DISTRIBUTION OF χ^2- ANALOGUE FOR NORMAL POPULATION WITH CLASS INTERVALS DEFINED IN TERMS OF SAMPLE MEDIAN

By A. R. Roy and S. G. Mohanty

Indian Council of Agricultural Research, New Delhi

Introduction

In the usual formulation of χ^2 statistic in the uniparametric case, the different classes in which the sample observations are grouped are fixed arbitrarily in advance and the sample is replaced by its frequency classification as determined by these classes. In doing this, a part of the information supplied by the individual sample observations is lost. An attempt was made by Lehman and Chernoff[1] to restore this information by using the maximum likelihood estimate $\hat{\theta}$ of the parameter θ in the population form. They obtained an estimate of the probability of a single observation falling in the i-th class interval S_i, $(i = 1, 2, \ldots, k)$ by substituting $\hat{\theta}$ for θ in this probability and constructed the statistic

$$\hat{R}_n = \sum_{i=1}^{k} \frac{(m_i - n\hat{p}_i)^2}{n\hat{p}_i} \tag{1}$$

where m_i is the number of sample observations falling in S_i and n is the number of sample observations. The asymptotic distribution of $\hat{R}_n$ as $n \to \infty$ is given as

$$\sum_{i=1}^{k-2} y_i^2 + \lambda y^2_{k-1} \tag{2}$$

where

$$y_i\text{'s } (i = 1, 2, \ldots, k-1)$$

are independently and normally distributed with mean 0 and variance 1, and λ is a constant between 0 and 1. It is to be mentioned that λ may depend on θ and therefore the distribution is not known to that extent.

The arbitrariness in the definition of the class intervals giving the frequency structure of the sample is retained in $\hat{R}_n$. This was removed by Roy[2] by defining the class intervals S_i's in terms of the sample observations. Let $\hat{\theta}$ be an estimate of θ based on a random sample of n independent observations

$$x_1, x_2, \ldots, x_n$$

with the property that there exists a function f such that

$$\left.\begin{aligned} &\hat{\theta} - \theta = \frac{1}{n}\sum_{\alpha=1}^{n} f(x_\alpha) + \epsilon \\ &\text{where} \\ &\text{(i)}\ \ \epsilon = o_p\left(\frac{1}{\sqrt{n}}\right), \\ &\text{(ii)}\ \ E[f(x)] = 0 \text{ for all } \theta, \\ &\text{and} \\ &\text{(iii) Var. } [f(x)] \text{ is finite for all } \theta. \end{aligned}\right\} \tag{3}$$

Then S_i is defined in terms of the estimate $\hat{\theta}$. Let

$$p_i^*(\hat{\theta}) = p_r\{x \epsilon S_i(\hat{\theta}) / \theta = \hat{\theta}\}$$

where x is independent of the observations in the original sample, *i.e.*, $p_i^*(\hat{\theta})$ is the probability of x lying in $S_i(\hat{\theta})$ when θ takes the value $\hat{\theta}$.

Then constructing the statistic

$$R_n^* = \sum_{i=1}^{k} \frac{\{m_i - np_i^*(\hat{\theta})\}^2}{np_i^*(\hat{\theta})} \tag{4}$$

where m_i is the number of sample observations x_α, $\alpha = 1, 2, \ldots, n$ falling in $S_i(\hat{\theta})$, Roy has proved[2] that the asymptotic distribution of R_n^* is given by

$$\sum_{i=1}^{k-2} y_i^2 + \lambda' y_{k-1}^2 \tag{5}$$

where y_i's $(i = 1, \ldots, k-1)$ are independently and normally distributed with mean 0 and variance 1 and λ' is a constant lying between 0 and 1. It has been seen that λ' can be made independent of θ when

θ is either a location or a scale parameter, by suitably defining S_i's in terms of $\hat{\theta}$.

It may be pointed out that the above result holds, subject to the condition that $\hat{\theta}$ is expressed as at (3). Such estimates exist, *e.g.*, maximum likelihood estimates. In this paper the asymptotic distribution of R_n^* as defined at (4) is derived by defining the class intervals S_i's ($i = 1, 2, \ldots, k$) in terms of the sample median which is an estimate of θ not satisfying the property (3). This is done for the normal population with unknown mean, and known variance. It is seen that the asymptotic distribution of R_n^* is not of the form as at (5) except in a special case, but it is still independent of the mean.

