THE LIKELIHOOD TEST OF INDEPENDENCE IN CONTINGENCY TABLES

By S. S. Wilks

J. Neyman and E. S. Pearson[1] have applied the principle of the ratio of likelihoods to the problem of determining criteria for testing various hypotheses about the group frequencies in problems dealing with grouped data. In particular, they have discussed the fundamental χ^2 problem, the test of goodness of fit, the hypothesis that two samples of grouped data are from the same population, and the hypothesis of independence in contingency tables. In their treatment of these problems, these authors have started from the limiting form of the probability of an observed set of frequencies and have shown that approximately each of the appropriate λ's is a function of the minimum value of a corresponding χ_1^2. The distribution of this minimum value is found, from which the significance test is made.

In certain cases the exact values of the λ's are relatively simple functions of the observations which can be as conveniently calculated as the corresponding χ^2's. The purpose of this note is to consider the exact expressions for the λ's and find their asymptotic distributions in large samples for the following hypotheses: (1) that a sample of grouped data is from a population with specified group frequencies (i.e., the fundamental χ^2 problem) ,(2) that several samples of grouped data are from the same population, and (3) that there is independence in a contingency table.

1. **The fundamental χ^2 problem.** Let $p_1, p_2, \cdots p_k$ be the probabilities of the mutually exclusive events $E_1, E_2, \cdots E_k$ respectively. In a sample of N events the probability that $E_1, E_2, \cdots E_k$ will occur $n_1, n_2, \cdots n_k$ times respectively, is given by

$$C = \frac{N!}{n_1!\, n_2! \cdots n_k!}\, p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}. \tag{1}$$

If we let Ω be the class of all sets of values of the p's such that their sum is unity, there is only one set of p's that maximize C, namely, $p_j = n_j/N$ $(j = 1, 2, \cdots k)$. The maximum of C is

$$C(\Omega \max) = \frac{N!}{n_1!\, n_2! \cdots n_k!} \cdot \frac{n_1^{n_1} n_2^{n_2} \cdots n_k^{n_k}}{N^N}. \tag{2}$$

[1] Biometrika, vol. 20A (1928), pp. 263–294.

The likelihood of the hypothesis that the sample is from a population specified by p's having the values $p_1, p_2, \cdots p_k$ is defined as

$$\lambda_s = \frac{C}{C(\Omega \max)} = \left(\frac{Np_1}{n_1}\right)^{n_1}\left(\frac{Np_2}{n_2}\right)^{n_2} \cdots \left(\frac{Np_k}{n_k}\right)^{n_k}. \tag{3}$$

λ_s is a quantity which clearly lies between 0 and 1. It will be 1 only when $p_j = n_j/N$ $(j = 1, 2, \cdots k)$, (that is, when the hypothesis is rigorously supported by the sample) and tends to 0 as the sample values n_j/N diverge more and more from the hypothetical values p_j. The problem of making an exact test of significance of an observed value of λ_s would involve the computation of all terms of form (1) the n's of which make λ_s less than the observed value of λ_s. This, of course, is impracticable except perhaps for the binomial case with small values of N. However, if the n's are large we can find an approximate solution. If we let $x_j = \dfrac{n_j - Np_j}{\sqrt{N}}$ then except for terms of order $1/\sqrt{N}$ and higher, the x's are distributed according to the law

$$\frac{1}{\sqrt{(2\pi)^{k-1} p_1 p_2 \cdots p_k}}\, e^{-\frac{1}{2}\sum_j \frac{x_j^2}{p_j}}, \tag{4}$$

where $\Sigma_j x_j = 0$. Neglecting terms of order $1/\sqrt{N}$ and higher we easily find (using natural logarithms) $-2 \log \lambda_s = \sum_j \dfrac{x_j^2}{2p_j}$. Therefore, if $\theta = -2 \log \lambda_s$, θ is approximately distributed according to the function

$$\frac{(\frac{1}{2})^{\frac{k-1}{2}}\, \theta^{\frac{k-3}{2}}}{\Gamma\left(\dfrac{k-1}{2}\right)} e^{-\frac{\theta}{2}} \tag{5}$$

which is the χ^2 distribution with $k - 1$ degrees of freedom.

Since we have neglected terms of order $1/\sqrt{N}$ in obtaining (4) there is no theoretical reason why χ^2 should be used in preference to $-2 \log \lambda_s$ as the criterion for testing the hypothesis that the sample is from a population specified by $p_1, p_2, \cdots p_k$. Any practical advantage which $-2 \log \lambda_s$ may have will therefore justify its use.

