Seven proofs of the Pearson Chi-squared independence test and its graphical interpretation

Eric Benhamou [*, ‡] , Valentin Melot [†, ‡]

Abstract

This paper revisits the Pearson Chi-squared independence test. After presenting the underlying theory with modern notations and showing new way of deriving the proof, we describe an innovative and intuitive graphical presentation of this test. This enables not only interpreting visually the test but also measuring how close or far we are from accepting or rejecting the null hypothesis of non independence.

AMS 1991 subject classification: 62E10, 62E15

Keywords: causality, Pearson correlation coefficient, graphical heat map

[*]A.I. SQUARE CONNECT, 35 Boulevard d'Inkermann 92200 Neuilly sur Seine, France and LAMSADE, Université Paris Dauphine, Place du Maréchal de Lattre de Tassigny,75016 Paris, France. E-mail: eric.benhamou@aisquareconnect.com, eric.benhamou@dauphine.eu

[†]Ecole Normale Supérieure, 45 Rue d'Ulm, 75005 Paris, France. E-mail: valentin.melot@ens.fr

[‡]the authors would like to mention a fruitful discussion with Robin Ryder that originates this work

1. Introduction

A natural and common question in statistics is to state if two nominal (categorical) variables are independent or not. The traditional approach is to use the Pearson Chi-Square test of independence as developed in Pearson (1900). We can determine if there is a (statistically) significant relationship between these two nominal variables. Traditionally, the data are displayed in a contingency table where each row represents a category for one variable and each column represents a category for the other variable.

For example, say a researcher wants to examine the relationship between gender (male vs. female) and driver riskiness (dangerous or safe driver). The null (respectively the alternative) hypothesis for this test is that there is no relationship (respectively a relationship) between the two variables: gender and driver riskiness.

The chi-square test yields only an approximated p-value as this is an asymptotic test as we will see shortly. Hence, this only works when data-sets are large enough. For small sample sizes, Fisher's (as explained in Fisher (1922)) or Barnard's (as presented in Bernard (1945) or in Bernard (1947)) exact tests are more appropriate but more complex. The large interest of the Pearson Chi-Square test of independence is its simplicity and robustness as it only relies on two main assumptions: large sample size and independence of observations. It is a mainstream test, available in the core library of **R**: function *chisq.test* or in python (function *stats.pearsonr* of the scipy library). A natural question is then to graphically represent this test to illustrate if we are close or not to the null hypothesis. Although there has been lot of packages to represent contingency tables, there has been a lack of thought for representing this test visually. This has motivated us to write this work that leads to the writing of a short function in python to do it. We also revisited the proof done by Pearson in 1900 and show that this proof can be derived more elegantly with more recent mathematical tools, namely Cochran theorem that was only found in 1934 (see Cochran (1934)) and also the Sherman Morison matrix inversion formula (provided in 1949 by Sherman and Morrison (1949)).

The contribution of our paper are twofold: to provide first a modern proof of the Pearson chi square test and second a nice and graphical interpretation of this test. Compared to previous proofs as for instance in Buonocore and Pirozzi (2014), we are the first one to provide seven proofs for this seminal results with the use of a wide range of tools, like not only Cochran theorem but Sherman Morison formula, Sylvester's theorem and an elementary proof using De Moivre-Laplace theorem. We are also the first paper to suggest to use confidence interval in mosaic plots.

The paper is organized as follows. We first present with modern notation the underlying theory and the seven proofs. We then examine how to display on a single graphic both the

contingency tables and the test. We show that this is surprisingly simple yet powerful. We conclude with various examples to illustrate our graphical representation.

2. Null hypothesis asymptotic

Let $X_1, X_2, \ldots$ be independent samples from a multinomial(1, p) distribution, where p is a k-vector with nonnegative entries that sum to one. That is,

$$P(X_{ij} = 1) = 1 - P(X_{ij} = 0) = p_j \quad \text{for all } 1 \leq j \leq k \tag{1}$$

and each X_i consists of exactly $k-1$ zeros and a single one, where the one is in the component of the "success" category at trial i.

This equation implies in particular that $\mathrm{Var}(X_{ij}) = p_j(1-p_j)$. Furthermore, $\mathrm{Cov}(X_{ij}, X_{il}) = \mathbb{E}[X_{ij}X_{il}] - p_j p_l = -p_j p_l$ for $j \neq l$. Therefore, the random vector X_i has covariance matrix given by

$$\Sigma = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \ldots & -p_1p_k \\ -p_1p_2 & p_2(1-p_2) & \ldots & -p_2p_k \\ \vdots & \vdots & \ddots & \vdots \\ -p_1p_k & -p_2p_k & \ldots & p_k(1-p_k) \end{pmatrix} \tag{2}$$

Let us prove shortly that the asymptotic distribution of the Pearson chi-square statistic given by

$$\chi^2 = \sum_{j=1}^{k} \frac{(N_j - np_j)^2}{np_j} \tag{3}$$

where N_j is the random variable $n\bar{X}_j$, the number of successes in the jth category for trials $1, \ldots, n$ converges in distribution to the chi-square distribution with $k-1$ degrees of freedom. This will imply in particular that to test that two samples are from the same statistics, we can use a test of goodness of fit.

The proof is rather elementary and we provide below seven different methods. These proofs show that they are profound connections between binomial, multinomial, Poisson, normal and chi squared distribution for asymptotic cases. They also illustrate that this problem can be tackled with multiple mathematical tools like De Moivre-Laplace theorem that is an early and simpler version of the Central Limit theorem and a recursive induction, but also characteristic function and Lévy's continuity theorem, geometry and linear algebra reasoning that are at the foundation of the Cochran theorem.

3. Seven different proofs for the Pearson independence test

3.1. First Proof: Sherman Morison formula for direct computation

Since $\mathbb{E}[X_i] = p$, the central limit theorem implies

$$\sqrt{n}(\bar{X}_n - p) \xrightarrow[n\to\infty]{\mathrm{d}} N_k(0, \Sigma) \tag{4}$$

where the notation $\xrightarrow[n\to\infty]{\mathrm{d}}$ indicates convergence in distribution and $N_k(0,\Sigma)$ is the multi dimensional normal with k dimension and Σ as its covariance matrix. Note that Σ is not invertible as the sum of any jth column of Σ is null since it is equal to $p_j - p_j(p_1 + \ldots + p_k)$.