2. Asymptotic Distribution of R_n^*

Consider any normally distributed population $N(\theta, 1)$ with unknown mean θ and variance unity with three groups defined in terms of the median $\tilde{x}$ of a random sample of n observations

$$x_\alpha\text{'s},\ \alpha = 1, 2, \ldots, n.$$

Then the asymptotic distribution[3] of $\tilde{x}$ is normal with mean θ and variance

$$\frac{\pi}{2n}, \quad i.e., \quad \tilde{x} - \theta = 0_p\left(\frac{1}{\sqrt{n}}\right).$$

Let the range of x, *viz.*, $(-\infty, \infty)$ be divided into the following three class intervals

$$(-\infty,\ \tilde{x} + a),\ (\tilde{x} + a,\ \tilde{x} + b) \text{ and } (\tilde{x} + b,\ \infty)$$

where

$$a < o < b.$$

We then obtain

$$\left.\begin{aligned} p_1^*(\tilde{x}) &= \pi_1 \quad \text{where } \pi_1 = F(a) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a} e^{-t^2/2}\,dt. \\ p_2^*(\tilde{x}) &= \pi_2 \quad ,, \quad \pi_2 = F(b) - F(a) \\ \text{and} \\ p_3^*(\tilde{x}) &= \pi_3 \quad ,, \quad \pi_3 = 1 - F(b). \end{aligned}\right\} \quad (6)$$

It is clear that

$$\pi_1 + \pi_2 + \pi_3 = 1. \tag{7}$$

Again, we have by[2]

$$m_1 = \sum_{\alpha=1}^{n} b_1(x_\alpha) + \frac{n}{\sqrt{2\pi}} (\tilde{x} - \theta)\, e^{-a^2/2} + o_p(\sqrt{n}); \tag{8}$$

where

$$b_1(x_\alpha) = \begin{cases} 1 & \text{when } x_\alpha \leqslant \theta + a \\ 0 & \text{otherwise} \end{cases} \qquad \text{for } \alpha = 1, 2, \ldots, n.$$

Now

$$\frac{m_1 - np_1^*(\tilde{x})}{\sqrt{n}} = \frac{1}{\sqrt{n}} \left[\sum_{\alpha=1}^{n} \{b_1(x_\alpha) - \pi_1\} + ne^{-a^2/2} (\tilde{x} - \theta) \right] + o_p(1). \tag{9}$$

Then the asymptotic variance of

$$\frac{m_1 - np_1^*(\tilde{x})}{\sqrt{n}}$$

is

$$\pi_1(1-\pi_1) + \frac{e^{-a^2}}{4} + \sqrt{\frac{2}{\pi}}\, e^{-a^2/2}\, E\left\{ (\tilde{x} - \theta) \sum_{\alpha=1}^{n} b_1(x_\alpha) \right\}. \tag{10}$$

To evaluate the last term in (10) consider a function

$$g_1(x - \theta) = b_1(x) = \begin{cases} 1 & \text{when } x - \theta \leqslant a \\ 0 & \text{otherwise} \end{cases}$$

and let

$$U_1 = \frac{1}{n} \sum_{\alpha=1}^{n} g_1(x_\alpha - \theta).$$

We now obtain from the joint distribution of $\sqrt{n}\,(\tilde{x} - \theta)$ and $\sqrt{n}\,(U_1 - \pi_1)$ as discussed[4] in and obtain therefrom

$$E\left\{ \tilde{x} - \theta) \sum_{\alpha=1}^{n} b_1(x_\alpha) \right\} = -\sqrt{\frac{\pi}{2}}\, \pi_1 + 0_p\left(\frac{1}{n}\right) \tag{11}$$

and therefore (10) gives

$$\text{Asymp. Var.}\left(\frac{(m_1 - np_1^*(\tilde{x})}{\sqrt{n}}\right) = \pi_1(1 - 2\pi_1) + \tfrac{1}{4}\mu_1^2 \qquad (12)$$

where

$$\mu_1 = e^{-a^2/2} - 2\pi_1.$$

Similarly we can obtain

$$\text{Asymp. Var.}\left(\frac{m_2 - np_2^*(\tilde{x})}{\sqrt{n}}\right) = \pi_1(1 - 2\pi_1) + \pi_3(1 - \pi_3) + \tfrac{1}{4}\mu_2^2 \qquad (13)$$

