ON THE CHI-SQUARE DISTRIBUTION FOR SMALL SAMPLES[1]

By Paul G. Hoel

1. **Introduction.** The use of what is known as the χ^2 distribution function for testing goodness of fit involves two types of error. One arises from the fact that the derivation of this function is based upon rough approximations, while the other arises from using the integral of this continuous function in place of summing the proper terms of a discrete set. Both of these errors become increasingly important as the sample becomes small. The purpose of this paper is to investigate the nature of this first type of error by finding a better approximation than the customary one to what might be called the exact continuous χ^2 distribution function.

The method employed is that of generating or characteristic functions, and consists in expressing successively in expanded form the generating function of the multinomial, the distribution function of the multinomial, the generating function of χ^2, and the distribution function of χ^2. Only the first and second order terms of this final distribution function are evaluated explicitly because of the increasingly heavy algebra involved. By means of these second order terms, the nature of the error involved in the use of the customary first order approximation is investigated.

2. **The Generating Function of the Multinomial.** Consider $k + 1$ cells into which observations can fall, and let p_i be the probability that an observation will fall in cell i. If n observations are made, the probability that cell i will contain α_i of these observations is given by the multinomial

$$P = \frac{n!}{\alpha_1!\,\alpha_2!\cdots\alpha_{k+1}!}\,p_1^{\alpha_1}p_2^{\alpha_2}\cdots p_{k+1}^{\alpha_{k+1}},$$

where $\sum_{i=1}^{k+1}\alpha_i = n$. The generating function of this multinomial can be written as[2]

$$M = [p_1e^{t_1} + \cdots + p_ke^{t_k} + p]^n = \left[1 + \sum_{i=1}^{k} p_i(e^{t_i} - 1)\right]^n,$$

where α_{k+1} is chosen as the dependent variable and p_{k+1} is written as p.

---

[1] Presented to the American Mathematical Society, April 9, 1938.

[2] Cf. Darmois, *Statistique Mathematique*, pp. 237–242, for the methods used in this and the next two paragraphs.

Let $x_i = \dfrac{\alpha_i - np_i}{\sqrt{n}}$. The generating function of the x_i is obtained from that of the α_i by multiplying M by the proper factor to shift the origin to the mean and then replacing t_i by, say, $u_i/\sqrt{n}$ to compensate for the change in scale. Denoting this function by φ,

$$\varphi = e^{-n\sum_{i=1}^{k}\frac{p_i u_i}{\sqrt{n}}}\left[1 + \sum_{i=1}^{k} p_i(e^{u_i/\sqrt{n}} - 1)\right]^n.$$

Consequently,

$$\log \varphi = -\sqrt{n}\sum_{i=1}^{k} p_i u_i + n \log\left[1 + \sum_{i=1}^{k} p_i(e^{u_i/\sqrt{n}} - 1)\right].$$

Since the range of the u_i may be selected sufficiently small for convergence, the logarithm on the right may be expanded in powers of the summation, which in turn may be expanded in powers of the u_i. Terms containing $1/n^{\frac{1}{2}q}$ as a factor will be homogeneous in the u_i of degree $q + 2$. Writing down only the terms of order $1/n$ and lower, this double expansion gives

