J.Statistics
1995(1), 11-24

ASYMPTOTICALLY ROBUST HOMOGENEITY TESTS UNDER COMPLEX SURVEYS-I

BY

Muhammad Khalid Pervaiz

ABSTRACT

The superpopulation approach with unrestrictive assumptions is adopted. Asymptotically robust tests for homogeneity of variances; i,e. standard error, grouping and jackknife for cluster samples form finite populations consisting on separate clusters are obtained. The test statistics are extended for stratified cluster samples as well.

Keywords: Asymptotically robust, superpopulation, cluster samples, stratified cluster samples, likelihood ratio, Taylor expansion, central limit law, complex samples, consistent, stratum.

1 INTRODUCTION

The first approach to the problem of testing the equality of variances under normality was made by Neyman and Pearson (1931) using the likelihood ratio statistic. Bartlett (1937) suggested modifications to the likelihood ratio test which improves the approximation to the chi-square. Further refinements were discussed by Bishop and Nair (1939), Hartley (1940) and Box (1949). Plackett (1946) is a good review paper. Kendall & Stuart (1967 p. 465-69; 1968 p. 97-105) and Plackett (1960, Chapter 5) are also worth seeing. Cochran (1941) and Hartley (1950) introduced methods for the purpose. Pearson (1931), Geary (1947) Finch (1950), Gayen (1950) and Box (1953)

Department of Statistics, Government College, Lahore

$e = 1,2,\ldots,m_{ic}$

$\overline{X}_{i..} = \frac{1}{N_{oi}} \sum_{c=1}^{N_i} \sum_{e=1}^{m_{ic}} x_{ice}$ Mean of finite population

$\overline{x}_{i..} = \frac{1}{n_{oi}} \sum_{c=1}^{n_i} \sum_{e=1}^{m_{ic}} x_{ice}$ Sample mean

$S_i^2 = \frac{1}{N_{oi}} \sum_{c=1}^{N_i} \sum_{e=1}^{m_{ic}} (x_{ice} - \overline{X}_{i..})^2$ Variance of finite population

$s_i^2 = \frac{1}{n_{oi}} \sum_{c=1}^{n_i} \sum_{e=1}^{m_{ic}} (x_{ice} - \overline{x}_{i..})^2$ Sample variance

2.2 Sampling design

The n_i clusters are selected form N_i clusters by simple random sampling techinque. The samples are selected independently from both finite populations. All subunits within each selected cluster are included in a sample.

2.3 Null hypothesis to be tested

The superpopulation approach with unrestrictive assumptions is adopted; e.g. Pervaiz (1989). It is assumed that x_{ice} are random variables, which implies that S_i^2 are also random variables. Therefore it is assumed that S_i^2 coverge to σ_i^2 as the population size N_{oi} goes to infinity. The finite populations with variances S_i^2's may be assumed to be random samples from infinite populations called superpopulations with variances σ_i^2. Then hypothesis of interest is:

$$H_{o1}: \sigma_1^2 = \sigma_2^2 \qquad vs \quad H_{A1}: \sigma_1^2 \neq \sigma_2^2$$

The approach to inference adopted and consisdered suitable is the p-distribution making use of the asymptotic arguments.

confirmed that the tests are non-robust against non-normality of the parent populations.

For non-normal populations Box(1953) intoduced grouping test for the equality of variances. Miller (1968) provided justification for the use of jackknife for testing hypothesis on variances.

The aim of this paper is to introduce a homogeneity test based on Taylor expansion estimation of variance under cluster sampling and stratified cluster sampling designs when finite populations consist of separate clusters; and to obtain grouping and jackknife tests under the said sampling designs. We omit the finite population corrections form the estimators of variance of sample variance, (e.g. Cochran, 1977, p.39).

2 ASYMPTOTICALLY ROBUST TEST STATISTICS UNDER CLUSTER SAMPLING

2.1 Notation

The suffix i denotes the finite population, i=1,2.

