GOODNESS-OF-FIT TEST USING PAIRED χ^2 -VALUES

Osamu Miyatake* and Minoru Ichimura**

ABSTRACT

A defect inherent in the traditional goodness-of-fit χ^2 -test is pointed out, see Miyatake (1985), in which the number of classes cannot be fixed a priori. Let a population be divided into ℓ classes and $\chi_{\ell-1}^2$ be the χ^2 -value obtained from a hypothesis H_0 and observed frequency in each class. We assume that ℓ is even, and put together these classes two by two, getting a coarse division of k classes ($k = \ell/2$). Let χ_{k-1}^2 be the χ^2 -value obtained from this coarse division under the same H_0 and observed frequencies. Then we get the decomposition $\chi_{\ell-1}^2 = \chi_{k-1}^2 + \chi_{\ell-k}^2$, where $\chi_{\ell-k}^2$ is a χ^2 -variable of degree of freedom $\ell - k$. It will be shown that the use of the pair $(\chi_{k-1}^2, \chi_{\ell-k}^2)$ is more effective in the goodness-of-fit test, in comparison with the traditional method using the single $\chi_{\ell-1}^2$.

1. Introduction

As an example, let us consider the following problem: Independent n points are distributed on interval $[0, 1]$. First we divide this interval into ℓ parts. The hypothesis H_0 is that a point falls on the i -th subinterval with probability p_i . Let n_i be the observed number

of points in the i -th subinterval. Then we get the value of $\chi_{\ell-1}^2 = \sum_{i=1}^{\ell} (n_i - np_i)^2 / np_i$. On

the other hand, putting together the subintervals two by two in succession, we get χ_{k-1}^2 , where $k = \ell/2$. Let α be the significance level, and hereafter we use the letters z , x , y , z_α and x_α instead of $\chi_{\ell-1}^2$, χ_{k-1}^2 , $\chi_{\ell-k}^2$, $\chi_{\ell-1}^2(\alpha)$ and $\chi_{k-1}^2(\alpha)$, respectively. Then, according to the traditional method, hypothesis H_0 can be tested by using x as well as z . If, in this case, H_0 is always rejected by z when it is rejected by x , there would be no problem. This is, however, not the case. That is, the probability for the validity of the simultaneous inequalities $x > x_\alpha$ and $z < z_\alpha$ comes to a considerable amount, e.g., it comes to about 3% when $\alpha = 0.05$, irrespectively of the partition number ℓ . The purpose of the present paper is to remove this inconvenience as well as possible.

Variable z is, as shown in Appendix A, decomposed into two parts as

$$z = x + y. \quad (1)$$

Here $y = \chi_{\ell-k}^2$ is positive and obeys the χ^2 -distribution of degree of freedom $\ell - k$. As shown in §2, y is a sum of variance in each of k coarse subintervals. Eq. (1) means, therefore, that

* Wakamatudai

** Okayama University of Science, Okayama 700, Japan

Key words : Goodness-of-fit test; Chi-square tests.

the total variance z is decomposed into a sum of a between-class variance x and a within-class variance y . When the value of z is fixed, the point (x, y) can move freely on the straight line (1). Accordingly, the pair (x, y) contains more information about the distribution of points, in comparison with the single z . In the present paper, we apply the pair (x, y) to the goodness-of-fit test.

In §2, the decomposition (1) is given explicitly. In §3, our problem is explained by using the decomposition (1). In §4, we determine the critical region for the point (x, y) . In §5, the comparison between the traditional χ^2 -test and our method which uses the pair (x, y) will be made by using some typical distributions.

