Scientific Research Publishing

Applied Mathematics, 2025, 16(3), 262-274
https://www.scirp.org/journal/am
ISSN Online: 2152-7393
ISSN Print: 2152-7385

$\frac{1}{n}$-Proportional Correction to Tests of Independence

Jan Vrbik

Mathematics Department, Brock University, St. Catharines, Canada
Email: jvrbik@brocku.ca

How to cite this paper: Vrbik, J. (2025) $\frac{1}{n}$-Proportional Correction to Tests of Independence. *Applied Mathematics*, **16**, 262-274.
https://doi.org/10.4236/am.2025.163013

Received: February 22, 2025
Accepted: March 24, 2025
Published: March 27, 2025

Abstract

It is known that both the Pearson and G-test of independence have the same asymptotic distribution, namely χ^2 with $(K-1)(M-1)$ degrees of freedom, where K and M are the number of levels, respectively, of two attributes. However, when extending the accuracy ot this approximation by $\frac{1}{n}$-proportional terms, the resulting corrections differ quite dramatically. The purpose of this article is to derive each of these corrections and demonstrate their use in practical application.

Keywords

Test of Independence, Asymptotic Distribution, Chi-Square Test, Contingency Tables, G-Test, Kronecker Product

1. Introduction

When testing whether two nominal-scale attributes (one having K discrete values, the other one M) are independent, the traditional (Pearson's) test uses the following test statistic

$$U := \sum_{i=1}^{K}\sum_{j=1}^{M} \frac{\left(X_{i,j} - \frac{X_{i,\bullet}X_{\bullet,j}}{n}\right)^2}{\frac{X_{i,\bullet}X_{\bullet,j}}{n}} \tag{1}$$

where n is the total number of observations, $X_{i,j}$ is the number of those resulting in the i^{th} value of the first attribute and in the j^{th} value of the second attribute, $X_{i,\bullet} := \sum_{j=1}^{M} X_{i,j}$ and $X_{\bullet,j} := \sum_{i=1}^{K} X_{i,j}$. The null hypothesis claims

that the probability of an observation being of the $(i,j)^{\text{th}}$ type is $p_i q_j$ where $\sum_{j=1}^{M} p_i = 1$ and $\sum_{i=1}^{K} q_j = 1$; note that the multinomial probability mass function of the $X_{i,j}$'s is therefore

$$f(\boldsymbol{x}) = \binom{n}{x_{1,1}, x_{1,2}, \cdots, x_{K,M}} (p_1 q_1)^{x_{1,1}} (p_1 q_2)^{x_{1,2}} \cdots (p_K q_M)^{x_{K,M}} \tag{2}$$

having $K + M - 2$ algebraically independent parameters [1].

This model implies that the conditional probability of an observation's second attribute being equal to j, given that the value of its first attribute is i, is given by q_j (for any j and i), *i.e.* that the attributes are independent of each other. The test also assumes that the p_i and q_j probabilities are unknown and are estimated by their maximum-likelihood estimators, namely $\frac{X_{i,\bullet}}{n}$ and $\frac{X_{\bullet,j}}{n}$ respectively. The alternate hypothesis allows the probabilities of the $X_{i,j}$'s to have any non-negative values, as long as these add up to 1, thus resulting in $KM - 1$ algebraically independent parameters. Failing the null hypothesis makes the above test statistic *increase* in value; the critical region is always an upper tail of the corresponding distribution.

Another test statistic proposed decades later for testing the same hypothesis is (using the same notation, for future convenience)

$$U := 2 \sum_{i=1}^{K} \sum_{j=1}^{M} X_{i,j} \ln\left(\frac{n X_{i,j}}{X_{\bullet,j} X_{i,\bullet}} \right) \tag{3}$$

of the so-called **G test** [2] (note that when $X_{i,j} = 0$, which happens occasionally, makes the corresponding term of the last expression equal to 0). It is well known (something we verify shortly) that, in the $n \to \infty$ limit, both Formula (1) and Formula (3) converge to the same expression, whose asymptotic distribution is chi-squared with $(K-1)(M-1)$ degrees of freedom [3].

The main objective of this article is to find $\frac{1}{n}$-proportional correction to this distribution, separately for each of our two test statistics.

Next comes a brief review of the basic mathematical tool used throughout this article.

2. Kronecker Product

In this section a, b, c and d are matrices of any dimensions (a dimension equal to 1 implies a column or row vector; both dimensions equal to 1 represents a scalar). Kronecker product $a \otimes b$ of two matrices is created by multiplying every element of a by the full matrix b and organizing the resulting blocks into a single matrix [4]. This product is clearly non-commutative, associative and distributive over addition; furthermore (assuming conformable dimensions for each matrix multiplication)

$$(a \otimes b)^{\mathrm{T}} = a^{\mathrm{T}} \otimes b^{\mathrm{T}}$$

$$(a\otimes b)^{-1}=a^{-1}\otimes b^{-1}$$

$$a\otimes b\circ c\otimes d=a\cdot c\otimes b\cdot d$$

where both $\circ$ and $\cdot$ indicate matrix multiplication; note that $\cdot$ takes precedence over $\otimes$ which, in turn, takes precedence over $\circ$ (this is the reason for the duplicate notation). When a and b are square matrices (K by K and M by M respectively)

$$\det(a\otimes b)=\det(a)^M\det(b)^K$$

We also need the following Woodbury's identity [5]

$$(a+b\cdot c)^{-1}=a^{-1}-a^{-1}\cdot b\cdot\left(I+c\cdot a^{-1}\cdot b\right)^{-1}\cdot c\cdot a^{-1}$$

where I stands for a (conformable) identity matrix, and the following Sylvester's identity [6]

$$\det(a+b\cdot c)=\det(a)\det\left(I+c\cdot a^{-1}\cdot b\right)$$

3. Asymptotic Theory

In this section we investigate the asymptotic distribution of Formula (1) and Formula (3) when $n\to\infty$, getting the same answer for both of these [7].

