À PROPOS DU TEST EXACT DE FISHER

Anthony Phan

Le test exact de Fisher est un test d'indépendance de variables ayant un nombre fini de valeurs possibles qui est employé lorsque les hypothèses d'applicabilité du test du χ^2 d'indépendance ne sont pas satisfaites, en particulier lorsque les effectifs d'un tableau de contingence sont trop petits. Nous nous limiterons au cas où chacune des variables n'a que deux valeurs possibles, c'est-à-dire à des tableaux de contingence de taille 2×2.

Table des matières

1. Préliminaire : tableaux 2×2 de sommes partielles fixées 1
2. Calculs de probabilités 2
3. Les tests d'indépendance 3
4. Le test exact de Fisher, méthode pratique 4
5. Exemples 5
6. Le test exact de Fisher, explications théoriques 6

1. Préliminaire : tableaux 2×2 de sommes partielles fixées

Soient $0 \leqslant k, \ell \leqslant n$ trois entiers. Nous considérons l'ensemble des tableaux à coefficients entiers

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \qquad \text{tels que } a+b+c+d=n,\ a+c=k \text{ et } a+b=\ell.$$

On a alors $b+d=n-k$ et $c+d=n-\ell$, et on constate aisément que

$$b=\ell-a, \qquad c=k-a \qquad \text{et} \qquad d=n-k-\ell+a,$$

de sorte que, si les sommes partielles sont fixées, on n'a qu'un seul degré de liberté, ci-dessus incarné par l'entier a. Puisque $b,\ c \geqslant 0$, on a $a \leqslant \min(k,\ell)$. Puisque $a,\ d \geqslant 0$, on a $a \geqslant \max(0, k+\ell-n)$. Finalement, on a nécessairement

$$\max(0, k+\ell-n) \leqslant a \leqslant \min(k,\ell).$$

On peut vérifier que toutes les valeurs de a satisfaisant ces inégalités conviennent. Nous noterons $M(n,k,\ell)$ l'ensemble de tels tableaux.

2. Calculs de probabilités

Nous considérons deux variables aléatoires X et Y *indépendantes* n'ayant chacunes que 2 valeurs possibles $\{x_1, x_2\}$ et $\{y_1, y_2\}$. Par commodité nous supposerons que $x_1 = 1$, $x_2 = 0$, $y_1 = 1$, $y_2 = 0$ et ainsi que les lois respectives de X et de Y sont les lois de Bernoulli $\mathcal{B}(1, \pi_X)$ et $\mathcal{B}(1, \pi_Y)$ — les indicatrices de deux caractères *indépendants* dans une certaine population. Bien entendu, π_X, $\pi_Y \in [0, 1]$ mais dans la discussion qui suivra nous nous restreindrons à π_X, $\pi_Y \in]0, 1[$, laissant en remarque les cas extrêmes.

Soient $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé, et $\big((X_1, Y_1), \ldots, (X_n, Y_n)\big)$ un échantillon du couple de variables aléatoires (X, Y) : les vecteurs $(X_1, Y_1), \ldots, (X_n, Y_n)$ sont indépendants de même loi que (X, Y). Par hypothèse d'indépendance de X et de Y, cela signifie que ces $2n$ variables aléatoires sont indépendantes et que leurs lois respectives sont ce que l'on pense. Nous notons

$$\begin{aligned}
A &= \mathrm{Card}\{1 \leqslant i \leqslant n : (X_i, Y_i) = (1, 1)\},\\
B &= \mathrm{Card}\{1 \leqslant i \leqslant n : (X_i, Y_i) = (0, 1)\},\\
C &= \mathrm{Card}\{1 \leqslant i \leqslant n : (X_i, Y_i) = (1, 0)\},\\
D &= \mathrm{Card}\{1 \leqslant i \leqslant n : (X_i, Y_i) = (0, 0)\}.
\end{aligned}$$