2. Decomposition of z

We tentatively assume that ℓ is even and put $k = \ell/2$. In Appendix B, we have shown that the nearer the value of k is to $(\ell + 1)/2$, the greater the information contained in the pair (x, y) is. We therefore take only the case $k = \ell/2$ into consideration in the text. In Eq. (1), z and x have the usual expressions. We write them as follows:

$$z = \sum_{s=1}^k \sum_{i=1}^2 (n_{si} - np_{si})^2 / np_{si}, \quad (2)$$

$$x = \sum_{s=1}^k (n_s - np_s)^2 / np_s. \quad (3)$$

Here suffix s refers to the coarse k -division, and suffix i to the fine ℓ -division, and we have put $p_s = p_{s1} + p_{s2}$ and $n_s = n_{s1} + n_{s2}$. The term y turns out to have the expression

$$y = \frac{1}{n} \sum_{s=1}^k \frac{1}{p_s} \left(\sqrt{p_{s1}/p_{s2}} n_{s2} - \sqrt{p_{s2}/p_{s1}} n_{s1} \right)^2. \quad (4)$$

A more general expression of y for arbitrary ℓ and k will be given in Appendix A. According to Cochran's theorem, see Cochran (1934), x and y are mutually independent. We prepare the x - y plane besides the current z -axis.

3. Detail of our problem

Let $f_n(\chi_n^2)$ be the probability density function (p.d.f.) of χ_n^2 , and $F(x, y)$ be the p.d.f. of point (x, y) on the x - y plane. Then we have

$$F(x, y) dx dy = P_r(x \leq X \leq x + dx, y \leq Y \leq y + dy) = f_{k-1}(x) f_{\ell-k}(y) dx dy,$$

and the whole domain of point (x, y) is the first quadrant on the x - y plane (Fig.1). We here introduce a new variable t by the transformation $x = tz$, $y = (1 - t)z$. Then we get the expression

$$F(x, y) dx dy = f_{t-1}(z) P(\ell, k, t) dt dz, \quad (5)$$

where

$$P(\ell, k, t) = \frac{1}{B(\frac{k-1}{2}, \frac{\ell-k}{2})} t^{\frac{k-1}{2}-1} (1-t)^{\frac{\ell-k}{2}-1}. \quad (6)$$

In this case, the first quadrant on the x - y plane is transformed onto the domain $0 \leq t \leq 1$, $0 \leq z < \infty$. If a point on the straight line $x + y = z$ starts from the point on the y -axis, moves along $x + y = z$, and then reaches the point on the x -axis, then the value of t starts from 0 and reaches 1, irrespectively of the value of z (Fig.1). Accordingly we have the relation

$$\iint_{x+y \geq z_\alpha} F(x, y) dx dy = \int_{z_\alpha}^{\infty} f_{\ell-1}(z) dz,$$

and the critical region $z \geq z_\alpha$ in the traditional method corresponds to the region $x + y \geq z_\alpha$ on the x - y plane.

We now consider the triangular domain B_α enclosed by the three straight lines $x = x_\alpha$, $y = 0$ and $x + y = z_\alpha$. Since the degree of freedom of x is smaller than that of z , the inequality $x_\alpha < z_\alpha$ always holds valid, so that the domain B_α is always formed. The falling of a point into B_α means that the hypothesis H_0 is rejected by x obtained from the coarse k -division and it can not be rejected by z obtained from the fine ℓ -division. And it turns out that the probability for $(x, y) \in B_{0.05}$ amounts to about 3%, irrespectively of the number ℓ of division. The purpose of the present paper is to remove this inconvenience as well as possible.

Fig. 1

4. Critical region on x - y plane

We consider the contour line C given by

$$C : F(x, y) = H. \quad (7)$$

Here positive constant H is determined by the condition

$$\iint_D F(x, y) dx dy = 1 - \alpha, \quad (8)$$

Fig. 2

where D is the domain enclosed by C . When $\ell \geq 8$, the surface $F(x, y)$ is unimodal and becomes maximum at the point $(x_0, y_0) = (k - 3, \ell - k - 2)$, and C is a convex closed curve. When $\ell = 4$ or 6 , the boundary of D consists of the curve C and at least one of the x - and y -axes (Fig.2(a),(b),(c)).