N_i	Number of clusters in finite population.
n_i	Number of clusters in sample.
m_{ic}	Observations in c-th cluster. c=1,2,.....,n_i.
$N_{oi} = \sum_{c=1}^{N_i} m_{ic}$	Finite population size.
$n_{oi} = \sum_{c=1}^{n_i} m_{ic}$	Sample size.
x_{ice}	e-th observation of c-th cluster.

$$z_{ice} = (x_{ice} - \bar{x}_{i..})^2$$

For proof see Pervaiz (1989 a)

Substituting $\hat{\Gamma}_1$ and $\hat{\Gamma}_2$, X_{TE} is approximately distributed as $t_{(n_1+n_2-2)}$ under H_{o1}

(b) GROUPING

The parent sample is divided into number of random groups, $n_i' = {}^{n_i}/_{L}$ (It is assumed that n_i are divisible by L). The variances for each group of clusters are computed and treated as asymptotically normal with equal means and variances. That being so under H_{o1}:

$$X_G = \frac{\bar{s}_1^2 - \bar{s}_2^2}{\sqrt{\frac{\hat{\Gamma}_1^+}{n_1'} + \frac{\hat{\Gamma}_2^+}{n_2'}}}$$

is approximately distributed as $t_{(n_1'+n_2'-2)}$. Where

$$\hat{\Gamma}_i^+ = \frac{1}{n_i' - 1}\sum_{g=1}^{n_i'}(s_{i,g}^2 - \bar{s}_i^2)^2 \qquad \text{and}$$

$$\bar{s}_i^2 = \frac{1}{n_i'}\sum_{g=1}^{n_i'} s_{i,g}^2$$

The asymptotic distribution of X_G^2 is χ_1^2 under H_{o1}.

(c) JACKKNIFE

Brillinger (1966) McCarthy (1966), and Mellor (1973) have suggested the application of jackknife with complex sampling design. Extensive discussion of jackknife method is given in Gray and Schucany (1972), Efron (1982) and Wolter (1985).

2.4 Properties of sample variances

Fuller (1975) gives central limit laws for complex samples with large sample sizes. Thus:

$$n_i^{\frac{1}{2}}(s_i^2 - \sigma_i^2) \xrightarrow{dist.} N(o, \Gamma_i) \text{ as } n_i \to \infty \qquad (2.4.1)$$

Furthermore s_1^2 and s_2^2 are independent.

2.5 Description of asympototically robust tests

(a) STANDARD ERROR

From (2.4.1) under H_{o1} the test statistic:

$$X_{TE} = \frac{s_1^2 - s_2^2}{\sqrt{\frac{\hat{\Gamma}_1}{n_1} + \frac{\hat{\Gamma}_2}{n_2}}}$$

is approximately distributed as $t_{(n_1+n_2-2)}$, if $\hat{\Gamma}_1$ and $\hat{\Gamma}_2$ are consistent estimates of the asymptotic variances of s_1^2 and s_2^2, respectively. The X_{TE}^2 is asymptotically distributed as χ_1^2 under H_{o1}. Theorem 1 provides the consistent estimates of Γ_i.

Theorem 1

The Taylor expansion estimate of Γ_i is:

$$\hat{\Gamma}_i = \frac{1}{n_i - 1}\sum_{c=1}^{n_i}\left(\bar{z}_{ic} - \bar{\bar{z}}_i\right)^2, \text{ where}$$

$$\bar{z}_{ic} = \frac{1}{m_{ic}}\sum_{e=1}^{m_{ic}} z_{ice};$$

$$\bar{\bar{z}}_i = \frac{1}{n_{oi}}\sum_{c=1}^{n_i}\sum_{e=1}^{m_{ic}} z_{ice} \quad \text{and}$$

of sample.

m_{ihc} Observations in c-th cluster in h-th stratum.

$N_i = \sum_{h=1}^{H_i} N_{ih}$ Total number of clusters.

$n_i = \sum_{h=1}^{H_i} n_{ih}$ Number of clusters in a sample.

$N_{oih} = \sum_{c=1}^{N_{ih}} m_{ihc}$ Number of observations in h-th stratum.

$n_{oih} = \sum_{c=1}^{n_{ih}} m_{ihc}$ Number of observations in a sample from h-th stratum.