$$\begin{aligned}
\log \varphi = {} & \frac{1}{2}\left[\sum_{i=1}^{k}(p_i - p_i^2)u_i^2 - 2\sum_{i<j} p_i p_j u_i u_j\right] \\
& + \frac{1}{\sqrt{n}}\left[\frac{1}{6}\sum_{i=1}^{k}(p_i - 3p_i^2 + 2p_i^3)u_i^3 - \frac{1}{2}\sum_{i\neq j}(p_i p_j - 2p_i^2 p_j)u_i^2 u_j\right. \\
& \left. + 2\sum_{i<j<l} p_i p_j p_l u_i u_j u_l\right] + \frac{1}{n}\left[\frac{1}{24}\sum_{i=1}^{k}(p_i - 7p_i^2 + 12p_i^3 - 6p_i^4)u_i^4\right. \\
& - \frac{1}{6}\sum_{i\neq j}(p_i p_j - 6p_i^2 p_j + 6p_i^3 p_j)u_i^3 u_j \qquad (1) \\
& - \frac{1}{4}\sum_{i<j}(p_i p_j - 2p_i^2 p_j - 2p_i p_j^2 + 6p_i^2 p_j^2)u_i^2 u_j^2 \\
& + \sum_{\substack{i<j<l \\ j<i<l \\ j<l<i}}(p_i p_j p_l - 3p_i^2 p_j p_l)u_i^2 u_j u_l \\
& \left. + 6\sum_{i<j<l<m} p_i p_j p_l p_m u_i u_j u_l u_m\right] + \cdots
\end{aligned}$$

Hence φ can be written in the form

$$\varphi = e^{\frac{1}{2}\left[\sum_{i=1}^{k}(p_i - p_i^2)u_i^2 - 2\sum_{i<j} p_i p_j u_i u_j\right]}\left[1 + \frac{A_1}{\sqrt{n}} + \frac{A_2}{n} + \cdots\right], \qquad (2)$$

where A_1 is the coefficient of $1/\sqrt{n}$ in (1), A_2 is the sum of the coefficient of $1/n$ and $A_1^2/2$, etc.

3. **The Multinomial Distribution Function.** If a distribution function can be expressed as

$$f(x_1, \cdots, x_k) = f_0 - \sum_{i=1}^{k} \alpha_i \frac{\partial f_0}{\partial x_i} + \sum_{i,j=1}^{k} \beta_{ij} \frac{\partial^2 f_0}{\partial x_i \partial x_j} - \cdots, \tag{3}$$

where f_0 is of the form $c_0 e^{-\frac{1}{2}\sum c_{ij}x_i x_j}$ with $|c_{ij}|$ positive definite, then its generating function[3] can be written as

$$F(u_1, \cdots, u_k) = F_0\left[1 + \sum_{i=1}^{k} \alpha_i u_i + \sum_{i,j=1}^{k} \beta_{ij} u_i u_j + \cdots\right], \tag{4}$$

where F_0 is the generating function of f_0 and is of the form $e^{\frac{1}{2}\sum a_{ij}u_i u_j}$ with $|a_{ij}|$ positive definite. Conversely,[4] if the generating function of a continuous distribution is of the form (4), then the distribution function can be expressed by means of (3). This relationship may be applied to (2) since it can be shown to be of the form (4).

The coefficients c_{ij} of f_0 corresponding to φ can be determined by making use of the fact that the moments of f_0 can be evaluated directly by integration or indirectly by differentiation of φ_0. It is sufficient here to equate expressions for second moments; thus

$$\left.\frac{\partial^2 \varphi_0}{\partial u_s \partial u_t}\right|_{u_i=0} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_s x_t c_0 e^{-\frac{1}{2}\sum c_{ij}x_i x_j} dx_1 \cdots dx_k.$$

Now

$$\left.\frac{\partial^2 \varphi_0}{\partial u_s \partial u_t}\right|_{u_i=0} = \begin{cases} -p_s p_t, & s \neq t \\ p_s - p_s^2, & s = t. \end{cases}$$

The value of the integral is known[5] to be c^{st}, the reciprocal of the element c_{st} in the determinant $|c_{ij}|$. Hence

$$c^{st} = \begin{cases} -p_s p_t, & s \neq t \\ p_s - p_s^2, & s = t. \end{cases}$$

But c_{st} can be obtained from c^{st}, since it is given by the reciprocal of c^{st}. Thus $c_{st} = \hat{c}^{st}/|c^{st}|$, where $\hat{c}^{st}$ denotes the cofactor of element c^{st} in $|c^{st}|$.

---

[3] Darmois, loc. cit., p. 242.

[4] See, for example, S. Kullback, *Annals of Mathematical Statistics*, vol. 5 (1934), pp. 263–307.