A first way to tackle the problem of inferring the distribution of the χ^2 statistic is to remove one dimension to the X vector to have a covariance matrix of full rank. More precisely, let us define for each sample i, the truncated vector variable $X_i^* = (X_{i1}, \ldots, X_{i,k-1})^T$. It is the $k-1$ vector consisting of the first $k-1$ components of X_i. Its covariance matrix is the sub matrix of Σ reduced to its first $k-1$ rows and columns. We call it Σ^*. Σ^* writes as the sum of two simple matrices

$$\Sigma^* = \underbrace{\begin{pmatrix} p_1 & 0 & \ldots & 0 \\ 0 & p_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & p_{k-1} \end{pmatrix}}_{A} - \underbrace{\begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_{k-1} \end{pmatrix}}_{b} \underbrace{\begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_{k-1} \end{pmatrix}^T}_{b^T} \tag{5}$$

We have trivially that $1 - b^T A^{-1} b = \sum_{i=1}^{k} p_i - \sum_{i=1}^{k-1} p_i = p_k$ and $A^{-1}b = (1, \ldots, 1)^T$ where the last vector is of $k-1$ dimension. Σ^* inverse, denoted by $(\Sigma^*)^{-1}$, is therefore given by the Sherman-Morrison formula

$$(\Sigma^*)^{-1} = \underbrace{\begin{pmatrix} 1/p_1 & 0 & \ldots & 0 \\ 0 & 1/p_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1/p_{k-1} \end{pmatrix}}_{A^{-1}} + \underbrace{\frac{1}{p_k} \begin{pmatrix} 1 & 1 & \ldots & 1 \\ 1 & 1 & \ldots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \ldots & 1 \end{pmatrix}}_{(1/(1-b^T A^{-1} b)(A^{-1} b b^T A^{-1})} \tag{6}$$

Let us write also p^* the $k-1$ vector of probability $p^* = (p_1, \ldots, p_{k-1})^T$. The χ^2 statistic

of equation (3) can be reformulated as follows:

$$\chi^2 = n \sum_{j=1}^{k} \frac{(\bar{X}_j - p_j)^2}{p_j} \tag{7}$$

$$= n \sum_{j=1}^{k-1} \frac{(\bar{X}_j - p_j)^2}{p_j} + \frac{(\bar{X}_k - p_k)^2}{p_k} \tag{8}$$

$$= n \sum_{j=1}^{k-1} \frac{(\bar{X}_j - p_j)^2}{p_j} + \frac{(\sum_{j=1}^{k-1}(\bar{X}_j - p_j))^2}{p_k} \tag{9}$$

where we have used in the last equation that $\sum_{j=1}^{k} \bar{X}_j - p_j = 0$

The latter equation can be rewritten in terms of matrix notation as

$$\chi^2 = n(\bar{X}^* - p^*)^T (\Sigma^*)^{-1} (\bar{X}^* - p^*) \tag{10}$$

The Central limit theorem states that $Y_n = \sqrt{n}(\Sigma^*)^{-1/2}(\bar{X}^* - p^*)$ converges in distribution to a normal variable $N_{k-1}(0, I_{k-1})$. The χ^2 statistic given by $(Y_n)^T Y_n$ converges in distribution to $\chi^2_\infty = N_{k-1}(0, I_{k-1})^T N_{k-1}(0, I_{k-1})$. χ^2_∞ is the sum of the squares of $k-1$ independent standard normal random variables, which is a chi square distribution with $k-1$ degree of freedom. This concludes the first proof. □

3.2. Second Proof: Cochran theorem

The second proof relies on the Cochran theorem. The start is the same. Since $\mathbb{E}[X_i] = p$, the central limit theorem implies

$$\sqrt{n}(\bar{X}_n - p) \xrightarrow[n\to\infty]{\mathrm{d}} N_k(0, \Sigma) \tag{11}$$

where N_k denotes the k multi dimensional normal as always in this paper. Let us denote by Z the standard reduced Gaussian variable corresponding to the basis implied by the k multi dimensional normal $N_k(0, I_k)$.

If we apply the Cochran theorem with the projection on the sub vectorial space F spanned by $\sqrt{p} = (\sqrt{p_1}, ..., \sqrt{p_k})^T$ (whose norm is obviously 1), we have that the projection matrix on F is given by

$$P_F = \sqrt{p}(\sqrt{p}^T \sqrt{p})^{-1} \sqrt{p}^T = \sqrt{p}\sqrt{p}^T$$

and that the projection on the orthogonal of F denoted by $F^\perp$ is given by $P_{F^\perp} = I - P_F = I - \sqrt{p}\sqrt{p}^T$.

Since F is spanned by one single vector, its dimension is 1, while its orthogonal, $F^{\perp}$, is of dimension $k-1$. The Cochran theorem states that the projection on $F^{\perp}$ of Z follows a normal whose distribution is given by $N(0, P_{F^{\perp}})$, and that the squared norm of the projection on $F^{\perp}$ follows a chi square distribution of dimension $k-1$.

We can notice that if we define $\Gamma = \text{diag}(p)$ and $A_n = \sqrt{n}\Gamma^{-1/2}(\bar{X} - p)$. The Chi-squared statistics can be rewritten as the norm of the stochastic vector A^n since

$$\chi^2 = (A_n)^T A_n \tag{12}$$

Using equation (11), we also know that A_n converges in distribution to $N(0, \Gamma^{-1/2}\Sigma\Gamma^{-1/2})$. Since $\Sigma = \Gamma - pp^T$, we have that

$$\Gamma^{-1/2}\Sigma\Gamma^{-1/2} = I_k - \Gamma^{-1/2}(pp^T)\Gamma^{-1/2} = I_k - (\Gamma^{-1/2}p)(\Gamma^{-1/2}p)^T \tag{13}$$

$$= I_k - \sqrt{p}\sqrt{p}^T = P_{F^{\perp}} \tag{14}$$

This states that A_n converges in distribution to the projection of Z on $F^{\perp}$. The statistics, χ^2, that is the squared norm of A_n converges in distribution to the squared norm of the projection of Z on $F^{\perp}$, whose distribution is a chi square distribution of dimension $k-1$. This proves that the statistics, χ^2, converges in distribution to a chi square distribution of dimension $k-1$. □

