AUSTRIAN JOURNAL OF STATISTICS
Volume 35 (2006), Number 2&3, 299–306

On Formulating Pearson's Chi-Squared Statistic in Two-Way Frequency Tables

Lynn Roy LaMotte
LSU School of Public Health, New Orleans, U.S.A.

Abstract: The standard form of Pearson's chi-squared statistic ignores variation due to estimating the mean vector in settings where the mean vector is not completely specified by the null hypothesis, as is the case when testing for homogeneity or independence in two-way tables. The root form of the statistic is formulated here with and without that additional variance included, resulting in somewhat different expressions.

Keywords: Approximate p-values, Independence, Homogeneity.

1 Introduction

Pearson (1900) derived the probability distribution of

$$\chi^2 = U^{\mathrm{T}}\{\mathrm{var}(U)\}^{-1}U\,, \tag{1}$$

where U is an n-variate normal random variable with mean 0 and variance-covariance matrix $\mathrm{var}(U)$. Although his notation and terminology are different – he wrote in terms of "a system of deviations from the means of n variables" – it is clear that this was his starting point.

After this derivation, he applied the results "to the problem of the fit of an observed to a theoretical frequency distribution". This involved substituting the first η of $\eta+1$ differences $o_i - e_i$, $i = 1, \ldots, \eta+1$, between observed frequencies o_i and known theoretical frequencies e_i in place of U. After considerable algebra, he arrived at the now-ubiquitous formulation for comparing observed and expected frequencies,

$$\chi^2 = \sum_{i=1}^{\eta+1} \frac{(o_i - e_i)^2}{e_i}\,. \tag{2}$$

In his Illustration II, Pearson used the observed total number of fives and sixes in 26,306 tosses of twelve dice to estimate the binomial parameter, from which estimated theoretical frequencies $\hat{e}_i$ for each of the thirteen possibilities $i = 0, \ldots, 12$ were then computed. The differences $o_i - e_i$ became $o_i - \hat{e}_i$. Still, following a two-plus-page rationalization, he used (2) directly, simply replacing e_i by $\hat{e}_i$, instead of starting with the basic form (1) with $u = (o - \hat{e}) - E(O - \hat{e})$ and $\mathrm{var}(U) = \mathrm{var}(O - \hat{e})$ (denoting by O the random variable of which o is the realization). Derivations of $E(O - \hat{e})$ and $\mathrm{var}(O - \hat{e})$ in the context of Illustration II appear to be complicated, and they involve a nuisance parameter for which a value would have to be substituted in order to derive a formulation from (1). Something much more complicated than (2) might have resulted, which likely would not have stuck so well.

Pearson's justification for using (1) when the parent distribution must be estimated from the sample is based on having "a fairly numerous series". In recent decades, much attention has been given to the accuracy of probability approximations related to (2), and to so-called exact probability computations, in settings where sample sizes are not large. See Agresti (2001). Although the differences shown in the next sections are inconsequential for large samples, for small samples they can affect the accuracy of approximate p-values based on (2).

We shall examine the formulation of (1) in the familiar setting of two-way tables of frequencies, testing for independence or homogeneity when conditioning on both row and column marginal totals, one set of marginal totals, or neither. We shall see that (1) leads to (2) in some settings but not in others.

It is hoped that the notation used here and the background are familiar. Notation is defined, and some useful results given, in Appendix A.

Denote by F an $r \times c$ table of frequencies arising from n independent observations on the bivariate random variable (I, J), taking pairs of values $\{(i, j) : i = 1, \dots, r, j = 1, \dots, c\}$. The entry F_{ij} of F is the number of times that the pair (i, j) occurs among the n trials. Denote column sums and row sums by $N = (n_1, \dots, n_c)^{\mathrm{T}}$ and $M = (m_1, \dots, m_r)^{\mathrm{T}}$, respectively. Re-express F as a vector as $f = \mathrm{vec}(F)$. In this form, we shall use the symbol f to denote both the random variable and its realized value.

Let Π denote the $r \times c$ matrix with ij-th entry $\pi_{ij} = \mathrm{Pr}(I = i, J = j)$, $i = 1, \dots, r$, $j = 1, \dots, c$. Let π_R and π_C denote the row and column marginal probability distributions of (I, J), and let $\hat{\pi}_R = M/n$ and $\hat{\pi}_C = N/n$.

