Q}_{ms}/n \tag{19.6.5}$$

where $\mathbf{Q}_{ms} = \{\boldsymbol{\Delta}(\boldsymbol{\pi}) + (m_{0s} - 1)\boldsymbol{\Phi}\}$, $\boldsymbol{\Delta}(\boldsymbol{\pi}) = diag(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'$, $\boldsymbol{\Phi} = (\sigma_{ii'})$, and $m_{0s} = \sum_{h\in s} m_h^2/n$.

For large r, $\widehat{\boldsymbol{\pi}}$ follows the $k-1$ variate normal distribution with mean $\boldsymbol{\pi}$ and variance $\mathbf{Q}_{ms}/n$. Hence for testing the hypothesis $\mathbf{H}_0$: $\boldsymbol{\pi} = \boldsymbol{\pi}_0$ against alternative $\mathbf{H}_1$: $\boldsymbol{\pi} \neq \boldsymbol{\pi}_0$, the test statistic

$$X^2(\pi) = n(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0)'\mathbf{Q}_{\mathbf{ms}}^{-1}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0) \tag{19.6.6}$$

follows a chi-square distribution with $k-1$ df for large r when the null hypothesis $\mathbf{H}_0$ is true. Let $\overline{\lambda}_1, \ldots, \overline{\lambda}_{k-1}$ be the eigenvalues of $(\boldsymbol{\Delta}(\boldsymbol{\pi}))^{-1}\boldsymbol{\Phi}$, then we can write, following Rao and Scott (1981),

$$X^2(\pi) \cong \sum_{i=1}^{k-1}\{1 + (m_{s0} - 1)\overline{\lambda}_i\}Z_i^2 \tag{19.6.7}$$

Noting that $\boldsymbol{\Delta}(\boldsymbol{\pi}) - \boldsymbol{\Phi}$ is a nonnegative definite (Rao and Scott, 1981), so that $0 \leq \overline{\lambda}_i \leq 1$ for $i = 1,\ldots, k-1$, we find $X^2(\pi) \leq m_{0s}\sum_{i=1}^{k-1} Z_i^2 = m_{0s}\chi_{k-1}^2$. Hence treating $X^2(\pi)/m_{0s}$ as χ_{k-1}^2 under $\mathbf{H}_0$, we arrive at a conservative test. In case all m_h's are equal to m, $X^2(\pi)/m_{0s}$ reduces to $X^2(\pi)/m$.

19.6.1.1 A Simpler Model

Consider a special case of the model (Eq. 19.6.1) where

$$C_m\{\gamma_{hj}(i), \gamma_{hj'}(i')\} = \sigma_{ii'} = \begin{cases} \overline{\rho}\pi_i(1-\pi_i) & \text{for } i = i' \\ -\overline{\rho}\pi_i\pi_j & \text{for } i \neq i' \end{cases}$$

In this situation $\mathbf{\Phi} = \overline{\rho}\mathbf{\Delta}(\boldsymbol{\pi})$ and the asymptotic distribution of $X^2(\pi)/\{1 + (m_{s0} - 1)\overline{\rho}\}$ under $\mathbf{H}_0$ is an exact chi-square distribution with $k - 1$ df.

19.6.2 Brier Model

Brier (1978) considered two-stage sampling where the sample sizes for the ssu's m_h's are selected independently. Let m_{hi} be the number of sampled units from the hth fsu that belong to the ith category with $\sum_{i=1}^{k} m_{hi} = n_h$. Brier (1978) assumed that the vector $\mathbf{m_h} = (m_{h1}, \ldots, m_{hk-1})'$ follows a multinomial distribution with parameter $\mathbf{p_h} = (p_{h1}, \ldots, p_{h,k})'$, $0 < p_{hj} < 1, \sum_{j=1}^{k} p_{hj} = 1$. The parameter vector $\mathbf{p_h}$ is assumed to be distributed as the Dirichlet distribution with probability density function

$$f(\mathbf{p_h}|\boldsymbol{\pi}, \nu) = \frac{\Gamma\nu}{\prod_{i=1}^{k}\Gamma(\nu\pi_i)} \prod_{i=1}^{k} p_{hi}^{\nu\pi_i - 1}; \quad \nu, \pi_i > 0, \quad \sum_{i=1}^{k}\pi_i = 1 \tag{19.6.8}$$