2. The hypothesis that several samples of grouped data are from a common population. Let $p_{i1}, p_{i2}, \cdots p_{is}$ be the probabilities with which the mutually exclusive events $E_{i1}, E_{i2}, \cdots E_{is}$ occur, where $\Sigma_j p_{ij} = 1$ $(i = 1, 2, \cdots r)$. Then in a sample of N_i events the chance that $E_{i1}, E_{i2}, \cdots E_{is}$ will occur $n_{i1}, n_{i2}, \cdots n_{is}$ times respectively is given by an expression similar to (1). The chance of the joint occurrence of the r samples is

$$\frac{N_1!\, N_2! \cdots N_r!}{n_{11}!\, n_{12}! \cdots n_{rs}!}\, p_{11}^{n_{11}} p_{12}^{n_{12}} \cdots p_{rs}^{n_{rs}}. \tag{6}$$

We are interested in testing the hypothesis that the r samples are from the same population, that is, that the r sets of p's $p_{i1}, p_{i2}, \cdots p_{is}$ $(i = 1, 2, \cdots r)$ are the same. The likelihood criterion λ_c appropriate to this hypothesis is the ratio of the maximum (ω(max)) of (6) subject to the condition that the sets of p's are the same (that is, $p_{ij} = p_j$ say, $i = 1, 2, \cdots r$; $j = 1, 2, \cdots s$) to the maximum (Ω (max)) of (6) without this restriction.

For convenience let the observations be arranged in table form so that n_{ij} is the frequency in the i-th row and j-th column. Let $n_{i\cdot}$ and $n_{\cdot j}$ be the totals of the i-th row and j-th column respectively, and N the total of all observations. Thus $n_{i\cdot}$ is the same as N_i. The expression for λ_c will be

$$\lambda_c = \frac{n_{\cdot 1}^{n_{\cdot 1}} n_{\cdot 2}^{n_{\cdot 2}} \cdots n_{\cdot s}^{n_{\cdot s}} n_{1\cdot}^{n_{1\cdot}} n_{2\cdot}^{n_{2\cdot}} \cdots n_{r\cdot}^{n_{r\cdot}}}{N^N n_{11}^{n_{11}} n_{12}^{n_{12}} \cdots n_{rs}^{n_{rs}}}. \tag{7}$$

It can be shown analytically that λ_c lies between 0 and 1. It can be 1 only when $\frac{n_{1j}}{N_1} = \frac{n_{2j}}{N_2} = \cdots = \frac{n_{rj}}{N_r}$, $j = 1, 2, \cdots s$, that is, when the hypothesis of a common population is perfectly substantiated by the samples. Because of the fact that the n_{ij} are integers, it is clear that λ_c can be 1 only in exceptional cases, but it can take on values arbitrarily near 1 for sufficiently large values of the n_{ij}.

If the N_i are large, the quantities $x_{ij} = \frac{n_{ij} - N_i p_j}{\sqrt{N_i}}$ are approximately distributed according to the function

$$F = \left(\frac{1}{(2\pi)^{s-1} p_1 p_2 \cdots p_s}\right)^{\frac{r}{2}} e^{-\frac{1}{2}\sum_{i,j} \frac{x_{ij}^2}{p_j}}, \tag{8}$$

where $\Sigma_j x_{ij} = 0$, $i = 1, 2, \cdots r$. By neglecting terms of order $1/\sqrt{N}$ and higher, we find that

$$-2 \log \lambda_c = \sum_{i,j} \frac{(N x_{ij} - \sqrt{N_i}(\sum_i \sqrt{N_i} x_{ij}))^2}{N^2 p_j}. \tag{9}$$

Denoting the quantity on the right side of (9) by χ_0^2 it follows by straightforward analysis that the characteristic function $\varphi(t)$ of χ_0^2 defined by the $r(s-1)$-tuple integral $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{it\chi_0^2} F \, dx_{11} \cdots dx_{rs}$ has the value

$$(\tfrac{1}{2})^{\frac{(r-1)(s-1)}{2}} (\tfrac{1}{2} - it)^{-\frac{(r-1)(s-1)}{2}} \tag{10}$$

But it is well known that (10) is the characteristic function of any quantity distributed according to (5) with $(k-1)$ replaced by $(r-1)(s-1)$. This, of course, is the χ^2 distribution with $(r-1)(s-1)$ degrees of freedom.

