ON TESTING FOR NORMALITY

K. SARKADI

Introduction

This paper deals with the following two problems:

1. How to apply the test of normality or the homogenity test of BARTLETT or COCHRAN for the error term if we have one observation per cell in a two-way classification table?

2. How to extend the application of any test of goodness of fit for testing the normality on the basis of a simple sample if the expectation and the variance (or at least one of them) are unknown?

The first problem is of practical interest and was suggested to the author by P. WELLISCH[1]. It is known that most of the variance analysis methods start from the supposition of a normally distributed error term. However, as far as I know, in the textbooks on variance analysis no method of proving this supposition in the case of single observation per cell is treated.

Here the difficulty arises from the fact that forming the differences between the observed values and their predictions these will not be independent. It is known however, that dependent normally distributed variables can easily be linearly transformed into independent ones. Our method of solving the above problem is based on this fact.

For that reason we have to choose such linear transforms of the original values which are mutually uncorrelated with expectation 0 and common variance. In order that the distribution of the transformed variables should be near to the distribution of the error terms (even in the case of alternative hypothesis) it is necessary that each of the transformed values should be highly correlated with one of the original values. This problem is treated in §§ 2—3.

The second problem is of interest as well. At present the χ^2-test is the only one which is adapted to the case of unknown parameters. The transformations given in Sections 1 and 4 allow, however, to apply any test of goodness of fit for testing normality. If only the expectation is unknown the solution is based on the same principle as in the first problem. The general case requires nonlinear transformation.

Of course the methods may be applied if we have several samples the theoretical parameters of which are different and unknown. For this case DUNIN-BARKOVSKY and SMIRNOV [1] have given a transformation reducing the problem to simple goodness of fit test. But while the transformation in

[1] Secretariat of the Council of Plant-Variety Testing, Budapest.

[1] results for the goodness of fit test only as many data as is the number of samples, our transformation decreases the number of data only by the number of unknown constants.

Similar transformation is given for the case of the Gamma parent distribution at the end of § 4. This transformation eliminates the scale parameter from the distribution of the data.

§ 1. Simple sample with unknown expectation

Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be independent random variables with common variance σ^2 and common expectation μ. Let

$$\bar{\mathbf{x}}' = \frac{n\bar{\mathbf{x}} + \mathbf{x}_n\sqrt{n}}{n+\sqrt{n}} = \frac{\sum\limits_{i=1}^{n} \mathbf{x}_i + \mathbf{x}_n\sqrt{n}}{n+\sqrt{n}}.$$

It is easy to prove that the following differences:

$$\mathbf{y}_1 = \mathbf{x}_1 - \bar{\mathbf{x}}',\ \mathbf{y}_2 = \mathbf{x}_2 - \bar{\mathbf{x}}', \ldots, \mathbf{y}_{n-1} = \mathbf{x}_{n-1} - \bar{\mathbf{x}}' \tag{1}$$

have the common expectation 0 and common variance σ^2, further that they are uncorrelated. The first statement is trivial, the two latters can be e.g. easily seen from the fact that $\bar{\mathbf{x}}_{n-1} = (\mathbf{x}_1 + \mathbf{x}_2 + \ldots + \mathbf{x}_{n-1})/(n-1)$ and $(\bar{\mathbf{x}}_{n-1} - \mathbf{x}_n)/\sqrt{n}$ have the same variance $\sigma^2/(n-1)$, both of them are uncorrelated to $\mathbf{x}_i - \bar{\mathbf{x}}_{n-1}$ $(i = 1, 2, \ldots, n-1)$; and that $\mathbf{x}_i = \bar{\mathbf{x}}_{n-1} + (\mathbf{x}_i - \bar{\mathbf{x}}_{n-1})$, $\mathbf{y}_i = (\bar{\mathbf{x}}_{n-1} - \mathbf{x}_n)/\sqrt{n} + (\mathbf{x}_i - \bar{\mathbf{x}}_{n-1})$ and thus the random vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n-1}\}$ and $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{n-1}\}$ have the same dispersion matrix.