Rao and Scott (1981) showed that for $m_h = m$, $E_m(\mathbf{m}_h) = m\boldsymbol{\pi}$, and $E_m(\mathbf{m}_h - m\boldsymbol{\pi})(\mathbf{m}_h - m\boldsymbol{\pi})' = \frac{m(\nu + m)}{\nu + 1}\mathbf{\Delta}(\boldsymbol{\pi})$, where $\mathbf{\Delta}(\boldsymbol{\pi})$ is given in Eq. (19.6.5). Further noting that the mean and variance of $\mathbf{n} = (n_1, \ldots, n_{k-1})'$ as $E(\mathbf{n}) = n\boldsymbol{\pi}$ and $V(\mathbf{n}) = n[1 + (m-1)\overline{\rho}]\mathbf{\Delta}(\boldsymbol{\pi})$ with $\overline{\rho} = 1/(\nu + 1)$, we find that under H_0, the asymptotic distribution of the modified statistic $\frac{(\nu + 1)}{\nu + m}X^2(\pi)$ is chi-square with $k - 1$ df. Rao and Scott (1981) also extended Brier's results to multistage sampling when the ssu's of unequal sizes are selected by SRSWOR method.

19.7 CONCLUDING REMARKS

Categorical data analysis is used extensively in analyzing survey data. Practitioners use standard statistical packages such as SPSS, BMDP, SAS, etc. to compute chi-square test statistics for goodness of fit, tests of independence, and homogeneity. The software packages give erroneous results

if the data are collected through a complex survey design, as software computations are based on SRSWR sampling. Several methods of analyzing categorical data for complex survey designs are available in the literature. The use of Wald statistics and modifications of standard chi-square statistics are popular. The Wald statistic has limitations in that it requires estimation of covariances that are not generally available from the published report. Furthermore, if the number of cells is large but the cell frequencies are not large enough, the estimates of variance and covariances are unstable. The first-order correction proposed by Rao and Scott (1981) is quite effective because it needs only estimation of variances, while second-order correction needs not only estimation variances but also covariances. Testing of significance of goodness of fit of the log-linear model based on a complex survey data can be performed following the procedure outlined by Rao and Scott (1981). Hidiroglou and Rao (1987) provided with practical application of chi-square test for goodness of fit, homogeneity, and independency for the Community Health Survey data collected though complex survey designs. For further information, readers are referred to the works of Rao and Scott (1987), Rao and Thomas (1988), Roberts et al. (1987), and Fay (1985), among others.

19.8 EXERCISES

19.8.1 In a survey conducted in 2010, the district of Gaborone was stratified into 20 strata each containing 20 EAs. From each of the EAs, a sample of two EAs was selected by probability proportional to size without replacement (PPSWOR) sampling scheme taking the number of people in the EA as measure of size variable. From each of the selected EAs 30 individuals were selected by SRSWOR method. The sampled 1200 individuals were classified into three age groups. The following table gives estimated proportions, last 2001 census proportion and the estimated variance—covariance matrix of the estimated proportions for the survey. Test if there is any change in the distribution of age between the periods 2010 and 2001 using (i) Pearsonian chi-square, (ii) modified chi-square, (iii) Wald statistic, (iv) Rao—Scott first- and second-order corrections, (v) Fellegi, and (vi) F-corrected to Wald statistic.