La loi du vecteur (A, B, C, D) est la loi multinomiale de paramètres $(\pi_A, \pi_B, \pi_C, \pi_D)$ et n, où

$$\begin{aligned}
\pi_A &= \mathbb{P}\{(X_i, Y_i) = (1, 1)\} = \pi_X \pi_Y,\\
\pi_B &= \mathbb{P}\{(X_i, Y_i) = (0, 1)\} = (1 - \pi_X)\pi_Y,\\
\pi_C &= \mathbb{P}\{(X_i, Y_i) = (1, 0)\} = \pi_X(1 - \pi_Y),\\
\pi_D &= \mathbb{P}\{(X_i, Y_i) = (0, 0)\} = (1 - \pi_X)(1 - \pi_Y),
\end{aligned}$$

c'est-à-dire que, pour tous entiers a, b, c et d tels que $a + b + c + d = n$, on a

$$\mathbb{P}\{A = a,\, B = b,\, C = c,\, D = d\} = \frac{n\,!}{a\,!\,b\,!\,c\,!\,d\,!}\,\pi_A^a \pi_B^b \pi_C^c \pi_D^d.$$

Notons $K = A + C$ et $L = A + B$ qui sont respectivement le nombre de 1 parmi les X_i et parmi les Y_i. Ces deux variables aléatoires sont indépendantes de lois respectives les lois binomiales $\mathcal{B}(n, \pi_X)$ et $\mathcal{B}(n, \pi_Y)$: on a $\pi_X = \pi_A + \pi_C$ et $\pi_Y = \pi_A + \pi_B$ naturellement, mais l'indépendance vient de celle des X_i et des Y_i.

Soient k, $\ell \in \{0, \ldots, n\}$. Nous allons calculer des probabilités conditionnelles sachant $\{K = k,\, L = \ell\}$. On a

$$\mathbb{P}\{K = k,\, L = \ell\} = C_n^k\, C_n^\ell\, \pi_X^k (1 - \pi_X)^{n-k} \pi_Y^\ell (1 - \pi_Y)^{n-\ell}.$$

Alors, si a, b, c et d sont des entiers tels que $a + b + c + d = n$, $a + c = k$ et $a + b = \ell$, c'est-à-dire si le tableau

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

appartient à $M(n,k,\ell)$,

$$\begin{aligned}
\mathbb{P}\left\{\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \,\middle|\, K=k,\, L=\ell\right\} &= \mathbb{P}\left\{\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}\right\} \Big/ \mathbb{P}\{K=k,\, L=\ell\} \\
&= \frac{n\,!\;\pi_A^a \pi_B^b \pi_C^c \pi_D^d}{a\,!\;b\,!\;c\,!\;d\,!\,C_n^k\,C_n^\ell\,\pi_X^k(1-\pi_X)^{n-k}\pi_Y^\ell(1-\pi_Y)^{n-\ell}} \\
&= \frac{n\,!\;\pi_X^k(1-\pi_X)^{n-k}\pi_Y^\ell(1-\pi_Y)^{n-\ell}}{a\,!\;b\,!\;c\,!\;d\,!\,C_n^k\,C_n^\ell\,\pi_X^k(1-\pi_X)^{n-k}\pi_Y^\ell(1-\pi_Y)^{n-\ell}} \\
&= \frac{k\,!\;(n-k)\,!\;\ell\,!\;(n-\ell)\,!}{n\,!\;a\,!\;b\,!\;c\,!\;d\,!}\,;
\end{aligned}$$

si les entiers a, b, c et d ne satisfont pas les hypothèses précédentes, cette probabilité conditionnelle est alors nulle.