2 Independent Observations

With n independent observations, f follows a multinomial distribution with parameters n and $\mathrm{vec}(\Pi)$. Under the null hypothesis that the row and column categories are independent, $\Pi = \pi_R \pi_C^{\mathrm{T}}$. Estimating expected frequencies under the null hypothesis by $n\hat{\pi}_{Ri}\hat{\pi}_{Cj}$ and substituting directly into (2) gives

$$Q_0 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(F_{ij} - n\hat{\pi}_{Ri}\hat{\pi}_{Cj})^2}{n\hat{\pi}_{Ri}\hat{\pi}_{Cj}} \,. \tag{3}$$

It is possible that some $\hat{\pi}_{Ri}$ or $\hat{\pi}_{Cj}$ is 0, leaving the term that involves it undefined. Using (2) directly provides no resolution of this problem, but starting with Pearson's root form (1) resolves it neatly, as shown below.

Although this is now one of the most routine applications of Pearson's chi-squared statistic, the setting is different from the one for which Pearson derived the statistic. Not only must the expected frequencies be estimated, but a whole family of distributions of f satisfies the null hypothesis. The hypothesis of independence itself may be considered to be different from a hypothesis of goodness-of-fit. Still, it is true that, under the null hypothesis, the expected value of F_{ij} is $\pi_{Ri}\pi_{Cj}$, and so Q_0 looks like (2) with e_i estimated from the observed frequencies.

Now let us derive the statistic based on f, beginning with Pearson's original starting point (1). The vector f of frequencies follows a multinomial distribution with expected value $n\mathrm{vec}(\Pi)$ and variance-covariance matrix $n\{\mathrm{Diag}[\mathrm{vec}(\Pi)] - \mathrm{vec}(\Pi)\mathrm{vec}(\Pi)^{\mathrm{T}}\}$. Under the null hypothesis, $\Pi = \pi_R\pi_C^{\mathrm{T}}$, and so $\mathrm{vec}(\Pi) = \pi_C \otimes \pi_R$. Then the variance-covariance matrix of f under the null hypothesis is

$$\mathrm{var}_0(f) = n\{\mathrm{Diag}(\pi_C \otimes \pi_R) - (\pi_C \otimes \pi_R)(\pi_C \otimes \pi_R)^{\mathrm{T}}\}\,. \tag{4}$$

Substituting into (1), for any f, π_C, and π_R such that $f - n\pi_C \otimes \pi_R \in \mathrm{sp}\{\mathrm{var}_0(f)\}$, Proposition 3 in Appendix A shows that

$$\begin{aligned}
&(f - n\pi_C \otimes \pi_R)^{\mathrm{T}}\{\mathrm{var}_0(f)\}^{+}(f - n\pi_C \otimes \pi_R) \\
&\quad = (f - n\pi_C \otimes \pi_R)^{\mathrm{T}}\{n\mathrm{Diag}(\pi_C \otimes \pi_R)\}^{+}(f - n\pi_C \otimes \pi_R) \\
&\quad = \sum_{i=1}^{r}\sum_{j=1}^{c}(F_{ij} - n\pi_{Ri}\pi_{Cj})^2(n\pi_{Ri}\pi_{Cj})^{+}\,.
\end{aligned} \tag{5}$$

Values of the nuisance parameters π_R and π_C are not specified in the null hypothesis. The question, which values should be used for them, is interesting, but we shall simply substitute the estimates $\hat{\pi}_R$ and $\hat{\pi}_C$ for them. These preserve the desired asymptotic properties of the statistic, while others might not. (Cramér (1946, p. 426 and p. 442) notes that these estimates result from the "modified χ^2 minimum method".) Call the resulting statistic Q_1; with these substitutions, it is the same as Q_0.