where

$$\mu_2 = e^{-b^2/2} - e^{-a^2/2} - 2\pi_3 + 2\pi_1$$

and

$$\text{Asymp. Var.}\left(\frac{m_3 - np_3^*(\tilde{x})}{\sqrt{n}}\right) = \pi_3(1 - 2\pi_3) + \tfrac{1}{4}\mu_3^2 \qquad (14)$$

where

$$\mu_3 = 2\pi_3 - e^{-b^2/2}.$$

It is obvious that

$$\mu_1 + \mu_2 + \mu_3 = 0. \qquad (15)$$

Proceeding on the same line as for asymptotic variances, the asymptotic covariance of

$$\frac{m_1 - np_1^*(\tilde{x})}{\sqrt{n}} \quad \text{and} \quad \frac{m_2 - np_2^*(\tilde{x})}{\sqrt{n}}$$

is obtained as

$$\text{Asymp. Cov.}\left(\frac{m_1 - np_1^*(\tilde{x})}{\sqrt{n}}, \frac{m_2 - np_2^*(\tilde{x})}{\sqrt{n}}\right) = -\pi_1(1 - 2\pi_1) + \tfrac{1}{4}\mu_1\mu_2, \qquad (16)$$

the asymptotic covariance of

$$\frac{m_1 - np_1^*(\tilde{x})}{\sqrt{n}} \quad \text{and} \quad \frac{m_3 - np_3^*(\tilde{x})}{\sqrt{n}}$$

as

$$\text{Asymp. Cov.}\left(\frac{m_1 - np_1^*(\tilde{x})}{\sqrt{n}}, \frac{m_3 - np_3^*(\tilde{x})}{\sqrt{n}}\right) = \tfrac{1}{4}\mu_1\mu_3, \quad (17)$$

and the asymptotic covariance of

$$\frac{m_2 - np_2^*(\tilde{x})}{\sqrt{n}} \quad \text{and} \quad \frac{m_3 - np_3^*(\tilde{x})}{\sqrt{n}}$$

as

$$\text{Asymp. Cov.}\left(\frac{m_2 - np_2^*(\tilde{x})}{\sqrt{n}}, \frac{m_3 - np_3^*(x)}{\sqrt{n}}\right)$$
$$= -\pi_3(1 - 2\pi_3) + \tfrac{1}{4}\mu_2\mu_3. \quad (18)$$

Therefore, the asymptotic covariance function of

$$\left(\frac{m_1 - np_1^*(\tilde{x})}{\sqrt{n}}, \frac{m_2 - np_2^*(\tilde{x})}{\sqrt{n}}, \frac{m_3 - np_3^*(\tilde{x})}{\sqrt{n}}\right)$$

is given by

$$\sum = \begin{pmatrix} \pi_1(1-2\pi_1) + \tfrac{1}{4}\mu_1^2 & -\pi_1(1-2\pi_1) + \tfrac{1}{4}\mu_1\mu_2 & \tfrac{1}{4}\mu_1\mu_2 \\ -\pi_1(1-2\pi_1) + \tfrac{1}{4}\mu_1\mu_2 & \pi_1(1-2\pi_1) + \pi_3(1-2\pi_3) + \tfrac{1}{4}\mu_2^2 & -\pi_3(1-2\pi_3) + \tfrac{1}{4}\mu_2\mu_3 \\ \tfrac{1}{4}\mu_1\mu_3 & -\pi_3(1-2\pi_3) + \tfrac{1}{4}\mu_2\mu_3 & \pi_3(1-2\pi_3) + \tfrac{1}{4}\mu_3^2 \end{pmatrix} \quad (19)$$

where all rows and all columns add to 0.

The asymptotic distribution of R_n^* depends on the characteristic roots of the matrix $P^{-\frac{1}{2}} \Sigma P^{\frac{1}{2}}$ where

$$P = \begin{pmatrix} \pi_1 & 0 & 0 \\ 0 & \pi_2 & 0 \\ 0 & 0 & \pi_3 \end{pmatrix}. \quad (20)$$

The characteristic roots of $P^{-\frac{1}{2}} \Sigma P^{\frac{1}{2}}$ are the roots of the equation

$$|\Sigma - xP| = 0 \quad (21)$$

since P is non-singular. One root is obviously 0. The other two roots are the roots of the quadratic equation

$$Ax^2 - Bx + C = 0 \quad (22)$$

where

$$A = \pi_1\pi_2\pi_3,$$

$$B = 2\pi_1\pi_2{}^2\pi_3 + \pi_1\pi_3\,(\pi_1 - 2\pi_1{}^2 + \pi_3 - 2\pi_3{}^2) + \tfrac{1}{4}\pi_1\pi_2\pi_3\left(\frac{\mu_1{}^2}{\pi_1} + \frac{\mu_2{}^2}{\pi_2} + \frac{\mu_3{}^2}{\pi_3}\right),$$

and

$$C = \pi_1\pi_3\,(1 - 2\pi_1)\,(1 - 2\pi_3) + \tfrac{1}{4}\mu_3{}^2\pi_1\,(1 - 2\pi_1) + \tfrac{1}{4}\mu_1{}^2\pi_3\,(1 - 2\pi_3).$$