It will be noticed that the exact value of λ_c is a function of the observations n_{ij} which is independent of the p's, while the approximate value of $-2 \log \lambda_c$

as given by (9) involves the p's. Before (9) could be used practically, one would have to replace the p's by sample estimates, thus making further approximations necessary in order to get the distribution. If the usual estimates $p_j = n_{.j}/N$ are used for the p's in χ_0^2 we find that χ_0^2 reduces to

$$\sum_{i,j} \frac{\left(n_{ij} - \frac{n_{i.}\, n_{.j}}{N}\right)^2}{\frac{n_{i.}\, n_{.j}}{N}} \tag{11}$$

which is the familiar χ^2 function for testing independence in contingency tables. However, (11) differs from χ_0^2 by terms of the same order (i.e., $1/\sqrt{N_i}$) as those by which χ_0^2 differs from $-2 \log \lambda_c$. Since we have neglected terms of the same order in obtaining (8), there is no theoretical reason why (11) should be used rather than $-2 \log \lambda_c$ for testing the hypothesis that the m samples are from a common population.

3. **The hypothesis of independence in contingency tables.** We shall consider a sample of N observations which can be arranged in a two-way contingency table having r rows and s columns. Let p_{ij} be the probability that an observation will fall in the i-th row and j-th column. The probability that the sample of N items will be distributed so that n_{ij} will be the number falling in the i-th row and j-th column ($i = 1, 2, \cdots r$; $j = 1, 2, \cdots s$) is given by

$$\frac{N!}{n_{11}!\, n_{12}! \cdots n_{rs}!}\, p_{11}^{n_{11}} p_{12}^{n_{12}} \cdots p_{rs}^{n_{rs}}. \tag{12}$$

Here we are interested in testing the hypothesis that the classification by rows is independent of the classification by columns, that is, that p_{ij} is of the form $p_i q_j$ where

$$\Sigma_i p_i = 1, \qquad \Sigma_j q_j = 1. \tag{13}$$

For this hypothesis the appropriate likelihood criterion, say λ_c', is the ratio of the maximum (ω(max)) of (12) when $p_{ij} = p_i q_j$ restricted by the conditions (13) to the maximum (Ω (max)) of (12) subject only to the condition that $\sum_{i,j} p_{ij} = 1$. λ_c' turns out to be identical with λ_c in (7). When the hypothesis of independence is true, the approximate distribution of the quantity $-2 \log \lambda_c'$ is the same as that of $-2 \log \lambda_c$ when the hypothesis of a common population is true. To show that the distributions are the same we note that by placing

$$x_{ij} = \frac{n_{ij} - N p_i q_j}{\sqrt{N}}, \tag{14}$$

we find from (12) that the x_{ij} are approximately distributed according to the function

$$(15)\qquad \frac{1}{(2\pi)^{\frac{rs-1}{2}}(p_1p_2\cdots p_r)^{\frac{s}{2}}(q_1q_2\cdots q_s)^{\frac{r}{2}}}e^{-\frac{1}{2}\sum_{i,j}\frac{x_{ij}^2}{p_iq_j}}$$

where $\sum_{i,j} x_{ij} = 0$. To the same degree of approximation we find

$$(16)\qquad -2\log\lambda_c' = \sum_{i,j}\frac{(x_{ij} - p_i\sum_i x_{ij} - q_j\sum_j x_{ij})^2}{p_iq_j} = \chi_0'^2 .$$

Now the characteristic function of $\chi_0'^2$ can be shown without much difficulty to be identical with that of χ_0^2 as given by (10). The identity of the characteristic functions of $\chi_0'^2$ and χ_0^2 implies the identity of the asymptotic distributions of $-2\log\lambda_c'$ and $-2\log\lambda_c$. The problem of testing the hypothesis of a common population in several samples of grouped data is mathematically equivalent to that of testing the hypothesis of independence in contingency tables.

If the usual estimates $p_i = \frac{n_{i\cdot}}{N}$, $q_j = \frac{n_{\cdot j}}{N}$ are used in (16) we find that χ_0^2 becomes the expression given by (11). But (11) differs from $\chi_0'^2$ by terms of order $1/\sqrt{N}$ and higher. Therefore, $-2\log\lambda_c'$ and (11) can differ from each other only by terms of order $1/\sqrt{N}$ which is the order of approximation involved in getting (15) from (12). Thus, $-2\log\lambda_c'$ has as much validity as the usual criterion (11) for testing for independence in contingency tables.