Remarques. — *a*) Un point tout particulièrement intéressant est que les paramètres π_X et π_Y ont disparu de l'expression des probabilités conditionnelles.

b) La somme de ces probabilités conditionnelles sur l'ensemble des tableaux de $M(n,k,\ell)$ vaut 1.

c) Si l'un des paramètres est nul ou égal à 1, par exemple si $\pi_Y \in \{0,1\}$, alors les tableaux des effectifs ont tous (presque sûrement) une ligne de zéros (la première si $\pi_Y = 0$, la seconde si $\pi_Y = 1$). Le calcul de probabilité conditionnelle n'a de sens que lorsque ℓ est égal à 0 ou n suivant les cas, et donne 1 ($M(n,k,\ell)$ est réduit à un singleton dans ce cas). Notons que dans ce cas l'indépendance de X et de Y découle de la dégénérescence de la loi de Y.

d) Le cas où les deux lois sont dégénérées est immédiat : presque sûrement le tableau des effectifs a une case fixée égale à n, les autres étant nulles.

3. **Les tests d'indépendance**

Nous considérons un échantillon de taille n d'une certaine population sur laquelle sont définis deux caractères X et Y ayant deux valeurs possibles $\{x_1, x_2\}$ et $\{y_1, y_2\}$. Le tableau de contingence s'écrit

y \ x	x_1	x_2	total
y_1	a	b	ℓ
y_2	c	d	$n-\ell$
total	k	$n-k$	n

où a est le nombre d'individus dans l'échantillon pour lesquels le couple de caractères (X,Y) prend la valeur (x_1, y_1), etc.

Les sommes partielles $(k, n-k)$ et de $(\ell, n-\ell)$ n'ont pas nécessairement grand-chose à voir *simultanément* avec les lois de X et de Y, c'est-à-dire qu'*on ne s'attend pas nécessairement à avoir simultanément*

$$\mathbb{P}\{X=x_1\} \approx \frac{k}{n}\,, \qquad \mathbb{P}\{X=x_2\} \approx \frac{n-k}{n}$$

et

$$\mathbb{P}\{Y = y_1\} \approx \frac{\ell}{n}\,, \qquad \mathbb{P}\{Y = y_2\} \approx \frac{n-\ell}{n}\,.$$

Il suffit pour s'en convaincre de considérer les cas des études prospectives et retrospectives dans le cadre des essais cliniques qui correspondent au cadre qui suit.

Nous supposerons que les sommes partielles $(\ell, n-\ell)$ n'ont pas nécessairement de rapport avec la distribution de Y. C'est le cas, par exemple, lorsque, par commodité, l'échantillon a été formé en choisissant délibérément ou arbitrairement le nombre d'individus de caractères y_1 et ceux de caractère y_2.

En contrepartie, on aimerait que les effectifs a, b, c et d reflètent les distributions de X conditionnellement à Y :

$$\mathbb{P}\{X = x_1 \,|\, Y = y_1\} \approx \frac{a}{\ell}\,, \qquad \mathbb{P}\{X = x_2 \,|\, Y = y_1\} \approx \frac{b}{\ell}$$

et

$$\mathbb{P}\{X = x_1 \,|\, Y = y_2\} \approx \frac{c}{n-\ell}\,, \qquad \mathbb{P}\{X = x_2 \,|\, Y = y_2\} \approx \frac{d}{n-\ell}\,.$$

L'hypothétique indépendance de X envers Y peut être suggérée par l'éventuelle proximité $a/\ell \approx c/(n-\ell) \approx k/n$, et par voie de conséquence $b/\ell \approx d/(n-\ell) \approx (n-k)/n$. De là, on tire

$$a \approx \frac{k\ell}{n}\,, \qquad b \approx \frac{(n-k)\ell}{n}\,, \qquad c \approx \frac{k(n-\ell)}{n} \qquad \text{et} \qquad d \approx \frac{(n-k)(n-\ell)}{n}\,.$$

Ces relations sont remarquables car, contrairement à ce que nous avions supposé, il y a dans celles-ci une complète symétrie en X et en Y. Ceci n'est guère surprenant car, affirmer que la loi de X conditionnellement à Y ne dépend pas de Y — relation *a priori* dissymétrique — équivaut à affirmer que les variables X et Y sont indépendantes — qui est une relation symétrique.