3.3. Third Proof: Sylvester theorem and eigen values

The third proof relies on Sylvester theorem and is based on an explicit computation of the eigen values of the associated covariance matrix. If we write Z the k vector with coordinates given by $Z_i = \frac{N_i - np_i}{\sqrt{np_i}} = \sqrt{n}\frac{N_i/n - p_i}{\sqrt{p_i}}$, the central limit theorem states that vector Z converges in distribtuion to $N(0, \Omega)$, a multivariate normal distribution, whose covariance matrix is given by

$$\Omega = \text{Cov}(Z) = \begin{pmatrix} 1-p_1 & -\sqrt{p_1 p_2} & \cdots \\ -\sqrt{p_1 p_2} & 1-p_2 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}.$$

We can compute explicitly the characteristic polynomial of this matrix. The Sylvester theorem states that $\text{Det}(I_k - cr) = 1 - r^T c$ for any $c, r \in \mathbb{R}^k$ (k dimensional vectors). Therefore, a direct application of Sylvester theorem shows that $\text{Det}(\Omega - \lambda I) = (1-\lambda)^{k-1}\lambda$ as $\Omega = I - pp^T$ for $p = (\sqrt{p_1}, \sqrt{p_2}, \dots)$ and $\text{Det}(\Omega - \lambda I) = (1-\lambda)^k \text{Det}(I_n - \frac{1}{1-\lambda}pp^T)$. This implies that Ω has $k-1$ eigenvalues that are 1 and one that is 0 and that the distribution is

really $k-1$ dimensional embedded in k dimensions. In particular there is a rotation matrix A that makes

$$A\Omega A^T = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} I_{k-1} & 0_{k-1,1} \\ 0_{1,k-1} & 0_{1,1} \end{pmatrix}$$

where $0_{n,m}$ is the n rows, m columns matrix filled with 0.

Denote $W = AZ \sim N_k(0, A\Omega A^T)$. Then W is a vector $(W_1, W_2, \dots, W_{k-1}, 0)$ of iid. $\mathcal{N}(0,1)$ Gaussians with only $k-1$ non null coordinates (the first $k-1$ coordinates). The function $f(Z) = Z_1^2 + Z_2^2 + \dots$ is the norm $\|Z\|_2^2$, and hence it is invariant if we rotate its argument. This means $f(Z) = f(AZ) = f(W) = W_1^2 + W_2^2 + \dots + W_{k-1}^2$ is Chi-square distributed with $k-1$ degree of freedom □

3.4. Fourth Proof: Characteristic function

Let us use characteristic function. The χ^2 statistic' s characteristic function is:

$$\phi_\chi^2(t) = \mathbb{E}\left[e^{\mathrm{i}t\chi^2}\right] = \mathbb{E}\left[e^{\mathrm{i}t\sum_{j=1}^k \frac{(N_j - np_j)^2}{np_j}}\right] \tag{15}$$

Since $\sum_{j=1\dots k} N_j - np_j = 0$, we can reduce the sum to $k-1$ terms and show the real quadratic form as follows:

$$\phi_\chi^2(t) = \mathbb{E}\left[e^{\mathrm{i}t\left(\sum_{j=1}^{k-1} \frac{(N_j - np_j)^2}{np_j} + \frac{(\sum_{j=1}^{k-1} N_j - np_j)^2}{np_k}\right)}\right] \tag{16}$$

$$= \mathbb{E}\left[e^{\frac{\mathrm{i}t}{n}\left((N^* - np^*)^T(\Sigma^*)^{-1}(N^* - np^*)\right)}\right] \tag{17}$$

where $N^* = (N_1, \dots, N_{k-1})^T$ is a $k-1$ stochastic vector and $(\Sigma^*)^{-1}$ the $k-1$ squared symmetric matrix is given by

$$(\Sigma^*)^{-1} = \begin{pmatrix} 1/p_1 + 1/p_k & 1/p_k & \dots & 1/p_k \\ 1/p_k & 1/p_2 + 1/p_k & \dots & 1/p_k \\ \vdots & \vdots & \ddots & \vdots \\ 1/p_k & 1/p_k & \dots & 1/p_{k-1} + 1/p_k \end{pmatrix} \tag{18}$$

The variance matrix of N^* is easily computed and given by :

$$\Sigma^* = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \dots & -p_1p_{k-1} \\ -p_1p_2 & p_2(1-p_2) & \dots & -p_2p_{k-1} \\ \vdots & \vdots & \ddots & \vdots \\ -p_1p_{k-1} & -p_2p_{k-1} & \dots & p_{k-1}(1-p_{k-1}) \end{pmatrix} \tag{19}$$

It is remarkable that looking at the characteristic function provides the inverse of Σ^* without effort as one can validate that $(\Sigma^*)^{-1}\Sigma^* = I_k$. The Central limit theorem for Binomial (which is also referred to as De Moivre-Laplace's theorem) states that $N^* - np^* \xrightarrow[n\to\infty]{\text{d}} N(0, \text{Var}(N^*)) = N(0, \Sigma^*)$. Hence, $(\Sigma^*)^{-1/2}(N^* - np^*) \xrightarrow[n\to\infty]{\text{d}} N(0, I_{k-1})$. Let us denote by $Z = (Z_1, \dots, Z_{k-1})^T$ the corresponding $k-1$ dimensional standard normal distribution. Taking the limit in the characteristic function thanks to Lévy's continuity theorem, we have therefore that

$$\phi^2_\chi(t) \xrightarrow[n\to\infty]{\text{d}} \mathbb{E}\left[\prod_{j=1,\dots,k-1} e^{\text{i}tZ_j^2}\right] = \left(\mathbb{E}\left[e^{\text{i}tU^2}\right]\right)^{k-1} \tag{20}$$

where U is a standard normal $N(0,1)$. We have $\mathbb{E}\left[e^{\text{i}tU^2}\right] = \frac{1}{1-2it}$ which leads to $\phi_X(\chi^2) \underset{n\to\infty}{\to} \frac{1}{(1-2it)^{k-1}}$ which concludes the proof as the characteristic function of a $\chi^2(k-1)$ is precisely $\frac{1}{(1-2it)^{k-1}}$ □