Let us first consider the case $\ell \geq 8$, and tentatively take the outside of C as the critical region R . Then, in general, R becomes narrow in the region where C approaches the both axes, and $F(x, y)$ becomes steep around there. We should exclude such narrow regions from the critical region. For the fluctuation of point (x, y) owing to the finiteness of sample size makes the test uncertain. We also exclude the neighbourhood of the origin $(0, 0)$, referring to the one-sided test used in the traditional χ^2 -test. We have, therefore, taken up the domain R_α determined in the following way as the critical region: R_α is the outside of domain D_α enclosed by a closed curve C_α which consists of both axes, vertical line $x = x_u$, horizontal line $y = y_u$ and contour line C as defined by (7) (Fig.3,(a)).

Here constant H_α is used instead of H . Constants x_u, y_u and H_α are determined so that the condition (8) may hold valid. The values of x_u, y_u and $H_{0.05}$ for $\ell = 4(2)18$ are given on Table 1.

As seen from Fig.3,(a), the non-critical region D_α juts out partially into the region $x + y \geq z_\alpha$, and recedes partially from the line $x + y = z_\alpha$. Accordingly it happens that the hypothesis cannot be rejected even when z is greater than z_α and it can be rejected even when z is smaller than z_α . This circumstance is shown in Fig.3,(b). In the inset (b), the ratio of the height of the shaded part to that of $f_{\ell-1}(z)$ shows the proportion of the rejection of H_0 for each z . That is, in our case, the proportion of the rejection increases gradually, as z increases.

Fig. 3

$\alpha = 0.05$			
$\ell \ k$	H_α	x_u	y_u
4 2	0.0030215	*	*
6 3	0.0023510	7.88166	*
8 4	0.0013075	10.3965	12.7616
10 5	0.0010781	12.1476	14.2044
12 6	0.0009157	13.8390	15.7292
14 7	0.0007999	15.4643	17.2460
16 8	0.0007123	17.0397	18.7437
18 9	0.0006436	18.5746	20.2194

Table 1

When $\ell = 6$, the contour line C determined by conditions (7) and (8) dose not close. In this case, the outside of curve C_α is taken as the critical region R_α , where C_α consists of the

Goodness-of-fit test using paired χ^2 -values

segment $x = x_u (0 \leq y \leq y_0)$ and a part of C . Constants x_u and H_α are to be determined so that conditions (7) and (8) may hold valid. When $\ell = 4$, both x_u and y_u do not appear, and we take C , H and D in (7) and (8) as C_α , H_α and D_α , respectively.

5. Power of test

The comparison between the traditional method and ours has been made by using some typical distributions on interval $[0, 1]$. We have taken up the following four families of p.d.f.:

$$\begin{aligned} p_1(x) &= \lambda e^{-\lambda x} / (1 - e^{-\lambda}), & p_2(x) &= \lambda(x - 0.5)^2 + 1 - \lambda/12, \\ p_3(x) &= \frac{1}{A} \frac{\lambda}{\lambda^2(x - 0.5)^2 + 1}, & p_4(x) &= \frac{1}{B} \frac{\lambda}{\lambda^2(x - 0.25)^2 + 1}, \end{aligned} \quad (9)$$

where λ is a parameter, and A and B are the normalization constants. Curves of these functions are shown on Fig.4. Special random numbers $\{r_i\}$ corresponding to these p.d.f.'s are generated through the relation $u_i = \int_0^{r_i} p_k(x) dx$, where u_i 's are uniform random numbers.

Fig. 4

For every fixed value of λ , we have executed the following test: For a sequence $\{r_i\}$, we set the hypothesis H_0 that the distribution is uniform. Interval $[0, 1]$ is divided into $\ell (= 12)$ equal parts, and $n (= 288)$ r_i 's have been used to get one z and one pair (x, y) . The total numbers N of z 's and (x, y) 's are 3200 each. We have fixed α at 0.05 and counted