Ainsi, un test d'indépendance consiste, quelles que soient — ou presque ! — les modalités de formation de l'échantillon, à comparer les tableaux

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \qquad \text{et} \qquad \begin{pmatrix} k\ell/n & (n-k)\ell/n \\ k(n-\ell)/n & (n-k)(n-\ell)/n \end{pmatrix}.$$

Ce dernier tableau est appelée tableau des *effectifs attendus* et correspondrait (noter qu'il n'y a pas de raison pour que ses coefficients soient entiers) à ce que les effectifs observés reflètent l'indépendance éventuelle de X et de Y.

Le « presque ! » précédent vient de ce qu'on s'attend dans la discussion précédente à ce que les effectifs observés pour la variable X correspondent à sa loi. Ce n'est plus le cas par la suite notamment parce qu'une telle attente ne peut avoir lieu que lorsque les effectifs des deux sous-échantillons de X sont assez grands.

4. Le test exact de Fisher, méthode pratique

Soient $n \in \mathbb{N}^*$, $0 \leqslant k, \ell \leqslant n$ deux entiers et $\alpha \in \,]0,1[$. On sélectionne un échantillon de taille n de (X, Y) pour lequel on a k occurences de x_1 et ℓ occurences de y_1.

On teste

$$H_0 : X \text{ et } Y \text{ sont indépendantes}$$

contre

$$H_1 : X \text{ et } Y \text{ ne sont pas indépendantes.}$$

On dresse la liste des tableaux de $M(n, k, \ell)$ et à chacun d'entre eux on associe la probabilité

$$\frac{k\,!\,(n-k)\,!\,\ell\,!\,(n-\ell)\,!}{n\,!\,a\,!\,b\,!\,c\,!\,d\,!}$$

correspondant à ses coefficients. Soient p_{coupure} la probabilité correspondant au tableau des effectifs de l'échantillon, et p la somme des probabilités des tableaux qui sont inférieures ou égales à p_{coupure} (cette somme comprend le terme p_{coupure}).

Si $p \geqslant \alpha$, on conserve l'hypothèse H_0. Si $p < \alpha$, on rejette l'hypothèse H_0 : le test est alors significatif et, au niveau de confiance $1 - \alpha$, on rejette l'hypothèse que X et Y puissent être indépendantes.

5. **Exemples**

Dans les deux exemples qui suivent les calculs ne sont que très approximatifs car ils ont été réalisés en virgule fixe avec une précision égale à 1/65536. L'accumulation des erreurs n'aidant pas, les divers résultats ne sont qu'illustratifs.

Test exact de Fisher

Tables originale et attendue :

$$\left(\begin{array}{rr|r} 1 & 5 & 6 \\ 3.375 & 2.625 & \\ 8 & 2 & 10 \\ 5.625 & 4.375 & \\ \hline 9 & 7 & 16 \end{array}\right)$$

Probabilité de coupure : 0.0235.

Tables de mêmes sommes par lignes et colonnes et leurs probabilités :

$\begin{pmatrix} 0 & 6 \\ 9 & 1 \end{pmatrix}$ probabilité : 0.00076 ;

$\begin{pmatrix} 1 & 5 \\ 8 & 2 \end{pmatrix}$ probabilité : 0.0235 ;

$\begin{pmatrix} 2 & 4 \\ 7 & 3 \end{pmatrix}$ probabilité : 0.15717 ;

$\begin{pmatrix} 3 & 3 \\ 6 & 4 \end{pmatrix}$ probabilité : 0.36697 ;

$\begin{pmatrix} 4 & 2 \\ 5 & 5 \end{pmatrix}$ probabilité : 0.33035 ;

$\begin{pmatrix} 5 & 1 \\ 4 & 6 \end{pmatrix}$ probabilité : 0.11002 ;

$\begin{pmatrix} 6 & 0 \\ 3 & 7 \end{pmatrix}$ probabilité : 0.01038.

p-valeur bilatérale : 0.03464.