Proposition 3 requires the inclusion relation $z \in \mathrm{sp}(D - DAD)$, which here becomes $f - n\pi_C \otimes \pi_R \in \mathrm{sp}\{\mathrm{var}_0(f)\}$. That it be satisfied here requires only that $F_{ij} - n\pi_{Ri}\pi_{Cj} = 0$ whenever $\pi_{Ri}\pi_{Cj} = 0$, that is, that $\pi_{Ri}\pi_{Cj} = 0$ implies that $F_{ij} = 0$, which is satisfied when $\hat{\pi}_R$ and $\hat{\pi}_C$ are used. This fact is used in subsequent developments, but it will not be discussed again.

It is straightforward to show that $E_0(f - n\hat{\pi}_C \otimes \hat{\pi}_R) = 0$. One way to derive an expression for its variance is to uncondition (9). This gives

$$\mathrm{var}_0(f - n\hat{\pi}_C \otimes \hat{\pi}_R) = \left(\frac{n-1}{n}\right) n\{\mathrm{Diag}(\pi_C) - \pi_C\pi_C^{\mathrm{T}}\} \otimes \{\mathrm{Diag}(\pi_R) - \pi_R\pi_R^{\mathrm{T}}\}\,. \tag{6}$$

Using $f - n\hat{\pi}_C \otimes \hat{\pi}_R$ and its estimated variance-covariance matrix in (1) results in

$$\begin{aligned}
Q_2 &= (f - n\hat{\pi}_C \otimes \hat{\pi}_R)^{\mathrm{T}}\{\hat{\mathrm{v}}\mathrm{ar}_0(f - n\hat{\pi}_C \otimes \hat{\pi}_R)\}^{+}(f - n\hat{\pi}_C \otimes \hat{\pi}_R) \\
&= \left(\frac{n}{n-1}\right) Q_0\,,
\end{aligned} \tag{7}$$

upon applying Proposition 3 again.

3 Conditioning on Column Totals

Now consider the same setting, but conditional on the column totals. Denote the conditional expected value of the j-th column of F by $n_j\mu_j$. The null hypothesis is $\mathrm{H}_0 : \mu_1 = \cdots = \mu_c$. Under H_0, denote this common mean vector by μ_0, and note that it is the same as π_R.

Conventionally, Q_0 is used in this setting to test H_0. We shall derive Pearson's statistic from (1) based on f and on $f - \hat{E}_0(f|N)$. The conditional distribution of f is the product of multinomial distributions, and under H_0 all have the same vector of category probabilities π_R. Denote the j-th column of F by f_j, so that $f^{\mathrm{T}} = (f_1^{\mathrm{T}}, \ldots, f_c^{\mathrm{T}})$. Then $E_0(f_j|N) = n_j\mu_0 = n\hat{\pi}_{Cj}\pi_R$, and so we see that $f - E_0(f|N) = f - n\hat{\pi}_C \otimes \pi_R$. The variance-covariance matrix of f is block-diagonal, the j-th diagonal block being $\mathrm{var}_0(f_j|N) = n_j\{\mathrm{Diag}(\pi_R) - \pi_R\pi_R^{\mathrm{T}}\}$, that is,

$$\mathrm{var}_0(f|N) = n\mathrm{Diag}(\hat{\pi}_C) \otimes \{\mathrm{Diag}(\pi_R) - \pi_R\pi_R^{\mathrm{T}}\}\,. \tag{8}$$

Noting that, conditional on the column sums N, $f - n\hat{\pi}_C \otimes \hat{\pi}_R$ is a linear transformation of f, it may be shown that

$$\mathrm{var}_0(f - n\hat{\pi}_C \otimes \hat{\pi}_R|N) = n\{\mathrm{Diag}(\hat{\pi}_C) - \hat{\pi}_C\hat{\pi}_C^{\mathrm{T}}\} \otimes \{\mathrm{Diag}(\pi_R) - \pi_R\pi_R^{\mathrm{T}}\}. \tag{9}$$

With these expressions, substituting $\hat{\pi}_R$ for π_R and using Proposition 3, we have

$$\begin{aligned} Q_3 &= (f - n\hat{\pi}_C \otimes \hat{\pi}_R)^{\mathrm{T}}\{\hat{\mathrm{v}}\mathrm{ar}_0(f|N)\}^{+}(f - n\hat{\pi}_C \otimes \hat{\pi}_R) \\ &= Q_0 \end{aligned}$$

and

$$\begin{aligned} Q_4 &= (f - n\hat{\pi}_C \otimes \hat{\pi}_R)^{\mathrm{T}}\{\hat{\mathrm{v}}\mathrm{ar}_0(f - n\hat{\pi}_C \otimes \hat{\pi}_R|N)\}^{+}(f - n\hat{\pi}_C \otimes \hat{\pi}_R) \\ &= Q_0\,. \end{aligned}$$