3.1. Normal Limit of Formula (2)

To make our task easier, we assume that the sum all probabilities of the Multinomial distribution defined in Formula (2) equals to $r<1$, and thus allowing for the possibility of an extra outcome whose probability is $1-r$, denoting its observed total by W. This substantially simplifies subsequent development by removing the singularity of the variance-covariance (V-C) matrix of the $X_{i,j}$ variables, while letting us reach correct conclusions in the $r\to 1$ limit. Correspondingly modifying Formula (2) is easy: extend the bottom row of the multinomial coefficient by $w=n-\sum_{i=1}^{K}\sum_{j=1}^{M}x_{i,j}$ and further multiply by $(1-r)^w$. Since all variables must add up to n, the resulting $KM+1$ by $KM+1$ V-C matrix is singular, but the singularity disappears when replacing W by $n-\sum_{i=1}^{K}\sum_{j=1}^{M}X_{i,j}$ and considering the distribution of the remaining $X_{i,j}$'s only. The corresponding moments are

$$\mathbb{E}\left(X_{i,j}\right)=np_iq_j$$

$$\mathrm{Var}\left(X_{i,j}\right)=np_iq_j\left(1-p_iq_j\right)$$

$$\mathrm{Cov}\left(X_{i,j},X_{k,m}\right)=-np_iq_jp_kq_m\quad\text{when } i\neq k \text{ or } j\neq m$$

which remains correct even with the extra W. We then define

$$Y_{i,j}:=\frac{X_{i,j}-np_iq_j}{\sqrt{n}} \tag{4}$$

and substitute

$$x_{i,j}=np_iq_j+y_{i,j}\sqrt{n}$$

$$w=-\sum_{i=1}^{K}\sum_{j=1}^{M}y_{i,j}$$

into the *natural logarithm* of Formula (2), after is has been modified by introducing the extra W. Adding ln of the corresponding Jacobian (namely $KM\cdot\ln\sqrt{n}$, since the distribution of the $Y_{i,j}$'s is KM dimensional) and taking the $n\to\infty$ limit of the resulting expression (all limits of this article are routinely done by a computer equipped with Mathematica or a similar software) yields

$$\begin{aligned}&-KM\ \ln\sqrt{2\pi}-\frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{M}\ln\left(p_iq_j\right)-\frac{\ln(1-r)}{2}\\&-\sum_{i=1}^{K}\sum_{j=1}^{M}\frac{y_{i,j}^2}{2p_iq_j}-\frac{\left(\sum_{i=1}^{K}\sum_{j=1}^{M}y_{i,j}\right)^2}{2(1-r)}\end{aligned}\tag{5}$$

This identifies the asymptotic distribution of the $Y_{i,j}$'s to be multivariate Normal, with a probability density function (PDF) whose natural logarithm is given by the last expression. Its linear and quadratic moments agree with those of the original discrete distribution, being equal to 0 for the expected value of each $Y_{i,j}$ and to

$$p_i\delta_{i,k}q_j\delta_{j,m}-p_ip_kq_jq_m\tag{6}$$

for the covariance between $Y_{i,j}$ and $Y_{k.m}$ (for any combination of indices); the corresponding V-C matrix is thus

$$\mathbb{V}=\mathbb{P}\otimes\mathbb{Q}-\boldsymbol{p}\cdot\boldsymbol{p}^{\mathrm{T}}\otimes\boldsymbol{q}\cdot\boldsymbol{q}^{\mathrm{T}}\tag{7}$$

where $\boldsymbol{p}$ and $\boldsymbol{q}$ are column vectors whose elements are the p_i and q_j probabilities (respectively), and $\mathbb{P}$ and $\mathbb{Q}$ are diagonal matrices with the p_i and q_j probabilities (respectively) on the main diagonal. The PDF can then be written, in correspondence with Formula (5), as

$$\frac{\exp\left(-\dfrac{\boldsymbol{Y}^{\mathrm{T}}\circ\mathbb{A}\circ\boldsymbol{Y}}{2}\right)}{(2\pi)^{KM/2}}\sqrt{\det\mathbb{A}}\tag{8}$$

where $\boldsymbol{Y}$ is a column vector whose KM elements are $Y_{i,j}$ (organized in accordance with our Kronecker's product, *i.e.* $Y_{1,1},Y_{1,2},\cdots,Y_{1,M},Y_{2,1},\cdots,Y_{K,M}$),

$$\mathbb{A}=\mathbb{P}^{-1}\otimes\mathbb{Q}^{-1}+\frac{\mathbf{1}_k\cdot\mathbf{1}_k^{\mathrm{T}}\otimes\mathbf{1}_m\cdot\mathbf{1}_m^{\mathrm{T}}}{1-r}=\mathbb{P}^{-1}\otimes\mathbb{Q}^{-1}+\frac{\mathbf{1}_k\otimes\mathbf{1}_m\circ\mathbf{1}_k^{\mathrm{T}}\otimes\mathbf{1}_m^{\mathrm{T}}}{1-r}\tag{9}$$

and

$$\begin{aligned}\det(\mathbb{A})&=\det\left(\mathbb{P}^{-1}\otimes\mathbb{Q}^{-1}\right)\det\left(1+\frac{\mathbf{1}_k^{\mathrm{T}}\otimes\mathbf{1}_m^{\mathrm{T}}\circ\mathbb{P}\otimes\mathbb{Q}\circ\mathbf{1}_k\otimes\mathbf{1}_m}{1-r}\right)\\&=\frac{1}{\left(\prod_{i=1}^{K}p_i\right)^M\left(\prod_{j=1}^{M}q_j\right)^K(1-r)}\end{aligned}\tag{10}$$

the last two implied by Formula (5).