TEST EXACT DE FISHER

Tables originale et attendue :

$$\left(\begin{array}{cc|c} 5 & 5 & 10 \\ 3.0 & 7.0 & \\ 1 & 9 & 10 \\ 3.0 & 7.0 & \\ \hline 6 & 14 & 20 \end{array}\right)$$

Probabilité de coupure : 0.06485.

Tables de mêmes sommes par lignes et colonnes et leurs probabilités :

$\begin{pmatrix} 0 & 10 \\ 6 & 4 \end{pmatrix}$ probabilité : 0.00534 ;

$\begin{pmatrix} 1 & 9 \\ 5 & 5 \end{pmatrix}$ probabilité : 0.06485 ;

$\begin{pmatrix} 2 & 8 \\ 4 & 6 \end{pmatrix}$ probabilité : 0.24368 ;

$\begin{pmatrix} 3 & 7 \\ 3 & 7 \end{pmatrix}$ probabilité : 0.3714 ;

$\begin{pmatrix} 4 & 6 \\ 2 & 8 \end{pmatrix}$ probabilité : 0.24368 ;

$\begin{pmatrix} 5 & 5 \\ 1 & 9 \end{pmatrix}$ probabilité : 0.06485 ;

$\begin{pmatrix} 6 & 4 \\ 0 & 10 \end{pmatrix}$ probabilité : 0.00534.

p-valeur bilatérale : 0.14038.

Remarque. — Le premier test a été effectué à l'aide de la commande

```
\fisher{1}{5}{8}{2}
```

définie dans le fichier `fisher-mac.tex`.

6. **Le test exact de Fisher, explications théoriques**

Contexte théorique. — Considérons l'espace $E = \{x_1, x_2\} \times \{y_1, y_2\}$ muni de sa tribu discrète $\mathcal{E}$, d'applications coordonnées X et Y, et soit Θ l'ensemble des mesures de probabilités sur cet espace. On identifie cet ensemble à celui des quadruplets de réels positifs de somme égale à 1 :

$$\Theta = \left\{\theta = (\pi_{1,1}, \pi_{1,2}, \pi_{2,1}, \pi_{2,2}) : \pi_{i,j} \geqslant 0,\ i,j \in \{1,2\},\ \text{et}\ \sum_{i,j=1}^{2} \pi_{i,j} = 1\right\},$$

c'est-à-dire que, si $\theta \in \Theta$, $\pi_{i,j} = \theta\{x_i, y_j\}$ pour tous i, $j \in \{1, 2\}$. L'ensemble Θ est l'*ensemble des états de la nature* et $(E, \mathcal{E}, \Theta)$ est le *modèle statistique*.

Soient $n \in \mathbb{N}^*$, $\Omega = E^n$ muni de la tribu produit $\mathcal{A}$ qui est encore ici la tribu discrète, et (X_i, Y_i), $i = 1, \ldots, n$, les applications coordonnées. Pour tout $\theta \in \Theta$, nous notons $\mathbb{P}_\theta = \theta^{\otimes n}$ la probabilité produit. Par construction, dans $(\Omega, \mathcal{A}, \mathbb{P}_\theta)$, les n couples de variables aléatoires (X_i, Y_i) sont indépendants et chacun a pour loi la mesure de probabilité θ, ils forment ainsi l'échantillon du couple de variables aléatoires (X, Y).