[5] See, for example, Risser and Traynard, *Les Principes de la Statistique Mathematique*, p. 226.

$$|c^{st}| = \begin{vmatrix} p_1 - p_1^2 & -p_1p_2 & \cdots & -p_1p_k \\ & p_2 - p_2^2 & \cdots & -p_2p_k \\ & & \ddots & \vdots \\ & & & p_k - p_k^2 \end{vmatrix}$$

$$= (-1)^k p_1^2 p_2^2 \cdots p_k^2 \begin{vmatrix} 1 - \frac{1}{p_1} & 1 & \cdots & 1 \\ & 1 - \frac{1}{p_2} & \cdots & 1 \\ & & \ddots & \vdots \\ & & & 1 - \frac{1}{p_k} \end{vmatrix}$$

This determinant may be evaluated by subtracting the last column from each of the others and then expanding by minors of the last row. Thus

$$|c^{st}| = (-1)^k p_1^2 \cdots p_k^2 \left[(-1)^k \frac{1 - \sum_{i=1}^{k} p_i}{p_1 \cdots p_k} \right] = p_1 p_2 \cdots p_k p,$$

since $\sum_{i=1}^{k} p_i = 1 - p$ from probability considerations. To evaluate $\hat{c}^{st}$, delete row s and column t in $|c^{st}|$, then shift row t to the last row and column s to the last column. These shifts, together with the sign of the cofactor, change the sign of the resulting expression; hence

$$\hat{c}^{st} = -(-1)^{k-1} \frac{p_1^2 \cdots p_k^2}{p_s p_t} \begin{vmatrix} 1 - \frac{1}{p_1} & 1 & \cdots & 1 \\ & 1 - \frac{1}{p_2} & \cdots & 1 \\ & & \ddots & \vdots \\ & & 1 - \frac{1}{p_k} & 1 \\ & & & 1 \end{vmatrix} = p_1 p_2 \cdots p_k,$$

provided $s \neq t$. Since $\hat{c}^{ss}$ is merely $|c^{st}|$ after row s and column s have been deleted, it may be evaluated exactly as was $|c^{st}|$. Thus

$$\hat{c}^{ss} = (-1)^{k-1} \frac{p_1^2 \cdots p_k^2}{p_s^2} \left[(-1)^{k-1} p_s \frac{1 - \sum_{i \neq s}^{k} p_i}{p_1 \cdots p_k} \right] = p_1 p_2 \cdots p_k \left(1 + \frac{p}{p_s} \right).$$

Combining these results, $c_{st} = \frac{1}{p}$ for $s \neq t$ and $c_{ss} = \frac{1}{p} + \frac{1}{p_s}$; and therefore

$$f_0 = c_0 e^{-\frac{1}{2}\left[\sum_{i=1}^{k}\left(\frac{1}{p}+\frac{1}{p_i}\right)x_i^2+\frac{2}{p}\sum_{i<j}x_i x_j\right]}. \tag{5}$$

By computing the necessary derivatives of f_0, the explicit form of f, given by (3), can be obtained to the desired number of terms. Since such derivatives contain f_0 as a factor, f may be written as

$$f = f_0\left[1 - \frac{B_1}{\sqrt{n}} + \frac{B_2}{n} - \cdots\right], \tag{6}$$

where B_i is obtained from A_i of (2) by replacing terms in the u_i with the corresponding derivative of f_0 and then factoring out f_0.