Alternately, they can be derived from Formula (7) using the Woodbury and Silvester identities. Bypassing the details, we only verify that $\mathbb{A}$ is the inverse of $\mathbb{V}$, by multiplying

$$\left(\mathbb{P}^{-1}\otimes\mathbb{Q}^{-1}+\frac{\mathbf{1}_k\cdot\mathbf{1}_k^{\mathrm{T}}\otimes\mathbf{1}_m\cdot\mathbf{1}_m^{\mathrm{T}}}{1-r}\right)\circ\left(\mathbb{P}\otimes\mathbb{Q}-\boldsymbol{p}\cdot\boldsymbol{p}^{\mathrm{T}}\otimes\boldsymbol{q}\cdot\boldsymbol{q}^{\mathrm{T}}\right)$$

$$=\mathbb{I}_k\otimes\mathbb{I}_m-\mathbf{1}_k\cdot\boldsymbol{p}^{\mathrm{T}}\otimes\mathbf{1}_m\cdot\boldsymbol{q}^{\mathrm{T}}+\frac{\mathbf{1}_k\cdot\boldsymbol{p}^{\mathrm{T}}\otimes\mathbf{1}_m\cdot\boldsymbol{q}^{\mathrm{T}}}{1-r}-\frac{r\left(\mathbf{1}_k\cdot\boldsymbol{p}^{\mathrm{T}}\otimes\mathbf{1}_m\cdot\boldsymbol{q}^{\mathrm{T}}\right)}{1-r}$$

$$=\mathbb{I}_k\otimes\mathbb{I}_m$$

where $\mathbb{I}_k$ and $\mathbb{I}_m$ are k by k and m by m identity matrices respectively.

3.2. Asymptotic Distribution of *U*

We now take the $n\to\infty$ limit of Formula (1) and of Formula (3), first replacing $X_{i,j}$ with $Y_{i,j}$ by using Formula (4); this results in the same answer in both cases, namely in

$$\sum_{i=1}^{K}\sum_{j=1}^{M}\frac{Y_{i,j}^2}{p_iq_j}-\sum_{i=1}^{K}\frac{Y_{i,\bullet}^2}{p_i}-\sum_{j=1}^{M}\frac{Y_{\bullet,j}^2}{q_i} \tag{11}$$

where $Y_{\bullet,j}=\sum_{i=1}^{K}Y_{i,j}$ and $Y_{i,\bullet}:=\sum_{j=1}^{M}Y_{i.j}$. To get the corresponding moment generating function (MGF) of U, we need to compute the expected value of $\exp(tU)$, *i.e.*

$$\frac{\int\cdots\int_{-\infty}^{\infty}\exp\left(-\frac{\boldsymbol{Y}^{\mathrm{T}}\circ(\mathbb{A}-2t\mathbb{U})\circ\boldsymbol{Y}}{2}\right)\mathrm{d}\boldsymbol{Y}}{(2\pi)^{KM/2}}\sqrt{\det\mathbb{A}} \tag{12}$$

where

$$\mathbb{U}=\mathbb{P}^{-1}\otimes\mathbb{Q}^{-1}-\mathbb{P}^{-1}\otimes\mathbf{1}_m\cdot\mathbf{1}_m^{\mathrm{T}}-\mathbf{1}_k\cdot\mathbf{1}_k^{\mathrm{T}}\otimes\mathbb{Q}^{-1} \tag{13}$$

which is the matrix version of Formula (11); to see this, it helps to re-write Formula (11) as follows

$$\sum_{i=1}^{K}\sum_{k=1}^{K}\sum_{j=1}^{M}\sum_{m=1}^{M}Y_{i.j}\left(\frac{\delta_{i,k}\delta_{j,m}}{p_iq_j}-\frac{\delta_{i,k}}{p_i}-\frac{\delta_{j,m}}{q_j}\right)Y_{k,m}$$

To find the result of Formula (12) necessitates computing the determinant of $\mathbb{A}-2t\mathbb{U}$; this matrix can be expressed in the following form

$$(1-2t)\mathbb{P}^{-1}\otimes\mathbb{Q}^{-1}+\frac{\mathbf{1}_k\cdot\mathbf{1}_k^{\mathrm{T}}\otimes\mathbf{1}_m\cdot\mathbf{1}_m^{\mathrm{T}}}{1-r}+2t\left[\mathbb{P}^{-1}\otimes\mathbf{1}_m\quad\mathbf{1}_k\otimes\mathbb{Q}^{-1}\right]\circ\begin{bmatrix}\mathbb{I}_k\otimes\mathbf{1}_m^{\mathrm{T}}\\ \mathbf{1}_k^{\mathrm{T}}\otimes\mathbb{I}_m\end{bmatrix} \tag{14}$$

which follows from Formula (9) and Formula (13); the square brackets indicate that the matrix consists of two (and subsequently also of four) blocks.

Pre-multiplying Formula (14) by $\frac{\mathbb{P}\otimes\mathbb{Q}}{1-2t}$, whose determinant is

$$D_1 := \frac{\left(\prod_{i=1}^{K} p_i\right)^M \left(\prod_{j=1}^{M} q_j\right)^K}{(1-2t)^{KM}}$$

leaves us with the task of finding the determinant of

$$\begin{aligned}\mathbb{C} &:= \mathbb{I}_k \otimes \mathbb{I}_m + \frac{\boldsymbol{p}\cdot\mathbf{1}_k^{\mathrm{T}} \otimes \boldsymbol{q}\cdot\mathbf{1}_m^{\mathrm{T}}}{(1-r)(1-2t)} + \frac{2t}{1-2t}\left[\mathbb{I}_k \otimes \boldsymbol{q} \quad \boldsymbol{p} \otimes \mathbb{I}_m\right] \circ \begin{bmatrix} \mathbb{I}_k \otimes \mathbf{1}_m^{\mathrm{T}} \\ \mathbf{1}_k^{\mathrm{T}} \otimes \mathbb{I}_m \end{bmatrix} \\ &:= \mathbb{F} + \frac{2t}{1-2t}\mathbb{G}\circ\mathbb{H}\end{aligned} \tag{15}$$