Posons Θ_0 l'ensemble des $\theta \in \Theta$ pour lesquels $\pi_{1,1}\pi_{2,2} - \pi_{1,2}\pi_{2,1} = 0$, c'est-à-dire des mesures de probabilités pour lesquelles X et Y sont indépendantes : cette condition exprime que les deux colonnes (ou les deux lignes) du tableau $(\pi_{i,j})_{i,j=1,2}$ sont liées, donc que la loi de X sachant Y ne dépend pas de Y (ou que la loi de Y sachant X ne dépend pas de X). Prenons garde au passage de ne pas généraliser trop rapidement cette condition à des dimensions supérieures, c'est-à-dire au cas où X ou Y ont plus de deux valeurs possibles.

Hypothèses, statistique et règle de décision. — La mesure de probabilité $\theta \in \Theta$ est fixée mais inconnue. Soient $n \in \mathbb{N}^*$, $0 \leqslant k, \ell \leqslant n$ deux entiers et $\alpha \in]0, 1[$. Le test d'indépendance se formule selon :

$$H_0 : \theta \in \Theta_0 \qquad \text{contre} \qquad H_1 : \theta \notin \Theta_0.$$

La statistique du test est le tableau des effectifs

$$T = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

où

$$\begin{aligned} A &= \mathrm{Card}\{1 \leqslant i \leqslant n : (X_i, Y_i) = (x_1, y_1)\}, \\ B &= \mathrm{Card}\{1 \leqslant i \leqslant n : (X_i, Y_i) = (x_2, y_1)\}, \\ C &= \mathrm{Card}\{1 \leqslant i \leqslant n : (X_i, Y_i) = (x_1, y_2)\}, \\ D &= \mathrm{Card}\{1 \leqslant i \leqslant n : (X_i, Y_i) = (x_2, y_2)\}, \end{aligned}$$

restreint à l'événement $\{K = A + C = k, L = A + B = \ell\}$. La statistique T est donc à valeurs dans $M(n, k, \ell)$ et est considérée sous la mesure de probabilité $\mathbb{P}\{\,. \mid K = k, L = \ell\}$.

Nous définissons la région de rejet $R_\alpha \subset M(n, k, \ell)$ au niveau de risque α — et par là même, la région d'acceptation $A_\alpha = M(n, k, \ell) \backslash R_\alpha$. Munissons les éléments de $M(n, k, \ell)$

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \qquad \text{du coefficient} \qquad \frac{k\,!\,(n-k)\,!\,\ell\,!\,(n-\ell)\,!}{n\,!\,a\,!\,b\,!\,c\,!\,d\,!}$$

et munissons $M(n, k, \ell)$ de l'ordre $\preccurlyeq$ déduit :

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \preccurlyeq \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix} \iff \frac{k\,!\,(n-k)\,!\,\ell\,!\,(n-\ell)\,!}{n\,!\,a\,!\,b\,!\,c\,!\,d\,!} \leqslant \frac{k\,!\,(n-k)\,!\,\ell\,!\,(n-\ell)\,!}{n\,!\,a'\,!\,b'\,!\,c'\,!\,d'\,!}\,;$$

l'inégalité stricte $\prec$ étant définie de manière similaire. Notons que pour cet ordre il peut y avoir "égalité" de deux tables différentes. Les tableaux inférieurs ou égaux à un tableau fixé est l'ensemble de ses prédécesseurs et il est clair que $\preccurlyeq$ est compatible avec l'inclusion des ensembles de prédécesseurs. L'ensemble $\mathcal{P}$ des ensembles

de prédécesseurs est bien ordonné par l'inclusion. Nous pouvons donc poser sans ambiguïté

$$R_\alpha = \max\left\{P \in \mathcal{P} : \sum_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in P} \frac{k\,!\,(n-k)\,!\,\ell\,!\,(n-\ell)\,!}{n\,!\,a\,!\,b\,!\,c\,!\,d\,!} < \alpha\right\},$$

la région de rejet, et A_α son complémentaire, la région d'acceptation.

Ainsi, si la table observée t appartient à A_α, on conserve H_0 ; sinon t appartient à R_α auquel cas on rejette H_0.