It follows that if the distribution of $\mathbf{x}_i$ is normal then the variates $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{n-1}$ are mutually independent equally distributed normal variates.

The correlation coefficient of $\mathbf{x}_i$ and $\mathbf{y}_i$ is

$$\mathbf{R}\{\mathbf{x}_i, \mathbf{y}_i\} = 1 - \frac{1}{n+\sqrt{n}}.$$

The transformation (1) is optimal among all linear transformations into $n-1$ uncorrelated variates with 0 expectation in the sense that $\min\limits_i \mathbf{R}\{\mathbf{x}_i, \mathbf{y}_i\}$ is maximized by (1).

Proof. The statement is equivalent to the following: If in an n by n orthogonal matrix $C = \{c_{ij}\}$ $c_{n1} = c_{n2} = \ldots = c_{nn} = 1/\sqrt{n}$ then $\min\limits_{i \leqq n-1} c_{ii} \leqq 1 - 1/(n+\sqrt{n})$. (The equivalency is easy to be seen. The transformation matrix C gives a transformed vector whose first $n-1$ elements provide a linear transformation of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ into $n-1$ mutually uncorrelated variables $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{n-1}$ uncorrelated to $\bar{\mathbf{x}}$ as well, i.e. having 0 expectation independently of μ. Apart from constant factors, there is a 1 : 1 correspondance between all possible such transformations and the possible values of C.)

Evidently for some $i \leqq n-1$ $|c_{in}| \geqq 1/\sqrt{n}$. As

$$\mathbf{R}\{\mathbf{x}_i - \bar{\mathbf{x}}, \mathbf{y}_i\} = c_{ii}\sqrt{n/(n-1)},$$

$$\mathbf{R}\{\mathbf{x}_n - \overline{\mathbf{x}}, \mathbf{y}_i\} = c_{in}\sqrt{n/(n-1)}, \qquad \mathbf{R}\{\mathbf{x}_i - \overline{\mathbf{x}}, \mathbf{x}_n - \overline{\mathbf{x}}\} = -1/(n-1)$$

and from the geometrical interpretation of the linear functions of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ it is to be seen that[2]

$$\arccos \mathbf{R}\{\mathbf{x}_i - \overline{\mathbf{x}}, \mathbf{y}_i\} \geqq |\arccos \mathbf{R}\{\mathbf{x}_n - \overline{\mathbf{x}}, \mathbf{y}_i\} - \arccos \mathbf{R}\{\mathbf{x}_n - \overline{\mathbf{x}}, \mathbf{x}_i - \overline{\mathbf{x}}\}|$$

the following inequality is valid:

$$c_{ii} = \mathbf{R}\{\mathbf{x}_i, \mathbf{y}_i\} \leqq 1 - \frac{1}{n + \sqrt{n}}$$

and thus the above statement is proved.

The transformation given by (1) gives the possibility of applying any test of goodness of fit in case of unknown mean and known variance and we can take into account that the expectation is unknown not only in applying a χ^2-test but in that of KOLMOGOROV, SMIRNOV and RÉNYI too. In addition, if we apply the χ^2-test the diminishing of the degrees of freedom due to the estimation of the expectation can be avoided.

The element $\mathbf{x}_n$ plays a special role among the sample elements in the transformation. It is chosen for the sake of simplicity; of course any of the sample elements can be randomly chosen.

§ 2. Two-way classification, one observation per cell

Let the variates $\mathbf{x}_{ij}$ $(i = 1, 2, \ldots, s;\ j = 1, 2, \ldots, v)$ be independent with common variance σ^2 and with expectations

$$\mathbf{E}\{\mathbf{x}_{ij}\} = \mu_{..} + \mu_{i.} + \mu_{.j}$$

where the constants $\mu_{..}, \mu_{i.}, \mu_{.j}$ are unknown.