This needs to be done in two steps: for the determinant of $\mathbb{F}$, we get

$$D_2 := \det \mathbb{F} = 1 + \frac{\mathbf{1}_k^{\mathrm{T}} \otimes \mathbf{1}_m^{\mathrm{T}} \circ \boldsymbol{p} \otimes \boldsymbol{q}}{(1-r)(1-2t)} = 1 + \frac{r}{(1-r)(1-2t)} = \frac{1-2t(1-r)}{(1-r)(1-2t)}$$

while its inverse equals to

$$\mathbb{F}^{-1} = \mathbb{I}_k \otimes \mathbb{I}_m - \frac{\boldsymbol{p}\cdot\mathbf{1}_k^{\mathrm{T}} \otimes \boldsymbol{q}\cdot\mathbf{1}_m^{\mathrm{T}}}{1-2t(1-r)} \tag{16}$$

easily verifiable by simple multiplication, utilizing
$\boldsymbol{p}\cdot\mathbf{1}_k^{\mathrm{T}} \otimes \boldsymbol{q}\cdot\mathbf{1}_m^{\mathrm{T}} \circ \boldsymbol{p}\cdot\mathbf{1}_k^{\mathrm{T}} \otimes \boldsymbol{q}\cdot\mathbf{1}_m^{\mathrm{T}} = r\left(\boldsymbol{p}\cdot\mathbf{1}_k^{\mathrm{T}} \otimes \boldsymbol{q}\cdot\mathbf{1}_m^{\mathrm{T}}\right)$.

Sylvester's identity implies that the determinant of $\mathbb{C}$ is a product of $\det\mathbb{F}$ and the determinant of (at this point, we can start replacing r by 1)

$$\begin{aligned}&\mathbb{I}_{k+m} + \frac{2t}{1-2t}\begin{bmatrix} \mathbb{I}_k \otimes \mathbf{1}_m^{\mathrm{T}} \\ \mathbf{1}_k^{\mathrm{T}} \otimes \mathbb{I}_m \end{bmatrix} \circ \left(\mathbb{I}_k \otimes \mathbb{I}_m - \boldsymbol{p}\cdot\mathbf{1}_k^{\mathrm{T}} \otimes \boldsymbol{q}\cdot\mathbf{1}_m^{\mathrm{T}}\right) \circ \left[\mathbb{I}_k \otimes \boldsymbol{q} \quad \boldsymbol{p} \otimes \mathbb{I}_m\right] \\ &= \mathbb{I}_{k+m} + \frac{2t}{1-2t}\begin{bmatrix} \mathbb{I}_k \otimes 1 & \boldsymbol{p} \otimes \mathbf{1}_m^{\mathrm{T}} \\ \mathbf{1}_k^{\mathrm{T}} \otimes \boldsymbol{q} & 1 \otimes \mathbb{I}_m \end{bmatrix} - \frac{2t}{1-2t}\begin{bmatrix} \boldsymbol{p}\cdot\mathbf{1}_k^{\mathrm{T}} \otimes 1 & \boldsymbol{p} \otimes \mathbf{1}_m^{\mathrm{T}} \\ \mathbf{1}_k^{\mathrm{T}} \otimes \boldsymbol{q} & 1 \otimes \boldsymbol{q}\cdot\mathbf{1}_m^{\mathrm{T}} \end{bmatrix} \\ &= \frac{1}{1-2t}\begin{bmatrix} \left(\mathbb{I}_k - 2t\boldsymbol{p}\cdot\mathbf{1}_k^{\mathrm{T}}\right) \otimes 1 & \mathbf{0}_k \otimes \mathbf{0}_m^{\mathrm{T}} \\ \mathbf{0}_m \otimes \mathbf{0}_k^{\mathrm{T}} & 1 \otimes \left(\mathbb{I}_m - 2t\boldsymbol{q}\cdot\mathbf{1}_m^{\mathrm{T}}\right) \end{bmatrix}\end{aligned} \tag{17}$$

where $\mathbf{0}$ is a zero column-vector of indicated length and 1 is a scalar. The determinant of the last matrix is

$$D_3 := \frac{(1-2t)^2}{(1-2t)^{K+M}}$$

since the determinant of each main-diagonal block is equal to $1-2t$, due to Sylvester's identity.

In the $r \to 1$ limit Formula (12), whose value is given by
$\sqrt{\dfrac{\det\mathbb{A}}{\det(\mathbb{A}-2t\mathbb{U})}} = \sqrt{\dfrac{\det\mathbb{A}}{D_1 D_2 D_3}}$, then equals to

$$\left(\frac{1}{1-2t}\right)^{(K-1)(M-1)/2} \tag{18}$$

proving the well-known result that, to this level of approximation, the distribution of U is χ^2 with $(K-1)(M-1)$ degrees of freedom.

3.3. Inverting $\mathbb{A}-2t\mathbb{U}$

Making the integrand of Formula (12) into a PDF of a Normal distribution results in

$$\frac{\exp\left(-\frac{\boldsymbol{Y}^{\mathrm{T}}\circ(\mathbb{A}-2t\mathbb{U})\circ\boldsymbol{Y}}{2}\right)}{(2\pi)^{KM/2}}\sqrt{\det\mathbb{A}-2t\mathbb{U}} \tag{19}$$

whose higher (quartic and hexic in particular) moments are key to finding $\frac{1}{n}$-proportional corrections to the MGF of each Formula (1) and Formula (3). Furthermore, these moments are polynomial functions of its *second* moments (a special property of a multivariate Normal distribution), which in turn are simply the elements of $(\mathbb{A}-2t\mathbb{U})^{-1}$.

To compute this inverse, we first write Formula (14) as a product of $(1-2t)\mathbb{P}^{-1}\otimes\mathbb{Q}^{-1}$ (whose inverse is easy) and of $\mathbb{C}=\mathbb{F}+\frac{2t}{1-2t}\mathbb{G}\circ\mathbb{H}$, introduced in Formula (15). From Formula (16) we know that

$$\mathbb{F}^{-1}=\mathbb{I}_k\otimes\mathbb{I}_m-\boldsymbol{p}\cdot\mathbf{1}_k^{\mathrm{T}}\otimes\boldsymbol{q}\cdot\mathbf{1}_m^{\mathrm{T}}$$

as r is no longer needed and can be set equal to 1.