Risque de première espèce. — Supposons que θ appartienne à Θ_0, c'est-à-dire que l'hypothèse d'indépendance H_0 soit satisfaite. Nous avons vu qu'alors

$$\mathbb{P}\left\{T = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \,\middle|\, K = k,\, L = \ell\right\} = \frac{k\,!\,(n-k)\,!\,\ell\,!\,(n-\ell)\,!}{n\,!\,a\,!\,b\,!\,c\,!\,d\,!}$$

pour tout $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in M(n,k,\ell)$. Ainsi, par construction,

$$\mathbb{P}\{T \in R_\alpha \,|\, K = k,\, L = \ell\} < \alpha.$$

Cette probabilité est celle de rejeter à tort l'hypothèse H_0, c'est donc le risque de première espèce. Notons que la région de rejet a été définie comme un ensemble de tables ayant les plus faibles probabilités d'être observées lorsque H_0 est satisfaite, ensemble maximum pour la condition d'avoir une probabilité totale strictement inférieure à α.

Remarques. — *a*) Le risque de première espèce est la fonction définie sur Θ_0 qui pour tout $\theta \in \Theta_0$ a pour valeur la probabilité (qui dépend de θ) que la statistique appartiennent à la région de rejet. Ici, en notant que Θ_0 ne soit pas réduit à un seul point, cette fonction est constante (sur Θ_0).

b) Revenons sur les coefficients pour examiner comment ils varient en fonction de a. Si $\max(0, k+\ell-n) \leqslant a < \min(k,\ell)$, on a

$$\frac{\frac{k\,!\,(n-k)\,!\,\ell\,!\,(n-\ell)\,!}{n\,!\,(a+1)\,!\,(\ell-a-1)\,!\,(k-a-1)\,!\,(n-k-\ell+a+1)\,!}}{\frac{k\,!\,(n-k)\,!\,\ell\,!\,(n-\ell)\,!}{n\,!\,a\,!\,(\ell-a)\,!\,(k-a)\,!\,(n-k-\ell+a)\,!}} = \frac{(\ell-a)(k-a)}{(a+1)(n-k-\ell+a+1)}\,.$$

Ce quotient est strictement supérieur à 1 si et seulement si

$$\begin{aligned}
&(\ell-a)(k-a) > (a+1)(n-k-\ell+a+1)\\
\iff &k\ell-(k+\ell)a+a^2 > (n-k-\ell+1)a+a^2+n-k-\ell+a+1\\
\iff &k\ell+k+\ell-n-1 > (n+2)a\\
\iff &k\ell+k+\ell+1-(n+2) > (n+2)a\\
\iff &\frac{(k+1)(\ell+1)}{n+2}-1 > a.
\end{aligned}$$

De manière identique, le quotient est strictement inférieur, ou égal, à 1 si la condition précédente avec le changement correspondant de relation est satisfaite. Ceci nous montre que les coefficients croissent strictement avec a avant $(k-1)(\ell-1)/(n+2)-1$ et décroissent strictement après. La, ou les, valeur de a pour laquelle le coefficient est maximum correspond à une table proche de la table attendue. Plus a s'écarte de cette valeur, plus la table en diffère. Ceci indique que la région de rejet correspond à

aux tables les plus éloignées de la table attendue. Nous retrouvons ainsi le principe des tests d'indépendance tel qu'il est décrit dans la section 3.

Risque de seconde espèce et puissance du test. — La puissance du test est la fonction définie sur Θ_1 qui pour tout $\theta \in \Theta_1$ a pour valeur la probabilité (qui dépend de θ) que la statistique appartiennent à la région de rejet. C'est donc la fonction qui donne lorsque $\theta \in \Theta_1$ la probabilité de rejeter à raison H_0. Un test n'est intéressant que si cette puissance est grande, ou encore que l'infimum de cette fonction sur Θ_1 est grand. Le risque de seconde espèce est égal à 1 moins la puissance. C'est donc la fonction définie sur Θ_1 qui pour tout $\theta \in \Theta_1$ a pour valeur la probabilité que la statistique appartiennent à la région d'acceptation. Cette fonction indique ainsi quelles sont les probabilités de conserver à tort l'hypothèse H_0 et il importe que son maximum sur Θ_1 soit le plus petit possible.