Now we can define the following $(s-1)(v-1)$ uncorrelated variates:

$$\mathbf{y}_{ij} = \mathbf{x}_{ij} - \mathbf{x}'_{i.} - \mathbf{x}'_{.j} + \mathbf{x}'_{..} \tag{2}$$

$$(i = 1, 2, \ldots, s-1;\ j = 1, 2, \ldots, v-1)$$

where

$$\mathbf{x}'_{i.} = \frac{\sum_{j=1}^{v} \mathbf{x}_{ij} + \mathbf{x}_{iv}\sqrt{v}}{v + \sqrt{v}}$$

$$\mathbf{x}'_{.j} = \frac{\sum_{i=1}^{s} \mathbf{x}_{ij} + \mathbf{x}_{sj}\sqrt{s}}{s + \sqrt{s}}$$

$$\mathbf{x}'_{..} = \frac{\sum_{i=1}^{s}\sum_{j=1}^{v} \mathbf{x}_{ij} + \sqrt{v}\sum_{i=1}^{s} \mathbf{x}_{iv} + \sqrt{s}\sum_{j=1}^{v} \mathbf{x}_{sj} + \mathbf{x}_{sv}\sqrt{sv}}{(s+\sqrt{s})(v+\sqrt{v})}.$$

[2] In geometrical interpretation this is the triangle inequality in the spherical triangle determined by the vectors corresponding to the variates $\mathbf{x}_i - \overline{\mathbf{x}}$, $\mathbf{x}_n - \overline{\mathbf{x}}$, $\mathbf{y}_i$.

As shown below $\operatorname{cov}(\mathbf{y}_{ij}, \mathbf{y}_{kl}) = 0$ for $(i, j) \neq (k, l)$ and the variates $\mathbf{y}_{ij}$ have common expectation and variance $\mathbf{E}\{\mathbf{y}_{ij}\} = 0$ and $\mathbf{D}^2\{\mathbf{y}_{ij}\} = \sigma^2$.

Proof. The quantities $\mathbf{x}_{1j} - \mu_{1.}, \mathbf{x}_{2j} - \mu_{2.}, \ldots, \mathbf{x}_{sj} - \mu_{s.}$ have the common expectations $\mu_{.j} - \mu_{..}$ and variance σ^2. Thus applying the transformation (1), we obtain the transforms $\mathbf{y}'_{ij}$ having the expectation 0 and variance σ^2. The quantities $\mathbf{y}'_{ij}$ and $\mathbf{y}'_{kl}$ will be independent for $j \neq l$ because of the independency of $\mathbf{x}_{ij}$'s and uncorrelated for $j = l,\ i \neq k$ because of the property of the transformation. Now we apply the transformation (1) to the series $\mathbf{y}'_{i1}, \mathbf{y}'_{i2}, \ldots, \mathbf{y}'_{iv}$ which leads — after some calculation — to the quantities $\mathbf{y}_{ij}$ defined by (2). The quantities $\mathbf{y}_{ij}$ thus have the expectation 0 and variance σ^2. It follows in case of a normal parent distribution that the quantities $\mathbf{y}_{ij}$ will be mutually independent. As their first two moments do not depend on the assumption of normality they will be mutually uncorrelated in the non-normal case.

The correlation coefficient between $\mathbf{x}_{ij}$ and $\mathbf{y}_{ij}$ is

$$\mathbf{R}\{\mathbf{x}_{ij}, \mathbf{y}_{ij}\} = \left(1 - \frac{1}{s + \sqrt{s}}\right)\left(1 - \frac{1}{v + \sqrt{v}}\right). \tag{3}$$

$\mathbf{x}_{ij}$ will be called basic element if it has a highly correlated correspondent among the transformed values. The above transformation has the property that the basic elements form an $s - 1$ by $v - 1$ submatrix in the original matrix of the $\mathbf{x}_{ij}$ s. This property is advantageous both for the purpose of simplicity of the formulae and that of applying COCHRAN's or BARTLETT's test. But the author does not know whether the transformation is optimal in the sense of § 1. The intercorrelations betwen the basic elements can be decreased in absolute value by other choices of the basic elements. This is, e. g. the case for $s = v = 3$ if we choose $\mathbf{x}_{12}, \mathbf{x}_{13}, \mathbf{x}_{21}, \mathbf{x}_{31}$ for basic elements. One could expect that such choices may provide an increased $\min\limits_{i,j} \mathbf{R}\{\mathbf{x}_{ij}, \mathbf{y}_{ij}\}$ In the mentioned special case, however, this does not hold. The optimal transformation with the mentioned choice of basic elements is