3.5. Fifth Proof: Projection matrix and Pythagoras theorem

Even if this proof is relying on similar argument as the Cochran theorem, it slightly differs and have been first shown in Hunter (2015). We define $\Gamma = diag(p)$. The central limit theorem states that

$$\sqrt{n}\Gamma^{-1/2}(\bar{X} - p) \xrightarrow[n\to\infty]{\text{d}} N_k(0, \Gamma^{-1/2}\Sigma\Gamma^{-1/2}) \tag{21}$$

Noticing that $\Sigma = \Gamma - pp^T$, we have

$$\Gamma^{-1/2}\Sigma\Gamma^{-1/2} = I_k - \Gamma^{-1/2}pp^T\Gamma^{-1/2} = I_k - \sqrt{p}\sqrt{p}^T \tag{22}$$

Using the linearity and the commutativity property of the trace, we have

$$\text{Trace}(\Gamma^{-1/2}\Sigma\Gamma^{-1/2}) = \text{Trace}(I) - \text{Trace}(\sqrt{p}^T\sqrt{p}) = k-1 \tag{23}$$

since $\sqrt{p}^T\sqrt{p} = 1$. We can also notice that

$$(I_k - \sqrt{p}\sqrt{p}^T)^2 = I_k - 2\sqrt{p}\sqrt{p}^T + \sqrt{p}\sqrt{p}^T = I_k - \sqrt{p}\sqrt{p}^T \tag{24}$$

which means that $\Gamma^{-1/2}\Sigma\Gamma^{-1/2}$ is an idempotent matrix. Denoting by $A_n = \sqrt{n}\Gamma^{-1/2}(\bar{X} - p)$, we can notice that χ^2 is the squared norm of A_n: $\chi^2 = A_n^T A_n = \|A_n\|^2$. Since $I_k - \sqrt{p}\sqrt{p}^T$ is idempotent, it is a projection matrix of rank equal to its trace: $k - 1$. We can conclude using the following lemma found in Hunter (2015).

Lemma 3.1. *Suppose P is a projection matrix. Then if $Z \sim N_k(0, P)$, $Z^T Z \sim \chi^2(r)$ where r is the trace of P or equivalently the number of eigen values of P equal to 1 or equivalently the number of eigen values of P non equal to 0.*

□.

3.6. Sixth Proof: Generic induction with De Moivre-Laplace theorem

The sixth proof differs from previous ones in the spirit as it proves it using generic induction method. It also uses a weaker form of the central limit theorem (the De Moivre-Laplace theorem) that provides the asymptotic distribution for a binomial distribution. The goal here is to prove the following proposition.

Proposition 1. *if for $k > 2$, the Q_k statistics given by $Q_k = \sum_{j=1}^{k} \frac{(N_{j,k} - np_{j,k})^2}{np_{j,k}}$ converges in distribution to a $\chi^2(k-1)$ with $\sum_{j=1}^{k} p_{j,k} = 1$ and $\sum_{j=1}^{k} N_{j,k} = n$,*

then the $k+1$ statistic given by $Q_{k+1} = \sum_{j=1}^{k+1} \frac{(N_{j,k+1} - np_{j,k+1})^2}{np_{j,k+1}}$ converges in distribution to a $\chi^2(k)$ with $\sum_{j=1}^{k} p_{j,k+1} = 1$ and $\sum_{j=1}^{k} N_{j,k+1} = n$

To emphasize the fact that the probabilities and partition of n at rank k are different from the one at rank $k+1$, we have use the underscore notation for $p_{j,k}$ and $N_{j,k}$. However, when the underscore notation will be obvious, we will drop it to make the computation more readable. We will anyway warn the reader to make sure we do not loose him or her.

Proof. For k=2, Q_k writes as

$$Q_k = \frac{(N_{1,2} - np_{1,2})^2}{np_{1,2}} + \frac{(N_{2,2} - np_{2,2})^2}{np_{2,2}} = \left[\frac{N_{1,2} - np_{1,2}}{\sqrt{np_{1,2}(1-p_{1,2})}}\right]^2 \tag{25}$$

since $p_{1,2} + p_{2,2} = 1$ and $N_{1,2} + N_{2,2} = n$.

De Moivre-Laplace's theorem states that $X_n = \frac{N_{1,2} - np_{1,2}}{\sqrt{np_{1,2}(1-p_{1,2})}}$ converges in distribution to a standard normal distribution $N(0,1)$ since $N_{1,2} \sim Bin(n, p_{1,2})$. Hence, $Q_k \xrightarrow[n\to\infty]{\mathrm{d}} \chi^2(1)$.

Suppose the property to prove is true for $n = k$. We can create a general induction between Q_k and Q_{k+1} as follows:

$$\begin{aligned}
Q_{k+1} &= \sum_{j=1}^{k+1} \frac{(N_{j,k+1} - np_{j,k+1})^2}{np_{j,k+1}} \\
&= \sum_{j=1}^{k-1} \frac{(N_{j,k+1} - np_{j,k+1})^2}{np_{j,k+1}} + \frac{(N_{k,k+1} + N_{k+1,k+1} - n(p_{k,k+1} + p_{k+1,k+1}))^2}{n(p_{k,k+1} + p_{k+1,k+1})} + U^2 \\
&= \sum_{j=1}^{k} \frac{(N'_{j,k} - np'_{j,k})^2}{np'_{j,k}} + U^2 \\
&= Q'_k + U^2
\end{aligned}$$

where we have used the following notations

$$\begin{aligned}
N' &= (N_{1,k+1}, \ldots, N_{k-1,k+1}, N_{k,k+1} + N_{k+1,k+1}), \\
p' &= (p_{1,k+1}, \ldots, p_{k-1,k+1}, p_{k,k+1} + p_{k+1,k+1}), \\
U^2 &= \frac{(N_{k,k+1} - np_{k,k+1})^2}{np_{k,k+1}} + \frac{(N_{k+1,k+1} - np_{k+1,k+1})^2}{np_{k+1,k+1}} - \frac{(N_{k,k+1} + N_{k+1,k+1} - n(p_{k,k+1} + p_{k+1,k+1}))^2}{n(p_{k,k+1} + p_{k+1,k+1})}, \\
Q'_k &= \sum_{j=1}^{k} \frac{(N'_{j,k} - np'_{j,k})^2}{np'_{j,k}}
\end{aligned}$$