the numbers N_R of points z and (x, y) which fall in the respective critical regions R_α . On Fig.5, two values of N_R/N for z and (x, y) are put down using respective notations, dot and double circle, for every value of λ . Here only the result for $p_1(x)$ is shown. The results for other p.d.f.'s in (9) have been omitted here, because of the similar tendencies. Our result shown on Fig.5,(b) is explained as follows: If the hypothesis H_0 were correct for each value of λ , the ratio N_k/N should take values nearly equal to 0.05. That is, its values should be in the vicinity of the horizontal line $N_k/N = 0.05$. Two dotted lines on its both sides indicate the limits of confidence level 95%. It is seen that, when $\lambda < 0.1$, hypothesis H_0 cannot be rejected by the traditional method as well as ours. When $\lambda \geq 0.1$, both methods can generally reject H_0 . In this case, however, the double circles are always above the corresponding dots. This tendency shows that the rejection occurs more frequently in our method than in the traditional one.

$$p_1(x) = \lambda e^{-\lambda x} / (1 - e^{-\lambda})$$

Fig. 5

6. Conclusion

It has been pointed out that the traditional goodness-of-fit χ^2 -test in which the number of classes can be fixed arbitrarily contains a defect inherent in it. We have removed this defect as well as possible. When a population is divided into ℓ parts, we get the value of $\chi^2_{\ell-1}$ from a hypothesis on the distribution and observed frequencies in the ℓ parts. Putting together the ℓ parts two by two in succession, we get χ^2_{k-1} from the same hypothesis and frequencies. Here we assume that ℓ is even, and put $k = \ell/2$. The hypothesis H_0 can be tested by χ^2_{k-1} as well by $\chi^2_{\ell-1}$. In this case, the probability that H_0 is rejected by χ^2_{k-1} but not by $\chi^2_{\ell-1}$ amounts to nearly 3% for the confidence level 0.05. This inconvenience is the defect stated above.

Generally, the resolution of $\chi^2_{\ell-1} : \chi^2_{\ell-1} = \chi^2_{k-1} + \chi^2_{\ell-k}$ is obtained. Here $\chi^2_{\ell-k}$ obeys the χ^2 -distribution of degree of freedom $\ell - k$ and is independent of χ^2_{k-1} . Pair $(\chi^2_{k-1}, \chi^2_{\ell-k})$ contains more information about the distribution, in comparison with the single $\chi^2_{\ell-1}$. Our method is to use the pair $(\chi^2_{k-1}, \chi^2_{\ell-k})$ for the goodness-of-fit test. The critical region R_α corresponding to the confidence level α has been determined on the $(\chi^2_{k-1}, \chi^2_{\ell-k})$ -plane. Numerical values necessary for the test are given on Table 1. Comparing our method and the traditional one, it has turned out that ours is more effective for the goodness-of-fit test.

Our method is not necessarily handy, in comparison with the traditional one. It, however, seems to be satisfactory at the following two points: (1) we have removed the greater

part of the defect stated above, and, (2) viewed from the standpoint using only one variable, the regions for the rejection and non-rejection are separated at one point $\chi_{\ell-1}^2(\alpha)$ in the traditional method, while they overlap in the vicinity of $\chi_{\ell-1}^2(\alpha)$ in our method.

Though there are many ways in getting the coarse division, it has been shown that the nearer the value of k is to the median $(\ell + 1)/2$, the greater the information contained in the pair $(\chi_{k-1}^2, \chi_{\ell-k}^2)$ is.

Appendix A

The three terms in Eq. (1) are generally expressed as follows:

$$z = \chi_{\ell-1}^2 = \sum_{s=1}^k \sum_{i=1}^{\nu(s)} (n_{si} - np_{si})^2 / np_{si}, \quad (\text{A.1})$$

$$x = \chi_{k-1}^2 = \sum_{s=1}^k (n_s - np_s)^2 / np_s, \quad (\text{A.2})$$

$$y = \chi_{\ell-k}^2 = \frac{1}{n} \sum_{s=1}^k \frac{1}{p_s} \sum_{i < j} (\sqrt{p_{si}/p_{sj}} n_{sj} - \sqrt{p_{sj}/p_{si}} n_{si})^2. \quad (\text{A.3})$$