$N_{oi} = \sum_{h=1}^{H_i} N_{oih} = \sum_{h=1}^{H_i} \sum_{c=1}^{N_{ih}} m_{ihc}$ Finite population size.

$n_{oi} = \sum_{h=1}^{H_i} n_{oih} = \sum_{h=1}^{H_i} \sum_{c=1}^{n_{ih}} m_{ihc}$ Sample size.

$W_{ih} = \frac{N_{oih}}{N_{oi}}$ Stratum weight.

x_{ihce} e-th observation of c-th cluster of h-th stratum. $e = 1,2,\ldots\ldots, m_{ihc}$

$\overline{X}_{ih..} = \frac{1}{N_{oih}} \sum_{c=1}^{N_{ih}} \sum_{e=1}^{m_{ihc}} x_{ihce}$ Mean of h-th stratum.

$\overline{x}_{ih..} = \frac{1}{n_{oih}} \sum_{c=1}^{n_{ih}} \sum_{e=1}^{m_{ihc}} x_{ihce}$ Mean of h-th stratum of sample.

$\overline{X}_{i...} = \frac{1}{N_{oi}} \sum_{h=1}^{H_i} \sum_{c=1}^{N_{ih}} \sum_{e=1}^{m_{ihc}} x_{ihce}$ Mean of finite population.

$\overline{x}_{i...} = \sum_{h=1}^{H_i} \frac{W_{ih}}{n_{oih}} \sum_{c=1}^{n_{ih}} \sum_{e=1}^{m_{ihc}} x_{ihce}$ Weighted mean of sample. (3.1.1)

$S_i^2 = \frac{1}{N_{oi}} \sum_{h=1}^{H_i} \sum_{c=1}^{N_{ih}} \sum_{e=1}^{m_{ihc}} (x_{ihce} - \overline{X}_{i...})^2$ Variance of finite population.

The primary sampling units are comprised of clusters of elementary units. Let

$$s_{i,c}^2 = n_i s_i^2 - (n_i - 1)s_{i-c}^2$$

The s_{i-c}^2 are sample variances by using $n_i - 1$ clusters with c-th cluster omitted. The jackknife estimators are the average of $s_{i,c}^2$ are:

$$s_i^{2*} = n_i s_i^2 - \frac{n_i - 1}{n_i}\sum_{c=1}^{n_i} s_{i-c}^2 \qquad (2.5.1)$$

The $s_{i,c}^2$ are approximately independent and have asymptotically equal means and variances. That being so under H_{o1} the test statistic:

$$X_J = \frac{s_1^{2*} - s_2^{2*}}{\sqrt{\frac{\hat{\Gamma}_1^*}{n_1} + \frac{\hat{\Gamma}_2^*}{n_2}}}$$

is approximately distributed as $t_{(n_1+n_2-2)}$, Wolter (1985). Where

$$\hat{\Gamma}_i^* = \frac{1}{n_i - 1}\sum_{c=1}^{n_i}(s_{i,c}^2 - s_i^{2*})^2$$

The asymptotic distribution of X_J^2 is χ_1^2 under H_{o1}.

3. ASYMPTOTICALLY ROBUST TESTS UNDER STRATIFIED CLUSTER SAMPLING

3.1 Notation

The suffix i denotes the finite popultaion, i=1.2.

H_i	Number of strata.
N_{ih}	Total number of clusters in h-th stratum.
n_{ih}	Number of clusters in h-th stratum

is approximately distributed as $t_{(n_1+n_2-2)}$ under H_{o2}; if $\hat{V}_i$ are consistent estimators of V_i. The asymptotic distribution of X_{TE}^2 is χ_1^2 under H_{o2}. To estimate V_i by Taylor expansion method applying Theorem 1 define

$$z_{ihce} = (x_{ihce} - \bar{x}_{i...})^2;$$

$$\bar{z}_{ihc} = \frac{1}{m_{ihc}} \sum_{e=1}^{m_{ihc}} z_{ihce} \qquad \text{and}$$

$$\bar{z}_{ih} = \frac{1}{n_{ih}} \sum_{c=1}^{n_{ih}} \bar{z}_{ihc}$$

Then

$$\hat{V}_i = \sum_{h=1}^{H_i} (W_{ih})^2 \frac{1}{n_{ih}(n_{ih}-1)} \sum_{c=1}^{n_{ih}} (\bar{z}_{ihc} - \bar{z}_{ih})^2 \qquad (3.5.1)$$