$$\begin{aligned}
\mathbf{y}_{12} &= -\frac{1}{4}\left[(2 + \sqrt{6})\,\mathbf{u}_{22} + (4 - \sqrt{6})\,\mathbf{u}_{23} + (4 + \sqrt{6})\,\mathbf{u}_{32} + (2 - \sqrt{6})\,\mathbf{u}_{33}\right] \\
\mathbf{y}_{13} &= -\frac{1}{4}\left[(2 - \sqrt{6})\,\mathbf{u}_{22} + (4 + \sqrt{6})\,\mathbf{u}_{23} + (4 - \sqrt{6})\,\mathbf{u}_{32} + (2 + \sqrt{6})\,\mathbf{u}_{33}\right] \\
\mathbf{y}_{21} &= -\frac{1}{4}\left[(4 + \sqrt{6})\,\mathbf{u}_{22} + (2 + \sqrt{6})\,\mathbf{u}_{23} + (2 - \sqrt{6})\,\mathbf{u}_{32} + (4 - \sqrt{6})\,\mathbf{u}_{33}\right] \\
\mathbf{y}_{31} &= -\frac{1}{4}\left[(4 - \sqrt{6})\,\mathbf{u}_{22} + (2 - \sqrt{6})\,\mathbf{u}_{23} + (2 + \sqrt{6})\,\mathbf{u}_{32} + (4 + \sqrt{6})\,\mathbf{u}_{33}\right]
\end{aligned} \tag{4}$$

where

$$\mathbf{u}_{ij} = \mathbf{x}_{ij} - \frac{1}{3}\sum_{k=1}^{3}\mathbf{x}_{kj} - \frac{1}{3}\sum_{l=1}^{3}\mathbf{x}_{il} + \frac{1}{9}\sum_{k=1}^{3}\sum_{l=1}^{3}\mathbf{x}_{kl}$$

In this case

$$\mathbf{R}\{\mathbf{x}_{ij}, \mathbf{y}_{ij}\} = \frac{1+\sqrt{6}}{6} \tag{5}$$

for $(i, j) = (1, 2), (1, 3), (2, 1), (3, 1)$. Formula (3) gives for this case $\mathbf{R}\{\mathbf{x}_{ij}, \mathbf{y}_{ij}\} = (2+\sqrt{3})/6$ which is larger than (5).

The transformation (2) gives the possibility of testing the normality of the error term in our case. In addition it allows the testing of homogenity of variance between rows or between columns or between different tables of data with COCHRAN's or BARTLETT's criterion.

Evidently any row and column may play the role of the v-th row and s-th column in the written formulae. The choosing of them however, must not depend on the actual values.

§ 3. The general case of variance analysis

The case of n-way classification can be treated in the same way. In principle, the method can be extended for any case of variance analysis.

§ 4. Transforms independent from variance

In this Section there are given transforms which are independent not only from the unknown expectation but from the unknown variance too. In case of normality tests in general not only the expectation but the variance is unknown too. Thus our transformation gives the possibility of applying any test of goodness of fit for the general case of testing normality. Let us suppose we have performed the transformation given in Sections 2, 3 or 4 and we have ν variates $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_\nu$ with 0 expectation and variance σ^2. Suppose they are independent and normally distributed. We define the following transforms:

$$\mathbf{z}_i = \operatorname{sg} \mathbf{y}_i \, I_{\mathbf{t}_i}\left(\frac{1}{2}, \frac{v-i}{2}\right) \qquad (i = 1, 2, \ldots, \nu - 1) \tag{6}$$

where

$$\mathbf{t}_i = \frac{\mathbf{y}_i^2}{\mathbf{y}_i^2 + \mathbf{y}_{i+1}^2 + \ldots + \mathbf{y}_\nu^2}$$

and

$$I_t(p, q) = \frac{\int\limits_0^t x^{p-1}(1-x)^{q-1}\,dx}{\int\limits_0^1 x^{p-1}(1-x)^{q-1}\,dx}$$

is the incomplete Beta-function tabulated in [3].