Here the partition number ℓ is arbitrary, and ℓ subintervals are collected into k groups, the s -th of which consists of $\nu(s)$ subintervals. Accordingly we have $\sum_{s=1}^k \nu(s) = \ell$. p_s and n_s

in (A.2) and (A.3) are given by $p_s = \sum_{i=1}^{\nu(s)} p_{si}$ and $n_s = \sum_{i=1}^{\nu(s)} n_{si}$. Expressions (A.1) and (A.2) are the usual ones. The proof for (A.3) can be achieved, if it is shown that the difference $\Delta = \chi_{\ell-1}^2 - \chi_{k-1}^2 - \chi_{\ell-k}^2$ becomes zero when (A.1), (A.2) and (A.3) are inserted into the right hand side. To prove this, we put $\Delta = \sum_{s=1}^k T_s / np_s$. Then T_s becomes

$$T_s = np_s \sum_{i=1}^{\nu(s)} (n_{si} - np_{si})^2 / np_{si} - (n_s - np_s)^2 - \sum_{i < j} (\sqrt{p_{si}/p_{sj}} n_{sj} - \sqrt{p_{sj}/p_{si}} n_{si})^2.$$

and it can readily be shown that T_s vanishes identically.

Appendix B

Some considerations on the information contained in the pair (x, y) will be made in this Appendix. From relation (5) in the text, it is seen that the information contained in the pair (x, y) is equal to that contained in the pair (z, t) . And the latter is equal to the sum of the informations contained in z and t , since z and t are mutually independent. As a result, the information contained in the single z is less than that contained in (x, y) .

Let $I(k)$ be the information contained in variable t . For simplicity, we assume that the observed value of t can be made infinitely sharp. Then we can put

$$I(k) = \lim_{\Delta t \rightarrow 0} \left(- \sum P(\ell, k, t_i) \Delta t \log(P(\ell, k, t_i) \Delta t) \right),$$

which is the analogy of $-\sum p_i \log p_i$ in the discrete case. Here we have divided the interval $0 \leq t \leq 1$ into equal parts of length Δt , and t_i is determined by the condition

$$P(\ell, k, t_i) \Delta t = \int_{\Delta t} P(\ell, k, t) dt. \quad I(k) \text{ is divided into two parts as}$$

$$I(k) = \lim_{\Delta t \rightarrow 0} \left(- \sum P(\ell, k, t_i) \Delta t \log P(\ell, k, t_i) - \sum P(\ell, k, t_i) \Delta t \log \Delta t \right)$$

$$= \bar{I}(k) + I_0,$$

where

$$\bar{I}(k) = - \int_0^1 P(\ell, k, t) dt \log P(\ell, k, t)$$

and

$$I_0 = \lim \sum -P(\ell, k, t_i) \Delta t \log \Delta t = \lim (-\log \Delta t)$$

$$= \lim \left(- \sum \Delta t \log \Delta t \right) = +\infty.$$

Here I_0 is the information contained in the uniform distribution, and the difference $\bar{I}(k) = I(k) - I_0$ is the information measured from I_0 which is independent of k .

From (6) in the text, we have the symmetry relation $\bar{I}(k) = \bar{I}(k')$ for k and k' satisfying the condition $k + k' = \ell + 1$. From numerical calculations, it is seen that the nearer the value of k is to the median $(\ell + 1)/2$, the greater $\bar{I}(k)$ is. For example, the following numerical values for $\bar{I}(k)$ are obtained for $\ell = 12$:

k	=	2 or 11	3 or 10	4 or 9	5 or 8	6 or 7
$\bar{I}(k)$	=	-1.235	-.726	-.435	-.299	-.241

REFERENCES

- Cochran, W.G.(1934). The distribution of quadratic forms in a normal system, with application to the analysis of variance. *Proc. Camb. Phil. Soc.*, Vol.30, 178-191.
Miyatake, O. et al.(1985). A close examination of χ^2 -test. *45th I.S.I.*, 29-30.

(Received December 1987; Revised December 1988.)