Substituting $\hat{V}_1$ and $\hat{V}_2$ given as (3.5.1) X_{TE} is approximately distributed as $t_{(n_1+n_2-2)}$ under H_{o2}.

(b) GROUPING

The parent sample from h-th stratum is randomly divided into groups of size L, i.e.

$$n_{ih} = Ln_{ih}' \text{ for } L \geq 2$$

It is assumed that n_{ih} are divisible by L. It is assumed that variance for each group of clusters of h-th stratum, is asymptotically normal with equal means and variances. That being so the test statistic

$$X_G = \frac{\bar{s}_1^2 - \bar{s}_2^2}{\sqrt{\hat{V}_1^+ + \hat{V}_2^+}}$$

$$s_i^2 = \sum_{h=1}^{H_i} \frac{W_{ih}}{n_{oih}} \sum_{c=1}^{n_{ih}} \sum_{e=1}^{m_{ihc}} (x_{ihce} - \bar{x}_{i...})^2$$ Weighted sample variance.

3.2 Sampling Design

The n_{ih} clusters are drawn by simple random sampling from N_{ih} clusters. Within each selected cluster all subunits are included in a sample. The samples are independent from both populations.

3.3 Null hypothesis to be tested

It is assumed that x_{ihce} are random variables, which implies that S_i^2's are also random variables. Furthermore it is assumed that S_i^2 converge to σ_i^2 and N_{oi} becomes larger. Thus the finite populations with variances S_i^2's may be viewed as sample from infinite populations called superpopulations with variances σ_i^2. The hypothesis of interest is:

$$H_{02}: \sigma_1^2 = \sigma_2^2 \quad vs \quad H_{A2}: \sigma_1^2 \neq \sigma_2^2$$

3.4 Properties of sample variances

From (2.4.1)

$$n_i^{\frac{1}{2}}(s_i^2 - \sigma_i^2) \xrightarrow{dist.} N(0, \Gamma_i) \; as \; n_i \to \infty \qquad (3.4.1)$$

Furthermore s_1^2 and s_2^2 are independent. Denote $V_i = n_i^{-1}\Gamma_i$

3.5 Description of asymptotically robust tests

(a) STANDARD ERROR

From (3.4.1) the test statistic:

$$X_{TE} = \frac{s_1^2 - s_2^2}{\sqrt{\hat{V}_1 + \hat{V}_2}}$$

4. CONCLUDING REMARKS

Obviously these tests can be extended to find the confidence intervals for two variances when finite population consist on separate clusters.

is approximately distributed as $t_{(n_1+n_2-2)}$ under H_{o2}. Where

$$\hat{V}_i^+ = \sum_{h=1}^{H_i} \frac{1}{n_{ih}(n_{ih}-1)} \sum_{g=1}^{n_{ih}} (s_{ih,g}^2 - \bar{s}_{ih}^2)^2 \quad \text{and}$$

$$\bar{s}_i^2 = \sum_{h=1}^{H_i} \frac{1}{n_{ih}} \sum_{g=1}^{n_{ih}} s_{ih,g}^2$$

The asymptotic distribution of X_G^2 is χ_1^2 under H_{o2}.