In this setting, conditioning on column totals, whether Pearson's statistic (1) is constructed from f, $\hat{E}_0(f|N)$, and $\hat{\mathrm{v}}\mathrm{ar}_0(f|N)$ or from $f - n\hat{\pi}_C \otimes \hat{\pi}_R$ and its estimated variance-covariance matrix, the result is Q_0 either way.

4 Conditioning on Both Row and Column Totals

Finally, consider the setting conditional on both row and column totals. The null hypothesis is that the c columnwise conditional expected values $\mu_1, \ldots, \mu_c$ are all the same. With $\hat{\pi}_C$ and $\hat{\pi}_R$ fixed, $\mathrm{var}_0(f|M,N)$ and $\mathrm{var}_0(f - n\hat{\pi}_C\hat{\pi}_R|M,N)$ are the same. An expression for $\mathrm{var}_0(f|M,N)$ can be derived, with care and perseverance, directly from the conditional distribution of f. It is

$$\begin{aligned} \mathrm{var}_0(f|M,N) &= \mathrm{var}_0(f - n\hat{\pi}_C \otimes \hat{\pi}_R|M,N) \\ &= \left(\frac{n}{n-1}\right) n\{\mathrm{Diag}(\hat{\pi}_C) - \hat{\pi}_C\hat{\pi}_C^{\mathrm{T}}\} \otimes \{\mathrm{Diag}(\hat{\pi}_R) - \hat{\pi}_R\hat{\pi}_R^{\mathrm{T}}\}\,. \end{aligned} \tag{10}$$

Note that, due to the conditioning, there are no nuisance parameters; the conditional distribution of f is completely specified under H_0. From (1), the statistic is

$$\begin{aligned} Q_5 &= (f - n\hat{\pi}_C\hat{\pi}_R)^{\mathrm{T}}\{\mathrm{var}_0(f|M,N)\}^{+}(f - n\hat{\pi}_C\hat{\pi}_R) \\ &= \left(\frac{n-1}{n}\right)(f - n\hat{\pi}_C\hat{\pi}_R)^{\mathrm{T}}\{n\mathrm{Diag}(\hat{\pi}_C \otimes \hat{\pi}_R)\}^{+}(f - n\hat{\pi}_C\hat{\pi}_R) \\ &= \left(\frac{n-1}{n}\right) Q_0\,. \end{aligned} \tag{11}$$

Steyn and Stumpf (1984) show (10) in their discussion of asymptotics, but it is the same as the expressions they derived earlier (pp. 144–145), rearranged in the compact form of Kronecker products of matrices. For 2×2 tables, Upton (1982) based an adjustment to the chi-squared statistic on the ratio $n/(n-1)$, implicitly following (1), basing the statistic on the difference between the two sample proportions and using the variance of that difference conditioning on both row and column totals.

Conditional on M and N, the distribution of Q_0 is completely specified by the null hypothesis, but intractable. Because Q_5 is a quadratic form in $f - n\hat{\pi}_C\hat{\pi}_R$, its expected value can be derived directly from the first two moments of f:

$$\begin{aligned} E_0(Q_5|M,N) &= \mathrm{tr}[(n\hat{D}_C \otimes \hat{D}_R)^{+} n\{(\hat{D}_C - \hat{\pi}_C\hat{\pi}_C^{\mathrm{T}}) \otimes (\hat{D}_R - \hat{\pi}_R\hat{\pi}_R^{\mathrm{T}})\}] \\ &= \mathrm{tr}\{(\hat{D}_C^{+}\hat{D}_C - \hat{D}_C^{+}\hat{\pi}_C\hat{\pi}_C^{\mathrm{T}}) \otimes (\hat{D}_R^{+}\hat{D}_R - \hat{D}_R^{+}\hat{\pi}_R\hat{\pi}_R^{\mathrm{T}})\} \\ &= (c_+ - 1)(r_+ - 1), \end{aligned}$$