To complete the exercise, we need to simplify

$$\mathbb{F}^{-1}-\frac{2t}{1-2t}\mathbb{F}^{-1}\circ\mathbb{G}\circ\left(\mathbb{I}_{k+m}+\frac{2t}{1-2t}\mathbb{H}\circ\mathbb{F}^{-1}\circ\mathbb{G}\right)^{-1}\circ\mathbb{H}\circ\mathbb{F}^{-1}$$

obtained from Woodbury's identity as a part of inverting $\mathbb{C}$. The last line of Formula (17) has already done exactly that for the matrix in parentheses; the corresponding inverse is

$$(1-2t)\begin{bmatrix}\mathbb{I}_k\otimes 1 & \mathbf{0}_k\otimes\mathbf{0}_m^{\mathrm{T}}\\ \mathbf{0}_m\otimes\mathbf{0}_k^{\mathrm{T}} & 1\otimes\mathbb{I}_m\end{bmatrix}+2t\begin{bmatrix}\boldsymbol{p}\cdot\mathbf{1}_k^{\mathrm{T}}\otimes 1 & \mathbf{0}_k\otimes\mathbf{0}_m^{\mathrm{T}}\\ \mathbf{0}_m\otimes\mathbf{0}_k^{\mathrm{T}} & 1\otimes\boldsymbol{q}\cdot\mathbf{1}_m^{\mathrm{T}}\end{bmatrix}$$

easily verifiable by simple matrix multiplication (note that both matrices in the last expression are idempotent).

This inverse still needs to be pre-multiplied by $[\mathbb{I}_k\otimes\boldsymbol{q}\quad\boldsymbol{p}\otimes\mathbb{I}_m]$ and post-multiplied by $\begin{bmatrix}\mathbb{I}_k\otimes\mathbf{1}_m^{\mathrm{T}}\\ \mathbf{1}_k^{\mathrm{T}}\otimes\mathbb{I}_m\end{bmatrix}$, thus getting

$$(1-2t)\left(\mathbb{I}_k\otimes\boldsymbol{q}\cdot\mathbf{1}_m^{\mathrm{T}}+\boldsymbol{p}\cdot\mathbf{1}_k^{\mathrm{T}}\otimes\mathbb{I}_m\right)+4t\boldsymbol{p}\cdot\mathbf{1}_k^{\mathrm{T}}\otimes\boldsymbol{q}\cdot\mathbf{1}_m^{\mathrm{T}}$$

and further pre and post-multiplied by $\mathbb{F}^{-1}$, resulting in

$$(1-2t)\left(\mathbb{I}_k\otimes\boldsymbol{q}\cdot\mathbf{1}_m^{\mathrm{T}}+\boldsymbol{p}\cdot\mathbf{1}_k^{\mathrm{T}}\otimes\mathbb{I}_m-2\boldsymbol{p}\cdot\mathbf{1}_k^{\mathrm{T}}\otimes\boldsymbol{q}\cdot\mathbf{1}_m^{\mathrm{T}}\right)+4t\mathbf{0}_k\cdot\mathbf{0}_k^{\mathrm{T}}\otimes\mathbf{0}_m\cdot\mathbf{0}_m^{\mathrm{T}}$$

Multiplying the last expression by $\frac{2t}{1-2t}$ and subtracting the result from $\mathbb{F}^{-1}$ yields $\mathbb{C}^{-1}$, namely

$$\mathbb{I}_k\otimes\mathbb{I}_m+(4t-1)\boldsymbol{p}\cdot\mathbf{1}_k^{\mathrm{T}}\otimes\boldsymbol{q}\cdot\mathbf{1}_m^{\mathrm{T}}-2t\left(\mathbb{I}_k\otimes\boldsymbol{q}\cdot\mathbf{1}_m^{\mathrm{T}}+\boldsymbol{p}\cdot\mathbf{1}_k^{\mathrm{T}}\otimes\mathbb{I}_m\right)$$

Finally, post-multiplying by $\frac{\mathbb{P}\otimes\mathbb{Q}}{1-2t}$ results in the following expression for $\left(\mathbb{A}-2t\mathbb{U}\right)^{-1}$

$$\frac{\mathbb{P}\otimes\mathbb{Q}}{1-2t}+\left(\frac{1}{1-2t}-2\right)\boldsymbol{p}\cdot\boldsymbol{p}^{\mathrm{T}}\otimes\boldsymbol{q}\cdot\boldsymbol{q}^{\mathrm{T}}+\left(1-\frac{1}{1-2t}\right)\left(\mathbb{P}\otimes\boldsymbol{q}\cdot\boldsymbol{q}^{\mathrm{T}}+\boldsymbol{p}\cdot\boldsymbol{p}^{\mathrm{T}}\otimes\mathbb{Q}\right)$$

and the corresponding formula for computing the desired second moments, namely

$$\begin{aligned}\mathbb{E}\left(Y_{i,j}Y_{k,m}\right)&=-2p_ip_kq_jq_m+\delta_{i,k}p_iq_jq_m+\delta_{j,m}p_ip_kq_j\\&+\frac{p_i\left(\delta_{i,k}-p_k\right)q_j\left(\delta_{j,m}-q_m\right)}{1-2t}\end{aligned}\tag{20}$$

(note the expected agreement with Formula (6) when $t=0$).

Consequently,

$$\begin{aligned}\mathbb{E}\left(Y_{i,j}Y_{k,\bullet}\right)&=p_iq_j\left(\delta_{i,k}-p_k\right)\\\mathbb{E}\left(Y_{i,j}Y_{\bullet,m}\right)&=p_iq_j\left(\delta_{j,m}-q_m\right)\\\mathbb{E}\left(Y_{i,\bullet}Y_{k,\bullet}\right)&=p_i\left(\delta_{i,k}-p_k\right)\\\mathbb{E}\left(Y_{\bullet,j}Y_{\bullet,m}\right)&=q_j\left(\delta_{j,m}-q_m\right)\\\mathbb{E}\left(Y_{i,\bullet}Y_{\bullet,j}\right)&=0\end{aligned}\tag{21}$$

provide the remaining moments needed to proceed with computing the desired $\frac{1}{n}$-proportional corrections. Note that fourth-order moments are then found from

$$\mathbb{E}\left(abcd\right)=\mathbb{E}\left(ab\right)\mathbb{E}\left(cd\right)+\mathbb{E}\left(ac\right)\mathbb{E}\left(bd\right)+\mathbb{E}\left(ad\right)\mathbb{E}\left(bc\right)$$