The λ_c' method can easily be extended to the case of contingency tables of higher order. For example, in a three-way table of r rows, s columns and t layers in which n_{ijk} is the number of items observed in the i-th row, j-th column and k-th layer, the λ_c' criterion for testing the hypothesis of independence, that is, that the probabilities p_{ijk} are of the form $p_{1i}p_{2j}p_{3k}$ is such that

$$(17)\qquad \begin{aligned}-2\log\lambda_c' = {} & 2\sum_{i,j,k}(n_{ijk}\log n_{ijk}) + 4N\log N - 2\sum_i(n_{i\cdot\cdot}\log n_{i\cdot\cdot}) \\ & - 2\textstyle\sum_j(n_{\cdot j\cdot}\log n_{\cdot j\cdot}) - 2\sum_k(n_{\cdot\cdot k}\log n_{\cdot\cdot k})\end{aligned}$$

where $n_{i\cdot\cdot} = \sum_{j\cdot k} n_{ijk}$, and so on. $-2\log\lambda'$ in this case is approximately distributed like χ^2 with $rst - r - s - t + 2$ degrees of freedom.

4. **Illustrative examples.** To illustrate the use of λ_s we shall consider the following example given by R. A. Fisher[2] dealing with de Winton and Bateson's data on results of interbreeding the hybrid (F_1) generation of *Primula* in which two factors are considered.

	Flat Leaves		Crimped Leaves		Total
	Normal Eye	Primrose Queen Eye	Lee's Eye	Primrose Queen Eye	
Observed (n_i)	328	122	77	33	560
Expected (Np_i)	315	105	105	35	560

[2] Statistical Methods for Research Workers, 4th ed. p. 84.

If the two factors are Mendelian, that is, segregate independently, the four classes of offspring resulting from interbreeding the F_1 generation are expected to appear in the ratio 9:3:3:1 (assuming all classes equally viable). We wish to test the hypothesis of a 9:3:3:1 ratio. It is found that

$$-2\log_e \lambda_s = 2\log_e 10\left[\sum_i n_i \log_{10} n_i - \sum_i n_i \log_{10}(Np_i)\right] = 11.50\,.$$

Entering Fisher's χ^2 table for $n = 3$, we find that the chance of exceeding the value 11.50 is less than .01, which is significant if we take $P = .05$ as the critical level of significant deviation. Thus, the observed frequencies cannot be reasonably explained as chance deviations from the 9:3:3:1 ratio.

The usual χ^2 method gives $\chi^2 = 10.87$ and $n = 3$ for the 9:3:3:1 hypothesis. The value of P in this case lies between .01 and .02. It follows from the theoretical discussion that 10.87 has no greater validity than 11.50 in testing this hypothesis.

We shall illustrate the use of λ_c by using another example given by Fisher dealing with Wachter's data for back-crosses in mice.

	Black Self	Black Piebald	Brown Self	Brown Piebald	Total
Coupling:					
F_1 Males	88	82	75	60	305
F_1 Females	38	34	30	21	123
Repulsion:					
F_1 Males	115	93	80	130	418
F_1 Females	96	88	95	79	358
Total	337	297	280	290	1204

The back-crosses were made according as the male or female parents of the F_1 generation were heterozygous in the two factors Black-Brown, Self-Piebald, and according to whether the two dominant genes came both from one parent (Coupling) or one from each parent (Repulsion). We wish to test the hypothesis that the proportions are independent of the matings used. We find

$$-2\log \lambda_c = 2\log_e 10\left[\sum_{i,j} n_{ij}\log_{10} n_{ij} + N\log_{10} N - \textstyle\sum_i n_{i\cdot}\log_{10} n_{i\cdot} - \sum_j n_{\cdot j}\log_{10} n_{\cdot j}\right] = 21.69\,.$$

Entering Fisher's χ^2 table for $n = 9$ we find that the chance of exceeding this value is less than .01. The departure from the hypothesis of independence is significant on basis of the $P = .05$ level. The χ^2 method gives the remarkably close result $\chi^2 = 21.83$, which, with $n = 9$ gives $P < .01$.

5. **Summary.** We have considered the exact expressions for the Neyman-Pearson λ criteria appropriate to the following hypotheses: (1) That a sample

of grouped data is from a population with specified group proportions (the fundamental χ^2 problem), (2) that several samples of grouped data are from a common population, (3) that there is independence in a contingency table. The quantity $-2 \log \lambda$ for each of these cases is approximately distributed like χ^2, the number of degrees of freedom being given in each case. It is shown that the usual χ^2 method of testing these hypotheses has no greater theoretical validity than the λ method. On the practical side, it is to be remarked that $-2 \log \lambda$ can be computed with fewer operations than χ^2. Two examples are given to illustrate the practical application of the λ method.

PRINCETON UNIVERSITY.