where $\hat{D}_C = \mathrm{Diag}(\hat{\pi}_C)$, $\hat{D}_R = \mathrm{Diag}(\hat{\pi}_R)$, and c_+ and r_+ are the numbers of non-zero column and row sums, respectively. Because $Q_0 = n/(n-1)Q_5$,

$$E_0(Q_0|M,N) = \frac{n}{n-1}(c_+ - 1)(r_+ - 1), \tag{12}$$

which agrees with expressions found by Haldane (1940), Steyn and Stumpf (1984), and others. Expressions for the expected values of Q_0 conditional on column sums or unconditional can be derived by unconditioning (12), but they depend on π_R and $\hat{\pi}_C$ or π_C and do not appear to be informative. It is perhaps interesting to note that (12) appears as $n(r-1)(c-1)/(n-1)$ in Haldane (1940) and Steyn and Stumpf (1984). In both references, though, it is clear that the derivation is for conditioning on both row and column totals, and that it is assumed that all row and column totals are positive. Cramér (1946) quotes the same result, first in what appears to be the unconditional setting (p. 443) and again in the column conditional setting (p. 447). The expression (12) is not correct when either or both conditions are relaxed, although there is very little difference except for small n. For example, for a $2 \times c$ table, the approximating chi-squared distribution has expected value $c - 1$. Unconditioning on the row sums, it may be seen that

$$E_0(Q_0|N) = \frac{n}{n-1}(c_+ - 1)(1 - \pi_{R1}^n - \pi_{R2}^n).$$

Depending on $\pi_{R1} = 1 - \pi_{R2}$, the expected value can be anything between 0 at $\pi_{R1} = 0$ or 1 and $n(c_+ - 1)(1 - 1/2^{n-1})/(n-1)$ at $\pi_{R1} = 1/2$.

5 Conclusion

In practice, the form (2) is used in all three of the settings addressed here. If the test statistic is developed from Pearson's root form (1) instead, different statistics result when there is no conditioning or when conditioning on both row and column totals. However, the differences are slight, involving factors $n/(n-1)$ and $(n-1)/n$, respectively, while the basic quadratic form persists throughout.

So far as I have been able to tell, the derivations of (2) from (1) for two-way tables have not appeared before. Rather, the canonical form (2) has simply been applied directly without going back to the root form. Perhaps that's not surprising – Pearson's (1900) derivation of (2) from (1) is a long and busy algebraic argument. The same sort of derivation in two-way tables appears at first to be considerably more complicated. However, by formulating the problem in terms of matrices and exploiting the structures entailed in Propositions 1–3 in the Appendix, the derivations becomes reasonably straightforward.

A Notation and Propositions

Denote transpose of a matrix A by A^{T}. If A has an inverse, denote it by A^{-1}. For a ν-vector a, $\mathrm{Diag}(a) = \mathrm{Diag}(a_1, \ldots, a_\nu)$ denotes the $\nu \times \nu$ diagonal matrix with diagonal elements $a_1, \ldots, a_\nu$. For a real number x, let x^+ be $1/x$ if $x \neq 0$ and 0 if $x = 0$. Denote the Moore-Penrose pseudoinverse of a matrix A by A^+. The Moore-Penrose pseudoinverse of a diagonal matrix $D = \mathrm{Diag}(a)$ is $D^+ = \mathrm{Diag}(a_1^+, \ldots, a_\nu^+)$. A symmetric matrix A has the spectral decomposition $A = P\Lambda P^{\mathrm{T}}$, where Λ is diagonal with diagonal elements that are the eigenvalues of A (repeated according to their multiplicities), and columns of P are corresponding orthonormal eigenvectors. Then $A^+ = P\Lambda^+ P^{\mathrm{T}}$. Denote the linear subspace spanned by the columns of a matrix A by $\mathrm{sp}(A)$. Denote the trace of a square matrix A by $\mathrm{tr}(A)$. Denote the Kronecker product of matrices A and B by $A \otimes B$. For an $r \times c$ matrix A, let $\mathrm{vec}(A)$ denote the column vector formed by concatenating the columns of A vertically, that is, the $((j-1)c + i)$-th entry in $\mathrm{vec}(A)$ is the (i, j)-th entry of A, $i = 1, \ldots, r$, $j = 1, \ldots, c$. Unless specified otherwise, vectors are column vectors, that is, matrices with one column. When a denotes a vector, its i-th element is denoted a_i. For vectors a and b, $\mathrm{vec}(ab^{\mathrm{T}}) = b \otimes a$. For a random variable U and a vector a, $\mathrm{var}(a \otimes U) = (aa^{\mathrm{T}}) \otimes \mathrm{var}(U)$. Denote expected value and variance of the random variable U under a null hypothesis (like independence or homogeneity) H_0 by $E_0(U)$ and $\mathrm{var}_0(U)$. For the most part, a random variable and a realization of it will be distinguished notationally by upper-case and lower-case letters, like U for the random variable and u for a realization. In the case of f, though, this convention is clumsy and so will be violated, and the meanings of expressions like $E_0(f)$ and $\mathrm{var}_0(f)$ must be taken from context.