(with subsequent special cases, such as $\mathbb{E}\left(a^2b^2\right)=2\mathbb{E}\left(ab\right)^2+\mathbb{E}\left(a^2\right)\mathbb{E}\left(b^2\right)$ etc.), and the sixth-order moments from

$$\mathbb{E}\left(abcdef\right)=\mathbb{E}\left(ab\right)\mathbb{E}\left(cd\right)\mathbb{E}\left(ef\right)+\cdots+\mathbb{E}\left(af\right)\mathbb{E}\left(cd\right)\mathbb{E}\left(be\right)$$

(where the RHS consists of 15 terms corresponding to all possible ways of pairing the six arguments), usually needed in its special form such as

$$\mathbb{E}\left(a^3b^3\right)=9\mathbb{E}\left(a^2\right)\mathbb{E}\left(b^2\right)\mathbb{E}\left(ab\right)+6\mathbb{E}\left(ab\right)^3$$

etc.

4. $\frac{1}{n}$-Proportional Corrections

To find the desired correction to the χ^2 distribution of the Pearson test, we must include $\frac{1}{\sqrt{n}}$ and $\frac{1}{n}$-proportional terms in the Formula (13) expansion, thus getting (listing the *extra* terms only)

$$\mathfrak{U}=\cdots+\frac{\sum_{i=1}^{K}\frac{Y_{i,\bullet}^3}{p_i^2}+\sum_{j=1}^{M}\frac{Y_{\bullet,j}^3}{q_j^2}}{\sqrt{n}}+\frac{\sum_{i=1}^{K}\sum_{j=1}^{M}\left(2\frac{Y_{i,j}Y_{i,\bullet}Y_{\bullet,j}}{p_iq_j}-\frac{Y_{i,j}^2Y_{i,\bullet}}{p_i^2q_j}-\frac{Y_{i,j}^2Y_{\bullet,j}}{p_iq_j^2}\right)}{\sqrt{n}}$$
$$+\frac{\sum_{i=1}^{K}\sum_{j=1}^{M}\left(-2\frac{Y_{i,j}Y_{i,\bullet}^2Y_{\bullet,j}}{p_i^2q_j}-2\frac{Y_{i,j}Y_{i,\bullet}Y_{\bullet,j}^2}{p_iq_j^2}+\frac{Y_{i,j}^2Y_{i,\bullet}^2}{p_i^3q_j}+\frac{Y_{i,j}^2Y_{\bullet,j}^2}{p_iq_j^3}+\frac{Y_{i,j}^2Y_{i,\bullet}Y_{\bullet,j}}{p_i^2q_j^2}+\frac{Y_{i,\bullet}^2Y_{\bullet,j}^2}{p_iq_j}\right)}{n}$$
$$+\frac{-\sum_{i=1}^{K}\frac{Y_{i,\bullet}^4}{p_i^3}-\sum_{j=1}^{M}\frac{Y_{\bullet,j}^4}{q_j^3}}{n}$$
$$:=\frac{u_1}{\sqrt{n}}+\frac{u_2}{\sqrt{n}}+\frac{u_3}{n}+\frac{u_4}{n}$$

Similarly, we need the extra terms in the expansion of the natural logarithm of the RHS of Formula (2); this time we get

$$\mathcal{A}=\cdots+\frac{\sum_{i=1}^{K}\sum_{j=1}^{M}\left(\frac{Y_{i,j}^3}{6p_i^2q_j^2}-\frac{Y_{i,j}}{2p_iq_j}\right)}{\sqrt{n}}+\frac{1}{12n}$$
$$-\frac{\sum_{i=1}^{K}\sum_{j=1}^{M}\left(\frac{Y_{i,j}^4}{12p_i^3q_j^3}-\frac{Y_{i,j}^2}{4p_i^2q_j^2}+\frac{1}{12p_iq_j}\right)}{n}$$
$$:=\frac{a_1}{\sqrt{n}}+\frac{1}{12n}+\frac{a_2}{n}$$

Combining these two, we expand $\exp\left(\mathcal{A}-2t\mathfrak{U}\right)$ up to and including $\frac{1}{n}$-proportional terms, getting

$$\frac{2\left(u_1^2+u_2^2+2u_1u_2\right)t^2-2t\left(u_3+u_4\right)-2t\left(u_1+u_2\right)a_1+\frac{1}{2}a_1^2+a_2+\frac{1}{12}}{n} \tag{22}$$

(having discarded $\frac{1}{\sqrt{n}}$-proportional terms, whose expected values are equal to zero).

What remains to be done is to take the expected value of Formula (22); this is done by first expanding every sum in this expression, then finding the expected value of each term of these expansions, finally followed by carrying out the summation itself (over up to four indices). This results in scores of individual contributions, even though, rather surprisingly, most of them cancelling each other upon subsequent simplification. This computation can be successfully completed only with a computer's help; the actual Mathematica program required many lines of code and is available upon request.

As an example, we demonstrate how to find $\mathbb{E}\left(u_1u_2\right)$. We start by ignoring the triple summations, multiplying only the *first* term of u_1 by u_2 (note that this requires using a different index for u_1) and expanding the resulting expected value, thus

$$\mathbb{E}\left(\frac{Y_{k,\bullet}^3}{p_k^2}\left(2\frac{Y_{i,j}Y_{i,\bullet}Y_{\bullet,j}}{p_iq_j}-\frac{Y_{i,j}^2Y_{i,\bullet}}{p_i^2q_j}-\frac{Y_{i,j}^2Y_{\bullet,j}}{p_iq_j^2}\right)\right)$$