By assumption, $Q'_k \xrightarrow[n\to\infty]{\mathrm{d}} \chi^2(k-1)$. The last remaining part is to prove that Q'_k and U are independent and that $U^2 \sim \chi^2(1)$. In the following to make notation lighter, we will drop the lower index $._{k+1}$. Let us denote by $T_k = N_k - np_k$, $T_{k+1} = N_{k+1} - np_{k+1}$, $q_k = \sqrt{p_k}$, $q_{k+1} = \sqrt{p_{k+1}}$, a straight computation leads to

$$\begin{aligned}
U^2 &= \frac{1}{n(p_k + p_{k+1})} \left[\frac{p_{k+1} T_k^2}{p_k} + \frac{p_k T_{k+1}^2}{p_{k+1}} - 2T_k T_{k+1} \right] \\
&= \left[\frac{1}{\sqrt{n(p_k + p_{k+1})}} \right]^2 \left[\frac{q_{k+1} T_k}{q_k} - \frac{q_k T_{k+1}}{q_{k+1}} \right]^2 \\
&= \left[\frac{p_{k+1} N_k - p_k N_{k+1}}{\sqrt{n p_k p_{k+1} (p_k + p_{k+1})}} \right]^2 \\
&= V^2
\end{aligned}$$

where $V = \frac{p_{k+1} N_k - p_k N_{k+1}}{\sqrt{n p_k p_{k+1} (p_k + p_{k+1})}}$. De Moivre-Laplace's theorem states that $N = (N_1, ..., N_{k+1})$ converges in distribution to a Gaussian vector, or consequently that V converges in distri-

bution to a normal distribution, as V is a linear combination of the coordinate of N. V's mean is simple to calculate and equal to 0 since

$$\mathbb{E}[p_{k+1}N_k - p_kN_{k+1}] = n(p_{k+1}p_k - p_kp_{k+1}).$$

V's variance is simple to calculate and equal to 1 since

$$\mathrm{Var}[p_{k+1}N_k - p_kN_{k+1}] = n\left[p_{k+1}^2(p_k(1-p_k)) + p_k^2(p_{k+1}(1-p_{k+1})) - 2p_k^2p_{k+1}^2\right].$$

Hence $V \xrightarrow[n\to\infty]{\mathrm{d}} N(0,1)$. And U^2 converges in distribution to a $\chi^2(1)$ distribution.

The final part is to prove the independence of U and $Q_k^{'}$ or equivalently the independence of $L = p_{k+1}N_k - p_kN_{k+1}$ and $Q_k^{'}$. $Q_k^{'}$ is composed of coordinates of $N^{'}$. So it is is sufficient to prove that L is independent of all the coordinates of $N^{'}$. Both $N^{'}$ and L are coordinates of a Gaussian vector. So, their independence is equivalent to a null covariance. Let us compute. For any $j \leq k+1$, $\mathrm{Cov}(N_j, L) = n(p_{k+1}p_jp_k - p_kp_jp_{k+1}) = 0$. The covariance with the last coordinate of $N^{'}$ is also null as it is $\mathrm{Cov}(N_k + N_{k+1}, L) = \mathrm{Cov}(N_k, L) + \mathrm{Cov}(N_{k+1}, L) = 0 + 0 = 0$. This concludes the proof. □ □

3.7. Seventh Proof: Connection with Poisson variables and geometry

This proof is due to Fisher (1922) and relives on geometry arguments. In the sequel, we shall write k to be an integer, I_k the identity matrix of order k and $Z = (Z_1, Z_2, \ldots, Z_k)$ a random vector with a multi dimensional standard normal distribution $N_k(0, I_k)$ of order dimension k. We shall assume the components of Z to be independent.

Fisher used a geometric argument to determine the distribution of the random variable

$$U := Z_1^2 + Z_2^2 + \ldots + Z_k^2$$

The values $(Z_1, Z_2, \ldots, Z_k)$ of any given sample of Z can be interpreted as co-ordinates of a point P in the k-dimensional Euclidean space $\mathbb{R}^k$.

The Euclidean distance of P from the origin O is written as U and it represents its L^2 norm defined by $U = \|OP\|^2$. One property of the Euclidean distance is to be unchanged by any rotation of the co-ordinates orthonormal axes.

The joint probability density function of the components of Z should therefore be proportional to $e^{-\|OP\|^2/2}$ and should remain constant on the k-dimensional hypersphere with radius $\sqrt{u} = \|OP\|$. A consequence is that the density $f_U(u)$ shall be obtained by integrating it between the two hyperspheres with radius u and $u + du$

$$f_U(u) = ce^{-\|OP\|^2/2} \frac{d}{d\,\|OP\|} \|OP\|^k \tag{26}$$

$$= ce^{u/2} \frac{d}{du} u^{k/2} \tag{27}$$

We shall also impose that there is c a suitable normalization constant to have a density summed to 1. Hence, the constant shall be the following:

$$f_U(u) = \frac{1}{2^{r/2}\Gamma(k/2)} e^{-u/2} u^{k/2-1} \tag{28}$$

Hence, $U \sim \chi^2(k)$.

In the particular case of $s \leq k$ linear (independent) constraints for the components of Z, we shall be able to generalize previous results. Each constraint defines a hyper-plane of $\mathbb{R}^k$ containing O, say π, and Z is forced to belong to it. The intersection of the generic hyper-sphere of $\mathbb{R}^k$ with π is a hyper-sphere of the space $\mathbb{R}^{k-1}$. The result of the s linear constraints will create a hypersphere of $\mathbb{R}^{k-s}$ from a generic hypersphere of $\mathbb{R}^k$. We can apply previous reasoning with the adaptation or replacing k by $k-s$. In other words, if there are $s \leq k$ independent linear constraints for Z, one has

$$U = \sum_{i=1}^{r} Z_i^2 \sim \chi^2(k-s). \tag{29}$$

Let us consider $V = (V_1, V_2, \ldots, V_k)$ consisting of k independent random variables with Poisson distribution with intensity parameter $np_{0,i}$ for $i = 1, 2 \ldots, k : V_i \sim \Pi(np_{0,i})$.