Since $I_t\left(\frac{1}{2}, \frac{\nu-i}{2}\right)$ is the distribution function of $\mathbf{t}_i$, the variates $\mathbf{z}_i$ defined by (6) are uniformly distributed in the interval $(-1, 1)$. According to a theorem of E. LUKÁCS [2] $\mathbf{y}_i^2 + \mathbf{y}_{i+1}^2 + \ldots + \mathbf{y}_\nu^2$ and $\mathbf{y}_i^2/(\mathbf{y}_{i+1}^2 + \ldots + \mathbf{y}_\nu^2)$

are independent which implies that the variates $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{\nu-1}$ are mutually independent.

If the alternative hypothesis holds the variates $\mathbf{z}_i$ are in general not identically distributed and not independent. If ν tends to infinity and i remains constant the distribution of $\mathbf{z}_i$ tends to the distribution of $2\,\Phi(\mathbf{y}_i) - 1$ where $\Phi(y) = (2\,\pi)^{-1/2} \int\limits_{-\infty}^{y} e^{-x^2/2}\,dx$.

We may apply any test of goodness of fit for the $\mathbf{z}_i$'s.

The distribution of $\mathbf{z}_i$'s with small $\nu - i$ for the alternative hypothesis requires further investigations. Probably the goodness of fit tests can be ameliorated for large series if we omit a few values from the end of the series of $\mathbf{z}_i$'s.

The results of this Section can be extended easily for the case of a Gamma parent distribution. If $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_\nu$ are independently distributed and have a common Gamma distribution with density function

$$\frac{\alpha^\lambda}{\Gamma(\lambda)}\, x^{\lambda-1}\, e^{-\alpha x}$$

where λ is known but α is unknown the transformation

$$\mathbf{z}_i = I_{\mathbf{t}_i}\big(\lambda, (\nu - i)\,\lambda\big)$$

$$(i = 1, 2, \ldots, \nu - 1)$$

can be applied, where

$$\mathbf{t}_i = \frac{\mathbf{w}_i}{\mathbf{w}_i + \mathbf{w}_{i+1} + \ldots + \mathbf{w}_\nu}.$$

The variates $\mathbf{z}_i$ defined by the above formula are uniformly distributed in the interval (0, 1). Their mutual independency follows again from the theorem of LUKÁCS [2].

The above transformation gives the possibility of applying any test of goodness of fit in the case of a Gamma distribution with unknown α and known λ.

(Received January 27, 1960.)

REFERENCES

[1] ДУНИН-БАРКОВСКИЙ, И. В — СМИРНОВ, Н. В.: *Теория вероятностей и математическая статистика в технике.* Гостехиздат, Москва, 1955, pp. 354—360.

[2] LUKÁCS, E.: "A characterization of the Gamma distribution." *Annals of Mathematical Statistics* **26** (1955) 319—324.

[3] PEARSON, K.: *Tables of the Incomplete Beta-Function.* Cambridge University Press, Cambridge, 1948.

О ПРОВЕРКЕ ГИПОТЕЗЫ НОРМАЛЬНОСТИ

K. SARKADI

Резюме

Работа занимается двумя следующими проблемами.

1. Как можно исследовать нормальность распределения остаточного члена или произвести пробу BARTLETT-а или COCHRAN-а относительно однородности, если в таблице двусторонней классификации имеется одно наблюдение в каждой ячейке?

2. Как можно применять любой метод проверки гипотезы нормальности, если не известно математическое ожидание, дисперсия или оба эти значения?

Даются преобразования, делающие возможным проведение вышеуказанных исследований.

Формула (1) в случае простой пробы, а формула (2) в случае таблицы двусторонней классификации с одним наблюдением в системе преобразуют величины в величины без корреляции, с нулевым математическим ожиданием и дисперсией, равной исходней. Применяя преобразование (6) к полученным таким образом величинам, получим независимые значения $\mathbf{z}_i$, распределение которых будет равномерным на отрезке $[-1, +1]$, если исходное распределение было нормальным.