(c) JACKKNIFE

Following Wolter (1985) Let

$$s_{ih,c}^2 = \{H_i(n_{ih}-1)+1\}s_i^2 - H_i(n_{ih}-1)\, s_{ih-c}^2$$

Where s_{ih-c}^2's are variances after deleting *(h,c)* th cluster. The jackknife estimators are the average of $s_{ih,c}^2$'s, i.e.

$$s_{ih}^{2*} = \frac{1}{n_{ih}} \sum_{c=1}^{n_{ih}} s_{ih,c}^2$$

The $s_{ih,c}^2$'s are approximately independent and have asympotically equal means and variances. That being so the test statistic:

$$X_J = \frac{s_1^{2*} - s_2^{2*}}{\sqrt{\hat{V}_1^* + \hat{V}_2^*}}$$

is approximately distributed as $t_{(n_1+n_2-2)}$ under H_{o2}. Where

$$\hat{V}_1^* = \sum_{h=1}^{H_i} \frac{1}{n_{ih}(n_{ih}-1)} \sum_{c=1}^{n_{ih}} (s_{ih,c}^2 - s_{ih}^{2*})^2 \quad \text{and} \quad s_i^{2*} \text{ are jackknife}$$

estimators based on full sample.

The asymptotic distrubution of X_J^2 is χ_1^2 under H_{o2}.

GEARY, R.C. (1947). *Testing for normality.* Biometrika 34, 209-42.

GRAY, H.L & SCHUCANY, W.R. (1972). *The Generalized Jackknife Statistics.* New York: Marcel Dekcer.

HARTLEY, H.O. (1940). *Testing the homogeneity of a set of variances.* Biometrika, 31, 249-55.

HARTLEY, H.O. (1950). *The use of range in Analysis of variance.* Biometrika 37, 271-80.

KENDALL, M. & STUART, A. (1967). *The Advanced Theory of Statistics.* V. 2, 2nd ed. Charles Griffin & Company Limited. London & High Wycombe.

KENDALL, M & STUART, A. (1968). *The Advanced Theory of Statistics.* V.3, 3rd ed. Charles Griffin & Company Limited. London & High Wycombe.

McGARTHEY, P.J. (1966). *Replications. An Approach to the Analysis of Data from Complex Surveys.* Vital and Health Statistics Series 2, NO. 14. U.S. Department of Health, Education and Welfare, Washington: U.S. Government Printing Office.

MELLOR, R.W. (1973). *Subsample replication variance estimators.* Ph.D. Thesis. Harvard University.

MILLER, R.G., JR. (1968). *Jackknife variances.* Ann. Math. Statist. 39, 567-82.

NEYMAN, J. & PEARSON, E.S. (1931). *On the problem of k samples.* Bull. Acad. Polon. Sci. 3, 460.

PEARSON, E.S (1931). *Analysis of variance in cases of non-normal variation.* Biometrika 23, 114-33.

——— **(1931).** *Note on tests for normality.* Biometrika 22, 423-24.

REFERENCES

BARTLETT, M.S (1937). *Properties of Sufficiency and Statistical Tests.* Proc. Roy. Soc. A, 160, 268-81.

BISHOP, D.J. & NAIR, U.S. (1939). *A note on certain methods of testing for the homogeneity of a set of estimated variances.* J.R. Statist. Soc. Suppl. 6, 89-99.

BOX, G.E.P. (1949). *A general distribution theory for a class of likelihood criteria.* Biometrika 36, 317-46.

BOX, G.E.P. (1953). *Non Normality and Tests on Variances.* Biometrika 40, 318-35.

BRILLINGER, D.R. (1966). *The application of the jackknife to the analysis of sample surveys.* Commentary 8, 74-80.

COCHRAN, W.G. (1941). *The distribution of the largest of a set of estimated variances as a fraction of their total.* Ann. Eugen. 11, 47.

COCHRAN, W.G. (1977). *Sampling Techniques.* 3rd ed. John Wiley & Sons, New York.

EFRON, B. (1982). *The jackknife, the Bootstrap, and Other Resampling Plans,* Philadelphia: Society for Industrial and Applied Mathematics.

FINCH, D.J. (1950). *The effect of non-normality on the Z-Test, when used to compare the variances in two populations.* Biometrika 37, 186-89.

FULLER, W.A. (1975). *Regression analysis for sample surveys. Sankhya.* The Indian Journal of Statistics 37 C, 117-32.

GAYEN, A.K. (1950). *The distribution of variance ratio in random samples of any size drawn from non-normal universes.* Biometrika 37, 236-55.