For propositions 1 and 2 below, let V be an $n \times n$ symmetric nonnegative-definite matrix of rank r, where r is an integer between 1 and n. Note that there exists an $n \times r$ matrix L such that $L^{\mathrm{T}}VL$ is positive definite. One choice is $L = P$, where columns of P comprise a basis for $\mathrm{sp}(V)$. Proposition 1, which may be proved easily, shows that for any such matrix L, columns of VL form a basis for $\mathrm{sp}(V)$.

Proposition 1. Let L be an $n \times r$ matrix. In order that $L^{\mathrm{T}}VL$ be nonsingular it is necessary and sufficient that columns of VL be linearly independent.

Proposition 2. If L_1 and L_2 are $n \times r$ matrices such that $L_1^{\mathrm{T}}VL_1$ and $L_2^{\mathrm{T}}VL_2$ are positive definite, then for any vector $u \in \mathrm{sp}(V)$,

$$(L_1^{\mathrm{T}}u)^{\mathrm{T}}(L_1^{\mathrm{T}}VL_1)^{-1}(L_1^{\mathrm{T}}u) = (L_2^{\mathrm{T}}u)^{\mathrm{T}}(L_2^{\mathrm{T}}VL_2)^{-1}(L_2^{\mathrm{T}}u) .$$

Proposition 2 may be proved by noting that, since VL_1 and VL_2 are both bases for $\mathrm{sp}(V)$, there exists a nonsingular matrix B such that $VL_1 = VL_2B$.

When an n-variate random variable U has mean vector 0 and variance-covariance matrix V having rank r, then positive density of U is restricted to $u \in \mathrm{sp}(V)$. The distribution of U can be represented in terms of the non-degenerate random variable $L^{\mathrm{T}}U$, where L is $n \times r$ and $L^{\mathrm{T}}VL$ is nonsingular. For $u \in \mathrm{sp}(V)$, u and $L^{\mathrm{T}}u$ are one-to-one. For two different such matrices L_1 and L_2, $L_1^{\mathrm{T}}U$ and $L_2^{\mathrm{T}}U$ may be different, but Proposition 2 shows that the indicated quadratic forms are the same. Let P be a matrix with columns that are orthonormal eigenvectors of V corresponding to its positive eigenvalues, so that P is $n \times r$ and $P^{\mathrm{T}}VP$ is nonsingular. Then, by Proposition 2, for any $n \times r$ matrix L such that $L^{\mathrm{T}}VL$ is nonsingular and any $u \in \mathrm{sp}(V)$,

$$(L^{\mathrm{T}}u)^{\mathrm{T}}(L^{\mathrm{T}}VL)^{-1}(L^{\mathrm{T}}u) = (P^{\mathrm{T}}u)^{\mathrm{T}}(P^{\mathrm{T}}VP)^{-1}(P^{\mathrm{T}}u) = u^{\mathrm{T}}V^{+}u\,.$$