Age	Estimated proportion $\hat{p}_i$	Census proportion
Below 20	0.525	0.410
21—60	0.355	0.490
61 and above	0.120	0.100

Variance–covariance matrix

$$= \widehat{\mathbf{V}}/1200 = 10^{-5}\begin{pmatrix} 12.500 & -3.384 & -9.115 \\ & 8.950 & -5.565 \\ & & 14.680 \end{pmatrix}$$

19.8.2 Suppose a sample of 1800 individuals is selected from the district Francistown using a sampling design similar to that in Exercise 19.8.1. The age distribution and variance covariance matrix of the estimated cell proportions are given as follows:

Age	Estimated proportion $\hat{p}_i$
Below 20	0.525
21–60	0.355
61 and above	0.120

Variance–covariance matrix

$$= \widehat{\mathbf{V}}/1800 = 10^{-5}\begin{pmatrix} 15.750 & -3.384 & -12.365 \\ & 25.500 & -5.566 \\ & & 17.931 \end{pmatrix}$$

Test if there is any significant difference in age distribution between the district Gaborone (given in Exercise 19.8.1) and Francistown using (i) modified chi-square and (ii) Wald statistic.

19.8.3 A sample of 1500 households was selected by a complex survey design and classified according to income and education (head of the house hold). The estimated cell probabilities are given as follows:

	Income group			
Education	**Poor**	**Middle**	**High**	**Total**
Nil	0.15	0.10	0.05	0.30
Primary	0.10	0.08	0.02	0.20
Secondary	0.10	0.15	0.05	0.30
Tertiary	0.02	0.08	0.10	0.20
Total	0.37	0.41	0.22	

Fit an appropriate log-linear model on the estimated cell probabilities and use the model test if income depends on the level of education.

19.8.4 A sample of 6000 factory employees was selected using a complex survey design and classified according to gender and income (monthly salary). The estimated cell probabilities and deff's (in brackets) are given in the following table:

Gender	Income level				
	Below $1000	**$1001 to $2500**	**$2501 to $5000**	**$5000 and above**	**Total**
Male	0.215	0.15	0.08	0.012	0.457
	(1.82)	(0.75)	(2.75)	(4.32)	(1.854)
Female	0.325	0.175	0.038	0.005	0.543
	(1.75)	(1.95)	(1.78)	(5.86)	(2.10)
Total	0.540	0.325	0.118	0.017	
	(1.85)	(2.85)	(3.05)	(2.15)	

Test if there is any discrimination of salaries between male and female workers using modified chi-square and Bonferroni test.

19.8.5 The sample of 600 employees is divided at random into three groups. For each group the estimated cell proportions are given in the following table:

Gender	Income level				
	Below $1000	**$1001 to $2500**	**$2501 to $5000**	**$5000 and above**	**Total**
Group 1					
Male	0.265	0.12	0.12	0.005	0.51
Female	0.215	0.135	0.137	0.003	0.49
Total	0.48	0.255	0.257	0.008	
Group 2					
Male	0.218	0.21	0.15	0.01	0.588
Female	0.155	0.125	0.127	0.005	0.412
Total	0.373	0.335	0.277	0.015	
Group 3					
Male	0.258	0.222	0.12	0.02	0.62
Female	0.125	0.095	0.15	0.01	0.38
Total	0.383	0.317	0.27	0.03	

Test if the distribution of the hypothesis of dependency of income on gender using (i) modified chi-square, (ii) Wald statistic, and (iii) Bonferroni test.

19.8.6 The HIV status (estimated cell probabilities with the deff's in brackets) of two districts based on large-scale surveys is given in the following table:

Gender	HIV status		
	Positive	**Negative**	**Total**
District 1			
Male	0.10	0.36	0.46
	(1.85)	(0.98)	(1.98)
Female	0.16	0.38	0.54
	(2.15)	(1.75)	(2.08)
Total	0.26	0.74	
	(2.12)	(1.95)	
District 2			
Male	0.06	0.40	0.46
	(1.85)	(1.75)	(1.08)
Female	0.11	0.43	0.54
	(3.25)	(1.65)	(2.15)
Total	0.27	0.83	
	(2.02)	(1.75)	

Test if there is any significant difference in the HIV infection pattern between the two districts.