With $(n_1, n_2, \ldots, n_k, n) \in \mathbb{N}^{k+1}$ and $N = \sum_{i=1}^{k} V_i$, at first, one has:

$$P(V_1 = n_1, V_2 = n_2, \ldots, V_k = n_k, N = n) = \prod_{i=1} k \frac{(np_{0,i})^{n_i}}{n_i!} e^{-np_{0,i}} \tag{30}$$

$$= e^{-n} n^n \prod_{i=1} k \frac{(p_{0,i})^{n_i}}{n_i!} \tag{31}$$

Note that, although in the left-hand side of (30), N appears as a random variable, the joint distribution is singular because $N = \sum_{i=1} kN_i$. The marginal distribution of N is easy to determine because N is the sum of k independent random variables with Poisson distribution and with mean $\mathbb{E}(N) = n$. Hence $N \sim \text{Poisson}(n)$. From this fact and from equation (30), one gets the conditional probability:

$$P(V_1 = n_1, V_2 = n_2, \ldots, V_k = n_k | N = n) = \frac{n!}{n_1! n_2! \ldots n_k!} \prod_{i=1} k (p_{0,i})^{n_i} \tag{32}$$

This proves that the distribution of V , under the condition $N = n$ is the same as the one defined by $N = (N_1, N_2, \ldots, N_k)$, the k binomial variables corresponding to the k categorical variables. Let us write

$$\tilde{Z}_i = \frac{V_i - np_{0,i}}{\sqrt{np_{0,i}}}$$

The Central Limit theorem (for Poisson variable with growing intensity parameter) states that the renormalized Poisson variables $\tilde{Z}_i$ converges to standard normals:

$$\tilde{Z}_i = \frac{V_i - np_{0,i}}{\sqrt{np_{0,i}}} \xrightarrow[n\to\infty]{\mathrm{d}} N(0,1)$$

These Gaussian variables are not independent as the variables $\tilde{Z}_i$ have to satisfy the constraint

$$\sum_{i=1}^{k} \sqrt{np_{0,i}}\tilde{Z}_i = 0$$

The above reasoning states that the squared norm of these Gaussian variables should have a $\chi(k-1)$ distribution as they have to satisfy one linear constraints. Wrapping everything up , we have that

$$\sum_{i=1}^{k} \tilde{Z}_i^2 = \sum_{i=1}^{k} \frac{(N - i - np_{0,i})}{np_{0,i}} \xrightarrow[n\to\infty]{\mathrm{d}} \chi^2(k-1)$$

which concludes the proof.

□

4. Graphical interpretation of the test

In the different proofs, we have seen that asymptotically, the variables $\frac{N_j - np_j}{\sqrt{np_j(1-pj)}}$ are normally distributed. This has given us the idea, for a two by two (2x2) contingency table, to not only draw the table with color and size to give an intuition of the relationship between the two variables but also to provide a confidence interval to illustrate whether the second categorical variable is close or not to the first one in the sense of the Pearson correlation test.

Indeed, in statistical graphics, mosaic display, attributed to Hartigan and Kleiner (1981), is a graphical method to show the values (cell frequencies) in a contingency table cross-classified by one or more factors. Figure 1 shows the basic form of a mosaic display for a two-way table of individuals. Canonical examples can be found in Friendly (1992), Friendly (1994) and Friendly (2002). Mosaic displays have become the primary graphical tool for visualization and analysis categorical data in the form of n-way contingency tables. Although

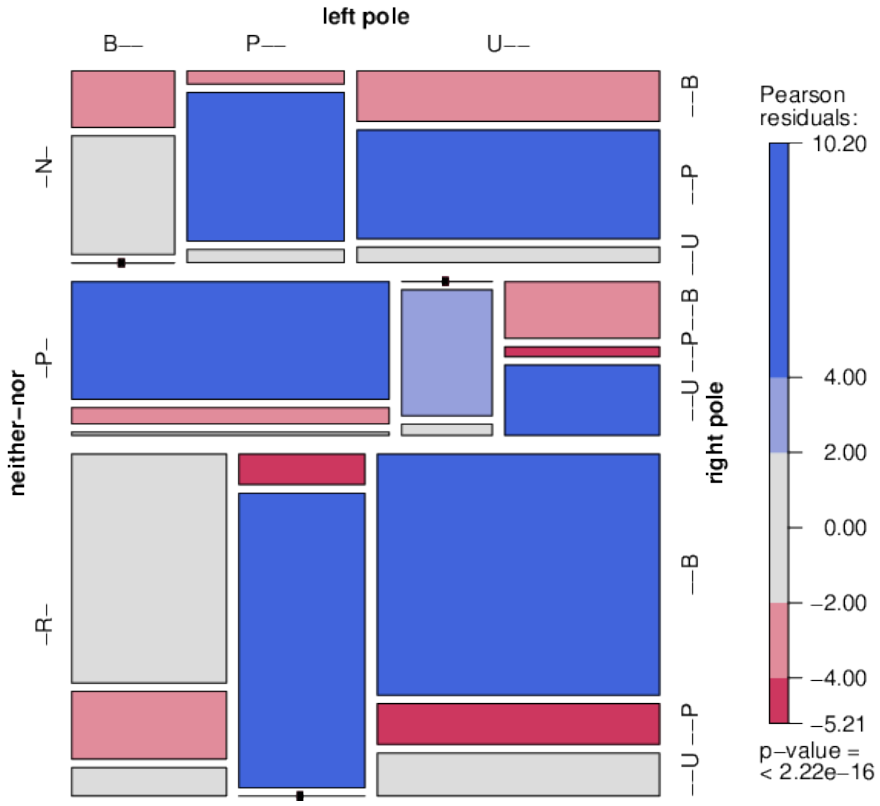

Fig. 1. Standard mosaic for a two-way contingency table. The color and the size provides information about the values of the contingency table. On the right side, the Pearson scale provides information about the residuals. However no confidence interval is provided.

they provide Pearson residuals, it appeared to us that adding a confidence interval could increased readability. This is what we have achieved as follows. The full source code used can be found in github.