Proposition 3. If A and D are symmetric $\nu \times \nu$ matrices and $DAD = DADAD$, then for any vector z in $\mathrm{sp}(D - DAD)$,

$$z^{\mathrm{T}}(D - DAD)^{+}z = z^{\mathrm{T}}D^{+}z\,.$$

Proof. Because $z \in \mathrm{sp}(D - DAD)$, there is a vector u such that $z = (D - DAD)u$. Then

$$z^{\mathrm{T}}(D - DAD)^{+}z = u^{\mathrm{T}}(D - DAD)u\,.$$

And

$$\begin{aligned} z^{\mathrm{T}}D^{+}z &= u^{\mathrm{T}}(D - DAD)D^{+}(D - DAD)u \\ &= u^{\mathrm{T}}(D - DAD - DAD + DADAD)u \\ &= u^{\mathrm{T}}(D - DAD)u = z^{\mathrm{T}}(D - DAD)^{+}z\,. \end{aligned}$$

It can be shown that, for symmetric matrices A and B, $(A \otimes B)^{+} = A^{+} \otimes B^{+}$. With that, it is straightforward to show that Proposition 3 extends to Kronecker products of the form $(D_1 - D_1A_1D_1) \otimes (D_2 - D_2A_2D_2)$, a form shared by all the variance-covariance matrices used in Section 2.

An m-category multinomial random variable f with category probabilities π and sample size n has variance-covariance matrix $n\{\mathrm{Diag}(\pi) - \pi\pi^{\mathrm{T}}\}$, and $\{\ \}$ has the properties of $D - DAD$ in Proposition 3, with $D = \mathrm{Diag}(\pi)$ and $A = 1_m 1_m^{\mathrm{T}}$, where 1_m denotes an m-vector of ones. Pearson avoided the singularity of the distribution of f by using only its first $m - 1$ entries. That is the same as using $L^{\mathrm{T}}U$ with L defined to be the first $m - 1$ columns of the $m \times m$ identity matrix. If all entries in π are positive, then $L^{\mathrm{T}}\mathrm{var}(f)L$ is positive definite. Whether we choose to use the first $m - 1$ components or the last $m - 1$, or to leave out the middle component instead, whatever L we choose such that $L^{\mathrm{T}}\mathrm{var}(f)L$ is nonsingular with the same rank as $\mathrm{var}(f)$,

$$\begin{aligned} \{L^{\mathrm{T}}(f - n\pi)\}^{\mathrm{T}}&[L^{\mathrm{T}}n\{\mathrm{Diag}(\pi) - \pi\pi^{\mathrm{T}}\}L]^{-1}\{L^{\mathrm{T}}(f - n\pi)\} \\ &= (f - n\pi)^{\mathrm{T}}\{n\mathrm{Diag}(\pi)\}^{+}(f - n\pi) \\ &= \sum_{i=1}^{m}(f_i - n\pi_i)^2(n\pi_i)^{+}\,, \end{aligned} \tag{13}$$

so long as $f - n\pi \in \mathrm{sp}[n\{\mathrm{Diag}(\pi) - \pi\pi^{\mathrm{T}}\}]$. If no π_i is 0, then (2) results. If some $\pi_j = 0$, then (13) holds provided that $f_j - n\pi_j = 0$.

References

Agresti, A. (2001). Exact inference for categorical data: recent advances and continuing controversies. *Statistics in Medicine*, *20*, 2709-2722.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton, N.J.: Princeton University Press.

Haldane, J. B. S. (1940). The mean and variance of χ^2, when used as a test of homogeneity, when expectations are small. *Biometrika*, *31*, 346-355.

Pearson, K. (1900). On the criterion that a given set of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos. Mag., Series 5*, *50*, 157-172.

Steyn, H. S., and Stumpf, R. H. (1984). Exact distributions associated with an $h \times k$ contingency table. *South African Statistical Journal*, *18*, 135-159.

Upton, G. J. G. (1982). A comparison of alternative tests for the 2×2 comparative trial. *Journal of the Royal Statistical Society A*, *145*, 86-105.

Author's address:

Lynn Roy LaMotte
LSU Health Sciences Center
School of Public Health
c/o Pennington Biomedical Research Center
2323 Kenilworth Parkway
Baton Rouge, LA 70808, USA

E-mail: lrlamo@cox.net