$$=\frac{6\mathbb{E}\left(Y_{i,\bullet}Y_{k,\bullet}\right)\mathbb{E}\left(Y_{i,j}Y_{k,\bullet}\right)^2-3\mathbb{E}\left(Y_{i,\bullet}Y_{k,\bullet}\right)\mathbb{E}\left(Y_{k,\bullet}^2\right)\mathbb{E}\left(Y_{i,j}^2\right)-6\mathbb{E}\left(Y_{i,j}Y_{i,\bullet}\right)\mathbb{E}\left(Y_{i,j}Y_{k,?}\right)\mathbb{E}\left(Y_{k,\bullet}^2\right)}{p_i^2q_jp_k^2}$$

$$-\frac{6\mathbb{E}\left(Y_{i,j}Y_{\bullet,j}\right)\mathbb{E}\left(Y_{i,j}Y_{k,\bullet}\right)\mathbb{E}\left(Y_{k,\bullet}^2\right)}{p_iq_j^2p_k^2}+\frac{6\mathbb{E}\left(Y_{i,j}Y_{\bullet,j}\right)\mathbb{E}\left(Y_{i,\bullet}Y_{k,\bullet}\right)\mathbb{E}\left(Y_{k,\bullet}^2\right)}{p_iq_jp_k^2}$$

where we have already discarded terms containing a zero factor such as $\mathbb{E}\left(Y_{i,\bullet}Y_{\bullet,j}\right)$. We then evaluate the expected values based on Formula (20) and Formula (21), and complete the i,j and k summations, resulting in

$$-12+18K+9K^2-15P_0+\frac{3(M-1)\left(K^2-P_0\right)}{1-2t}$$

where $P_0:=\sum_{i=1}^K p_i^{-1}$. To include the second term of u_1, we simply interchange K and M of the previous answer, replace P_0 by $Q_0:=\sum_{j=1}^M q_j^{-1}$, and add the result to the previous answer.

When this is completed for all terms of Formula (22), adding them and simplifying (many terms cancel out, as mentioned already) results in

$$\begin{aligned}&\frac{d}{2n}\left(\frac{1}{1-2t}-1\right)\\&+\frac{\tilde{P}\tilde{Q}-2(M-1)\tilde{P}-2(K-1)\tilde{Q}-2(K+M-3)d}{8n}\left(\frac{1}{(1-2t)^2}-\frac{2}{1-2t}+1\right)\\&+\frac{5\tilde{P}\tilde{Q}+2(M-1)(M-2)\tilde{P}+2(K-1)(K-2)\tilde{Q}+2d(K-2)(M-2)}{24n}\\&\times\left(\frac{1}{(1-2t)^3}-\frac{3}{(1-2t)^2}+\frac{3}{1-2t}-1\right)\end{aligned}\tag{23}$$

where $\tilde{P}:=P_0-K^2,\tilde{Q}:=Q_0-M^2$ and $d:=(K-1)(M-1)$; note that when $p_i=\frac{1}{K}$ for all i values, $\tilde{P}$ is then equal to zero, and similarly $q_j=\frac{1}{M}$ implies $\tilde{Q}=0$.

The last expression, further divided by $(1-2t)^d$, is the actual correction to the MGF of the test statistic Formula (13), implying that the corresponding PDF correction is a linear combination of χ^2 distributions; to get an explicit formula for this correction, make the following replacement in Formula (23)

$$\left(\frac{1}{1-2t}-1\right)\to\left(\frac{x}{d}-1\right)\chi_d^2(x)$$

$$\left(\frac{1}{(1-2t)^2}-\frac{2}{1-2t}+1\right)\to\left(\frac{x^2}{d(d+2)}-\frac{2x}{d}+1\right)\chi_d^2(x)$$

$$\left(\frac{1}{(1-2t)^3}-\frac{3}{(1-2t)^2}+\frac{3}{1-2t}-1\right)\to\left(\frac{x^3}{d(d+2)(d+4)}-\frac{3x^2}{d(d+2)}+\frac{3x}{d}-1\right)\chi_d^2(x)$$

where

$$\chi^2_d(x)=\frac{x^{d/2-1}\exp(-x/2)}{2^{d/2}\Gamma(d/2)}$$

is the PDF of the χ^2_d distribution. Note that each of the three corrections integrates to zero (the main reason for keeping them in this form).

These results are in perfect agreement with those obtained, using a different technique, by [8].

G-Test Results

This time we get the following extra terms in the expansion of Formula (3)

$$\mathfrak{U}=\cdots+\frac{2\sum_{i=1}^{K}\frac{Y_{i,\bullet}^3}{p_i^2}+2\sum_{i=1}^{M}\frac{Y_{\bullet,j}^3}{q_j^2}}{3\sqrt{n}}+\frac{-2\sum_{i=1}^{K}\sum_{j=1}^{M}\frac{Y_{i,j}^3}{p_i^2q_j^2}}{3\sqrt{n}}$$
$$+\frac{\sum_{i=1}^{K}\frac{Y_{i,\bullet}^4}{p_i^3}+\sum_{i=1}^{M}\frac{Y_{\bullet,j}^4}{q_j^3}}{3n}+\frac{-\sum_{i=1}^{K}\sum_{j=1}^{M}\frac{Y_{i,j}^4}{p_i^3q_j^3}}{3n}$$
$$:=\frac{u_1}{\sqrt{n}}+\frac{u_2}{\sqrt{n}}+\frac{u_3}{n}+\frac{u_4}{n}$$

We then repeat the steps of the previous section, getting the following (surprisingly simple) correction to the corresponding PDF

$$\frac{\tilde{P}\tilde{Q}+\left(M^2-1\right)\tilde{P}+\left(K^2-1\right)\tilde{Q}+\left(K^2-1\right)\left(M^2-1\right)}{12n}\left(\frac{x}{d}-1\right)\chi^2_d(x)$$

5. Monte-Carlo Verification and Conclusion

To demonstrate the utility of our correction formulas (but also their limitations, in case of the G-test), we generate a million observations from a contingency-table model and compare the results with the χ^2 approximation, both without and including the $\frac{1}{n}$-proportional correction.