The method works as follows: We first compute the total values of the our contingency table. This is the sum of all the values inside the table. We call this N. We estimate the probability $\hat{p}$ over all categories as follows: $\hat{p}$ = values of interest / N. Note that this probability is estimated on all data (under the assumption that both categories are from the same distribution). A confidence interval for the probability estimator is easily calculated from the standard normal distribution: $\delta_p = \text{pdf}(\text{ quantile }) \sqrt{(\hat{p}(1-\hat{p})\ N}$ where pdf stands for the traditional probability density function. We then just to add this confidence interval to our table centered around our estimated probability. In figure 2, we provide the mosaic plot for an initial table given by ideal values provided in 1.

We then change each of the four values with the following new values: 75, 100 and 200 to visualize the impact of a variable that becomes more and more independent. Corresponding

	C	D
A	50	50
B	50	50

Table 1: initial data for the contingency table

graphics are provided in 3, 4 and 5 with corresponding tables 2, 3 and 4. In the initial case (provided in figure mosaic-initial with corresponding values given in table 1), each individual square is of equal size and equal to one fourth of the total square. This is logical as all values are equal to 50. Since the dark red and dark green square lower side are within the confidence interval, at the middle of the confidence interval, we can graphically visualize that the two categorical are from the same distribution (with regards to the Pearson independence test). We will use this case as a benchmark and progressively change the second categorical variables to make it independent (or in the sense of the Pearson test, the second categorical variable will not come from the same distribution from a statistical point of view). The case with a modified value of 75 provides some hindsight about the graphical interpretation of the Pearson test). It is given by table 3. First of all the lower or upper side of the concerned squares lie within the confidence interval. This means that from a statistical point of view, the two categorical variables are from the same distribution. As we see that these two categorical variables are drifting away one from another, we can see that compared to our benchmark the two categorical variables are slightly different, though from the same distribution! Secondly, we see that the confidence interval is shifted upward or downward depending on the case. Because we always first show the categorical variable with more observation for all our figures (3, 4 and 5), the left and right column are the same, though x axis indexes are permuted. Figure 3 is an example of Pearson test where we fail to reject hypothesis H0 meaning we fail to reject H0: the two categorical variables are from the same distribution. Figure 4 is informative. In this case, we are slightly outside the confidence interval indicating that in this particular case, we can reject H0: the two categorical variables are from the same distribution. However, this is way different from the case of figure 5 where we are much more outside the confidence interval. Figure 4 corresponds to a modified value of 100 versus figure 5 that corresponds to a modified value of 200. This is precisely the interest of this graphical representation. We can measure by how much we are from failing to reject hypothesis H0. In all tables that provide the modified value (tables 2, 3 and 4), we have emphasized the modified value by writing it in bold.

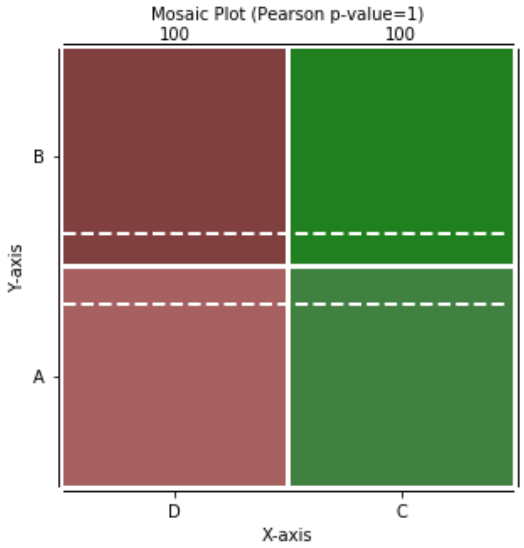

Fig. 2. The cases where all values are equal to 50. Each individual square is of same size. The dark red and dark green square lower side ore limits are within the confidence interval. They lie precisely at the middle of the confidence interval indicating that the two categorical variables are very very similar or from the same distribution (with regards to the Pearson independence test). We can conclude that these two categorical variables are statistically not independent. This case is our benchmark.

	C	D
A	**75**	50
B	50	50

	C	D
A	50	**75**
B	50	50

	C	D
A	50	50
B	**75**	50

	C	D
A	50	50
B	50	**75**

Table 2: Data for the various figures 3. We have changed one of the value to 75

	C	D
A	**100**	50
B	50	50

	C	D
A	50	**100**
B	50	50

	C	D
A	50	50
B	**100**	50

	C	D
A	50	50
B	50	**100**

Table 3: Data for the various figures 4. We have changed one of the value to 100

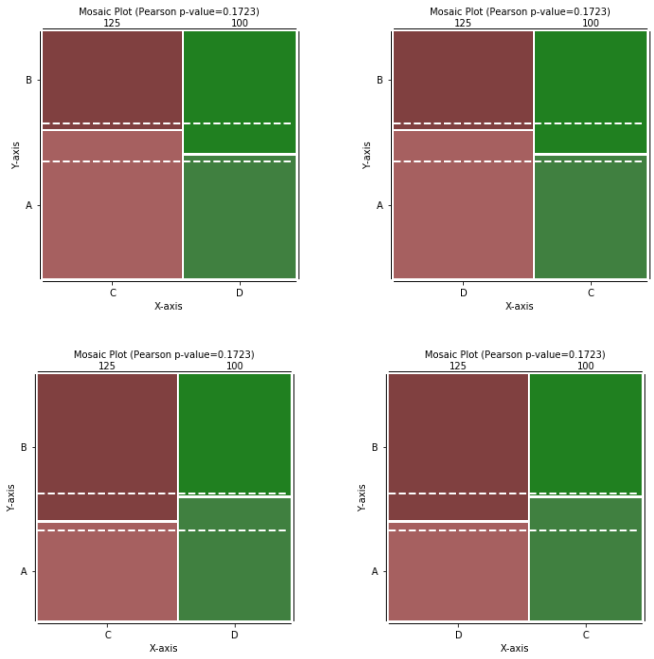

Fig. 3. We plot the four cases where we change one value to 75 and keep the other ones to 50. We start departing from the same distribution. However, in all four cases, the modified value is still within the confidence interval. It is further away from the middle of the large square but still within the confidence interval. It is worth noting that the estimated probability is computed on the full data. Hence, the confidence interval is shifted either upward or downward depending on the case. As we always plot in the first column the categorical variable with more observation, the graphics are the same on the right and left