4. **The Generating Function of χ^2.** Let this function be denoted by

$$G(t) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} e^{t\chi^2} f_0\left[1 - \frac{B_1}{\sqrt{n}} + \frac{B_2}{n} - \cdots\right] dx_1 \cdots dx_k.$$

Now

$$\chi^2 = \sum_{i=1}^{k+1}\frac{(\alpha_i - np_i)^2}{np_i} = \sum_{i=1}^{k+1}\frac{x_i^2}{p_i} = \sum_{i=1}^{k}\left(\frac{1}{p}+\frac{1}{p_i}\right)x_i^2 + \frac{2}{p}\sum_{i<j}x_i x_j;$$

consequently χ^2 is, except for a factor of $-\frac{1}{2}$, the quadratic form in f_0. Accordingly, letting $\theta = 1 - 2t$,

$$e^{t\chi^2} f_0 = e^{t\chi^2} c_0 e^{-\frac{1}{2}\chi^2} = c_0 e^{-\frac{1}{2}\theta\chi^2};$$

and hence

$$G(t) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} c_0 e^{-\frac{1}{2}\theta\chi^2}\left[1 - \frac{B_1}{\sqrt{n}} + \frac{B_2}{n} - \cdots\right] dx_1 \cdots dx_k.$$

Letting $z_i = x_i\sqrt{\theta}$ and denoting the value of B_i after this substitution by C_i,

$$\begin{aligned} G(t) &= \theta^{-\frac{1}{2}k}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f_0\left[1 - \frac{C_1}{\sqrt{n}} + \frac{C_2}{n} - \cdots\right] dz_1 \cdots dz_k, \\ &= \theta^{-\frac{1}{2}k}\left[1 + \frac{1}{n}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f_0 C_2\, dz_1 \cdots dz_k + \cdots\right], \end{aligned} \tag{7}$$

since the terms involving odd powers of $1/\sqrt{n}$ are of odd degree in the z_i and therefore vanish upon integration.

For the purposes of this paper only the integral which is the coefficient of $1/n$ needs to be evaluated. Since the algebra involved in this evaluation is heavy and the formulas become exceedingly long, only a few terms will be written out explicitly to indicate the procedure followed.

From (1), (2), and (6) it is clear that only fourth and sixth order derivatives of f_0 are needed. As examples

$$\frac{\partial^4 f_0}{\partial x_i^4} = f_0\left[D_i^4 - 6D_i^2\left(\frac{1}{p}+\frac{1}{p_i}\right) + 3\left(\frac{1}{p}+\frac{1}{p_i}\right)^2\right],$$

$$\frac{\partial^6 f_0}{\partial x_i^6} = f_0\left[D_i^6 - 15D_i^4\left(\frac{1}{p}+\frac{1}{p_i}\right) + 45D_i^2\left(\frac{1}{p}+\frac{1}{p_i}\right)^2 - 15\left(\frac{1}{p}+\frac{1}{p_i}\right)^3\right],$$

where $D_i = \frac{1}{p}\left[x_1 + \cdots + \left(1+\frac{p}{p_i}\right)x_i + \cdots + x_k\right]$. Following the procedure indicated in (6) and (7), this integral becomes

$$\begin{aligned}
&\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f_0\Bigg\{\frac{1}{24}\sum_{i=1}^{k}[p_i - 7p_i^2 + 12p_i^3 - 6p_i^4] \\
&\qquad\left[\frac{D_i^4}{\theta^2} - 6\frac{D_i^2}{\theta}\left(\frac{1}{p}+\frac{1}{p_i}\right) + 3\left(\frac{1}{p}+\frac{1}{p_i}\right)^2\right] \\
&\qquad + (\text{similar terms of degree 4 and lower in the } D_i) \\
&\qquad + \frac{1}{72}\sum_{i=1}^{k}[p_i^2 - 6p_i^3 + 13p_i^4 - 12p_i^5 + 4p_i^6] \\
&\qquad\left[\frac{D_i^6}{\theta^3} - 15\frac{D_i^4}{\theta^2}\left(\frac{1}{p}+\frac{1}{p_i}\right) + 45\frac{D_i^2}{\theta}\left(\frac{1}{p}+\frac{1}{p_i}\right)^2 - 15\left(\frac{1}{p}+\frac{1}{p_i}\right)^3\right] \\
&\qquad + (\text{similar terms of degree 6 and lower in the } D_i)\Bigg\}\, dz_1\cdots dz_k.
\end{aligned} \tag{8}$$