Our first example assumes that two attributes have 10 *equally likely* levels each (for a total of 100 cells), and the number of observations is equal to 150. The results are displayed in **Figure 1** for the Pearson test and in **Figure 2** for the G-test (the dotted lines show the basic χ^2 approximation, the solid lines include the respective corrections). From these graphs it is quite obvious that, in the former case, the correction terms are quite capable of removing the (still rather serious) error, while the G-test is so hopelessly inaccurate that repairing it becomes impossible (the coefficient of our formula has a value bigger than 5, being no longer 'small', resulting in making the 'corrected' PDF partly negative; a strong indication that the error is too large to be dealt with in this manner). The benefit of our correction is to clearly indicate a presence of an inacceptable error in the usual χ^2 approximation (suggesting a solution would require a separate study). We conclude that the G-test should not be used with large number of cells (unless the number of

observations is correspondingly large).

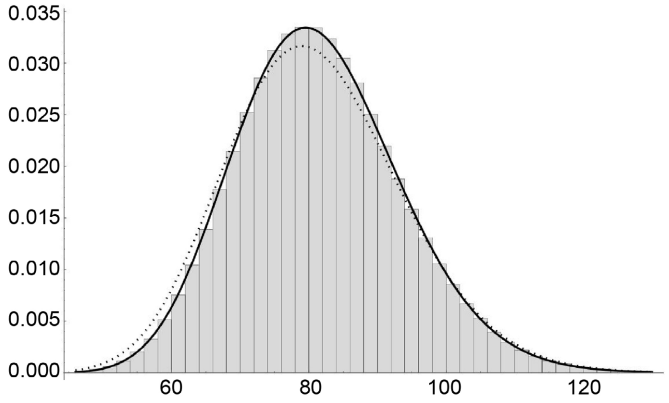

Figure 1. Pearson's test correction (n = 150).

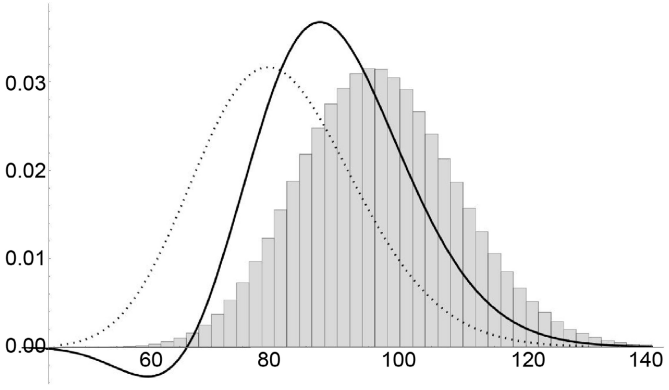

Figure 2. G-test and its error (n = 150).

Nevertheless, we can still confirm that our G-test correction works when the attributes have only 3 levels each (we will still make them equally likely) and n is equal to 35; this is indicated in **Figure 3**.

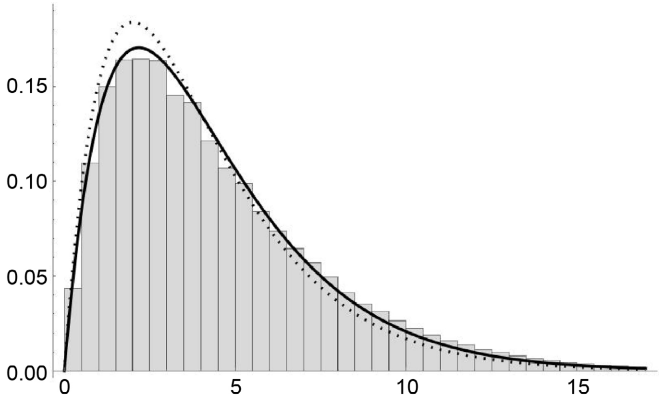

Figure 3. G-test corrected (n = 35).

Note that the empirical distribution (represented by a histogram) has some mi-

nor irregularities due to the exact distribution's discrete nature (correcting for these would require a new study). Nevertheless, the graph indicates that the right-tail region is not affected by this phenomenon (this is particularly true for the Pearson test, as seen in Figure 4). We should also mention that, in this same situation and using the same data, the Pearson test is still more accurate than the G-test, both when using the basic χ^2 approximation and using its corrected version. In conclusion, we find no reason for recommending the G-test over the Pearson test of independence.

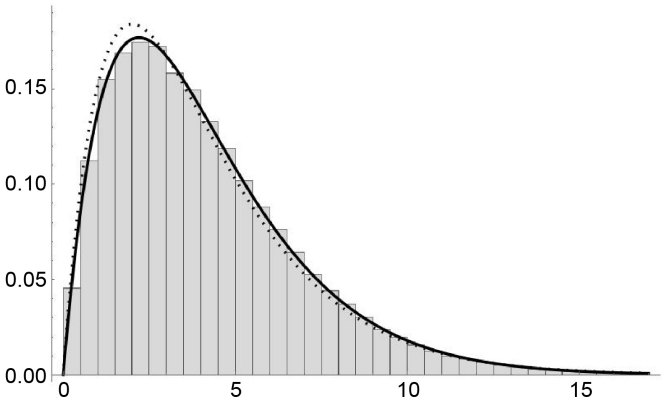

Figure 4. Correcting Pearson's test (n = 35).

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Kendall, M.G. and Stuart, A. (1973) The Advanced Theory of Statistics, Vol. 2, Chapter 33. Harper Publishing Company.

[2] MacDonald, H.J. (2014) G-Test of Goodness-of-Fit, Handbook of Biological Statistics. 3rd Edition, Sparky House Publishing, 53-58.

[3] Sokal, R.R. and Rohlf, F.J. (1981) Biometry: The Principles and Practice of Statistics in Biological Research. 2nd Edition, Freeman.

[4] Wikipedia. Kronecker Product.

[5] Wikipedia. Woodbury Matrix Identity.

[6] Wikipedia. Sylvester's Determinant Identity.

[7] Serfling, R.J. (1980). Approximation Theorems of Mathematical Statistics. Wiley. https://doi.org/10.1002/9780470316481

[8] Vrbik, J. (2014) Improving Accuracy of χ^2 Test of Independence. *Advances and Applications in Statistics*, **39**, 91-94.