	C	D
A	**200**	50
B	50	50

	C	D
A	50	**200**
B	50	50

	C	D
A	50	50
B	**200**	50

	C	D
A	50	50
B	50	**200**

Table 4: Data for the various figures 5. We have changed one of the value to 200

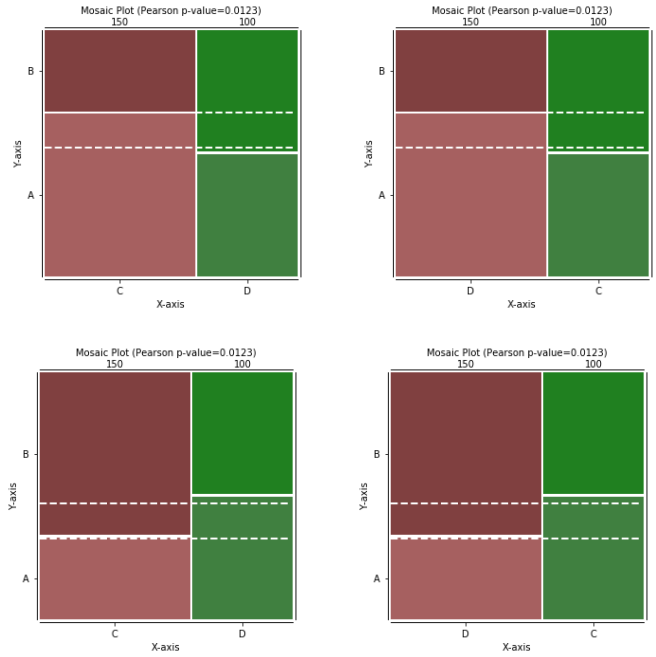

Fig. 4. The four cases where we have change one value to 100 and keep the other to 50. In this case, we are slightly outside from the interval confidence interval. This indicates in particular that if we had taken a higher error of type I, the test could have been successful.

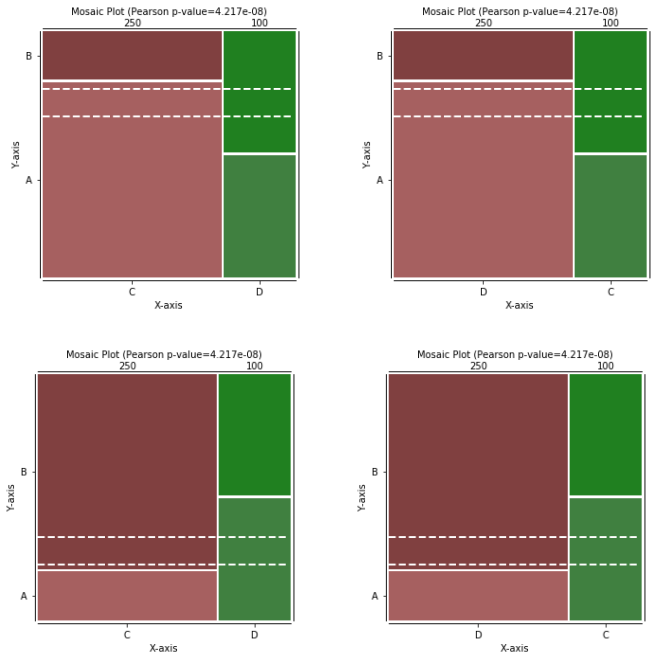

Fig. 5. The four cases where we have change one value to 200 and keep the other to 50. We should note that these figures are very different from figure 4. In our case, we are much more outside from the confidence interval than in the case of figure 4. This is precisely the interest of this graphical representation to be able to compare and dissociate the two cases. Values for these graphics are provided in table4.

5. Conclusion

In this paper, we have revisited the Pearson Chi-squared independence test. We have provided seven proofs for this seminal test. We also present an innovative and intuitive graphical representation of this test with a confidence interval. This enables not only interpreting visually the test but also measuring how close or far we are from accepting or rejecting the null hypothesis of non independence. Further work could be to extend these confidence interval interpretation to contingency tables larger than two by two.

References

Bernard, G., 1945. A new test for 2x2 tables. Nature p. 177.

Bernard, G., 1947. Significance tests for 2x2 tables. Biometrika pp. 123–138.

Buonocore, A., Pirozzi, E., 2014. On the pearson-fisher chi-squared theorem. Applied Mathematical Sciences, HIKARI Ltd, pp. 6733 – 6744.

Cochran, W. G., 1934. The distribution of quadratic forms in a normal system, with applications to the analysis of covariance. Mathematical Proceedings of the Cambridge Philosophical Society 30, 178–191.

Fisher, R. A., 1922. On the interpretation of χ^2 from contingency tables, and the calculation of P. Journal of the Royal Statistical Society 1, 87–94.

Friendly, M., 1992. Mosaic displays for loglinear models. Proceedings of the Statistical Graphics Section pp. 61–68.

Friendly, M., 1994. Mosaic displays for multi-way contingecy tables. Journal of the american statistical association 89, 190–200.

Friendly, M., 2002. A brief history of the mosaic display. Journal of Computational and Graphical Statistics 11, 89–107.

Hartigan, J., Kleiner, B., 1981. Mosaics for contingency tables. in Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface pp. 268–273.

Hunter, D. R., 2015. Notes for a graduate-level course in asymptotics for statisticians. Journal of the american statistical association.

Pearson, K., 1900. On the criterion that a given system 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. Philosophical Magazine 302, 157–175.

Sherman, J., Morrison, W. J., 1949. Adjustment of an inverse matrix corresponding to changes in the elements of a given column or a given row of the original matrix. Annals of Mathematical Statistics pp. 620–624.