When $\theta = 1$, the integral reduces to that of $f_0 B_2$, which in turn is the integral of a linear combination of derivatives of f_0. But the integral of such a derivative vanishes. As a result, if the integral of $f_0 D_i^2$ has been computed directly, that of $f_0 D_i^4$ and then that of $f_0 D_i^6$ can be found indirectly by equating the corresponding bracket to zero for $\theta = 1$. Similarly for the other terms of the above integral. As examples

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f_0 D_i^2\, dz_1\cdots dz_k = \frac{1}{p}+\frac{1}{p_i},$$

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f_0 D_i^4\, dz_1\cdots dz_k = 3\left(\frac{1}{p}+\frac{1}{p_i}\right)^2.$$

Upon evaluating all such integrals, (8) reduces to

$$\begin{aligned}
&\frac{1}{24}\sum_{i=1}^{k}(p_i - 7p_i^2 + 12p_i^3 - 6p_i^4)\,3\left(\frac{1}{p}+\frac{1}{p_i}\right)^2\left(\frac{1}{\theta}-1\right)^2 \\
&\qquad + \left(\text{similar terms all containing } \left(\frac{1}{\theta}-1\right)^2 \text{ as a factor}\right) \\
&\qquad + \frac{1}{72}\sum_{i=1}^{k}(p_i^2 - 6p_i^3 + 13p_i^4 - 12p_i^5 + 4p_i^6)\,15\left(\frac{1}{p}+\frac{1}{p_i}\right)^3\left(\frac{1}{\theta}-1\right)^3 \\
&\qquad + \left(\text{similar terms all containing } \left(\frac{1}{\theta}-1\right)^3 \text{ as a factor}\right).
\end{aligned} \tag{9}$$

In order to interpret these results, it is necessary to condense these various sums of probability expressions. If the terms are arranged in descending powers of the p_i, it will be discovered that certain combinations condense readily. The condensation in each case lies in recognizing combinations like

$$\sum_{i=1}^{k} p_i^4 + 4\sum_{i\neq j} p_i^3 p_j + 6\sum_{i<j} p_i^2 p_j^2 + 12\sum_{i<j<l} p_i^2 p_j p_l + 24\sum_{i<j<l<m} p_i p_j p_l p_m = \left(\sum_{i=1}^{k} p_i\right)^4.$$

However, some of the terms resulting from multiplying by $1/p_i$ above cannot be condensed in this fashion until they have been reduced to familiar sums by using relationships of the following type:

$$\sum_{i<j} p_i p_j\left(\frac{1}{p_i}+\frac{1}{p_j}\right) = (k-1)\sum_{i=1}^{k} p_i,$$

$$\sum_{i<j<l} p_i p_j p_l\left(\frac{1}{p_i}+\frac{1}{p_j}+\frac{1}{p_l}\right) = (k-2)\sum_{i<j} p_i p_j.$$

After all possible condensations have been made, (9) reduces to

$$\left(\frac{1}{\theta}-1\right)^2\frac{1}{8}\left[\sum_{i=1}^{k+1}\frac{1}{p_i} - (k^2+4k+1)\right] + \left(\frac{1}{\theta}-1\right)^3\frac{1}{24}\left[5\sum_{i=1}^{k+1}\frac{1}{p_i} - (3k^2+12k+5)\right].$$

As a result, the generating function of χ^2 can be written as

$$G(t) = \theta^{-\frac{1}{2}k} + \frac{S_1}{n}(\theta^{-\frac{1}{2}(k+4)} - 2\theta^{-\frac{1}{2}(k+2)} + \theta^{-\frac{1}{2}k}) + \frac{S_2}{n}(\theta^{-\frac{1}{2}(k+6)} - 3\theta^{-\frac{1}{2}(k+4)} + 3\theta^{-\frac{1}{2}(k+2)} - \theta^{-\frac{1}{2}k}) + (\text{terms involving higher powers of } 1/n), \tag{10}$$

where $S_1 = \frac{1}{8}\left[\sum_{i=1}^{k+1}\frac{1}{p_i} - (k^2+4k+1)\right]$ and $S_2 = \frac{1}{24}\left[5\sum_{i=1}^{k+1}\frac{1}{p_i} - (3k^2+12k+5)\right]$.