Equation (22) after some simplification can be expressed as

$$Ax^2 - \{A + C + \tfrac{1}{4}\,(\pi_1\mu_3 + \pi_3\mu_1)^2\}\,x + C = 0. \qquad (23)$$

Let the roots be λ_1 and λ_2. Then the asymptotic distribution of $R_n{}^*$ by^2 is given by

$$\lambda_1 y_1{}^2 + \lambda_2 y_2{}^2 \qquad (24)$$

where y_1 and y_2 are independently and normally distributed with mean 0 and variance 1. Since A, B, $C \geqslant 0$, both the roots of the equation (22) or (23) are non-negative and moreover, since

$$B = A + C + \tfrac{1}{4}\,(\pi_1\mu_3 + \pi_3\mu_1)^2 \geqslant A + C.$$

One root of (23) is $\geqslant 1$, equality holding when $B = A + C$, and the other < 1. But $B = A + C$, if and only if

$$\pi_1\mu_3 + \pi_3\mu_1 = 0 \qquad (25)$$

which is again true if and only if

$$-a = b.$$

This is the symmetrical case for which

$$\mu_1 = -\mu_3 = \mu^* \text{ (say)}$$

$$\pi_1 = \pi_3 = \pi^* \text{ (say)}.$$

In this case one of the two non-zero roots is 1 and the other is

$$\lambda^* = 1 - 2\pi^* + \frac{\mu^*}{2\pi^*}.$$

The asymptotic distribution of $R_n{}^*$ is therefore

$$y_1{}^2 + \lambda^* y_2{}^2 \qquad (26)$$

conforming to the case where $\hat{\theta}$ is of the form (3).

The case where $b > a > 0$ or $a < b < 0$ can be discussed analogously and it is seen that in these cases also one of the characteristic roots is zero, while the other two are positive, one being greater than 1 and the other less than 1. This confirms that the general form of the asymptotic distribution of R_n^* is as (24). It is interesting to note that one of the λ values in (24) is always greater than 1 in asymmetric cases.

3. Illustrations

Illustrations in support of the above results are given by actually computing the different roots in the particular cases where

$$a = -1, \qquad b = 0{\cdot}5 \qquad (a < 0 < b)$$

and

$$a = 0{\cdot}5, \qquad b = 1 \qquad (b > a > 0).$$

We get the two non-zero roots in the former case as

$$\lambda_1 = 1{\cdot}0558, \qquad \lambda_2 = 0{\cdot}6240$$

and in the latter as

$$\lambda_1 = 1{\cdot}0894, \qquad \lambda_2 = 0{\cdot}6225.$$

A generalisation of the problem by dividing the range into k intervals can be done likewise.

Summary

χ^2-statistic with classes based on random variables has been defined by A. R. Roy in Technical Report No. 1, Stanford University, 1955, entitled 'On χ^2-statistics with variable intervals' and its distribution has been considered when the random variables defining the classes follow certain properties. These properties are essentially those satisfied by the maximum likelihood estimates. An attempt has been made in this paper to find the distribution of the χ^2-statistic when the random variables defining the classes do not possess the above properties. This has been done by considering normal population with class intervals defined in terms of the sample median. The distribution so obtained is different from that based on the classes defined by the sample mean which follows the properties mentioned above.

REFERENCES

1. Chernoff, H. and Lehman, E. L. .. "The use of maximum likelihood estimates in test for goodness of fit," *Annals of Mathematical Statistics*, 1954, **25**.

2. Roy, A. R. .. "On statistics with variable intervals," *Technical Report No.* 1, Department of Statistics, Stanford University, Stanford, California, 1956.

3. Cramer, H. .. *Mathematical Methods of Statistics*, Princeton University Press, 1946.

4. Sukhatme, B. V. .. "Joint asymptotic distribution of the median and a U-Statistic," *Journal of the Royal Statistical Society*, 1957, **19 B** (1).