5. **The Distribution Function of χ^2.** It is well known that $\theta^{-\frac{1}{2}k} = (1-2t)^{-\frac{1}{2}k}$ is the generating function of what is commonly called the χ^2 distribution function with k degrees of freedom. If this distribution function is denoted by $F_k(\chi^2)$, then the distribution function corresponding to (10) can be written as

$$F_k + \frac{S_1}{n}(F_{k+4} - 2F_{k+2} + F_k) + \frac{S_2}{n}(F_{k+6} - 3F_{k+4} + 3F_{k+2} - F_k) + (\text{terms involving higher powers of } 1/n). \tag{11}$$

The customary test for goodness of fit involves the integral of $F_k(\chi^2)$ from χ to ∞, which has been tabled for values of χ^2 and k. The form of (11) is such that the integral of the term in $1/n$ is easily evaluated by means of this same table. However, for more accurate results and for theoretical reasons, it is more elucidating to express these integrals in a more compact form. This is accomplished by using familiar[6] expansions for the integral of $F_k(\chi^2)$. Denoting the integral of the explicit terms of (11) by P, it is easy to show that

$$P = P_1 + \frac{1}{n}[R_1 S_1 + R_2 S_2], \tag{12}$$

where P_1 is the customary tabled value for k degrees of freedom and

$$R_1 = \frac{e^{-\frac{1}{2}\chi^2}\chi^k}{2.4 \cdots (k+2)}[\chi^2 - (k+2)],$$
$$R_2 = \frac{e^{-\frac{1}{2}\chi^2}\chi^k}{2.4 \cdots (k+4)}[\chi^4 - 2(k+4)\chi^2 + (k+4)(k+2)], \tag{13}$$

for k even, while for k odd both R_1 and R_2 contain an additional factor of $\sqrt{2/\pi}$ and have $1.3 \cdots (k+2)$ and $1.3 \cdots (k+4)$ respectively for denominators.

6. **Conclusions.** In any given problem the second approximation P can be calculated easily by means of either (11) or (12) and compared with the customary first approximation P_1. However, the magnitude of this correction term is of primary interest when χ^2 is near a significance level and when one or more of n, k, and p_i is small because the accuracy of P_1 is questioned in those cases.

For χ^2 at the .05 level and for $2 \leqq k \leqq 16$, it is easily shown that $0 < R_1 < .08$ and $-.08 < R_2 < 0$. Clearly S_2 is positive, while S_1 will be positive if one or more of the p_i is sufficiently small. Consequently, for those cases of particular interest, the correction term is surprisingly small partly because R_1 and R_2 are so small and partly because they are of opposite sign.

To illustrate this viewpoint consider the following numerical example. Let $n = 10$, $k = 4$, $\chi^2 = 9.488$, $p_1 = p_2 = \frac{1}{20}$, $p_3 = \frac{4}{20}$, $p_4 = \frac{6}{20}$, $p_5 = \frac{8}{20}$. Then $S_1 = 2.23$, $S_2 = 6.38$, $R_1 = .056$, $R_2 = -.027$, $P_1 = .05$, and $P = .045$. The correction term of $-.005$ is very small in spite of the fact that this example is an extreme case to which the customary χ^2 test would not be applied.

As judged by the second order approximation obtained in this paper, the actual error comitted by using the customary first approximation is much smaller than the order of the neglected terms would indicate, and therefore the range of applicability of P_1 is wider than has been supposed. However, this investigation has considered only the error due to rough approximations and leaves untouched the second type of error indicated in the introductory paragraph.

OREGON STATE COLLEGE

[6] Risser and Traynard, loc. cit., p. 251.