Soit $\theta \in \Theta$. En toute généralité, nous avons

$$\begin{aligned}
\mathbb{P}\{K=k, L=\ell\} &= \sum\nolimits_{\binom{a\ b}{c\ d}\in M(n,k,\ell)} \frac{n\,!}{a\,!\,b\,!\,c\,!\,d\,!}\,\pi_{1,1}^a\pi_{2,1}^b\pi_{1,2}^c\pi_{2,2}^d \\
&= \sum_{a=\max(0,k+\ell-n)}^{\min(k,\ell)} \frac{n\,!\,\pi_{1,1}^a\pi_{2,1}^{\ell-a}\pi_{1,2}^{k-a}\pi_{2,2}^{n-k-\ell+a}}{a\,!\,(\ell-a)\,!\,(k-a)\,!\,(n-k-\ell+a)\,!} \\
&= \sum_{a=\max(0,k+\ell-n)}^{\min(k,\ell)} \frac{\big(\pi_{1,1}\pi_{2,2}/(\pi_{1,2}\pi_{2,1})\big)^a\,n\,!\,\pi_{2,1}^{\ell}\pi_{1,2}^{k}\pi_{2,2}^{n-k-\ell}}{a\,!\,(\ell-a)\,!\,(k-a)\,!\,(n-k-\ell+a)\,!}\,.
\end{aligned}$$

Ainsi pour tout $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in M(n,k,\ell)$

$$\begin{aligned}
\mathbb{P}\left\{T=\begin{pmatrix} a & b \\ c & d \end{pmatrix} \,\middle|\, K=k,\, L=\ell\right\} &= \frac{\mathbb{P}\big\{T=\binom{a\ b}{c\ d}\big\}}{\mathbb{P}\{K=k,\,L=\ell\}} \\
&= \frac{\frac{(\pi_{1,1}\pi_{2,2}/(\pi_{1,2}\pi_{2,1}))^a\,n\,!\,\pi_{2,1}^{\ell}\pi_{1,2}^{k}\pi_{2,2}^{n-k-\ell}}{a\,!\,(\ell-a)\,!\,(k-a)\,!\,(n-k-\ell+a)\,!}}{\sum_{a'=\max(0,k+\ell-n)}^{\min(k,\ell)} \frac{(\pi_{1,1}\pi_{2,2}/(\pi_{1,2}\pi_{2,1}))^{a'}\,n\,!\,\pi_{2,1}^{\ell}\pi_{1,2}^{k}\pi_{2,2}^{n-k-\ell}}{a'\,!\,(\ell-a')\,!\,(k-a')\,!\,(n-k-\ell+a')\,!}} \\
&= \frac{\frac{(\pi_{1,1}\pi_{2,2}/(\pi_{1,2}\pi_{2,1}))^a}{a\,!\,(\ell-a)\,!\,(k-a)\,!\,(n-k-\ell+a)\,!}}{\sum_{a'=\max(0,k+\ell-n)}^{\min(k,\ell)} \frac{(\pi_{1,1}\pi_{2,2}/(\pi_{1,2}\pi_{2,1}))^{a'}}{a'\,!\,(\ell-a')\,!\,(k-a')\,!\,(n-k-\ell+a')\,!}}\,.
\end{aligned}$$

Notons que lorsqu'il y a indépendance, on a $\pi_{1,1}\pi_{2,2}/(\pi_{1,2}\pi_{2,1}) = 1$ et l'expression restante se simplifie en ce que nous avons déjà rencontré à plusieurs reprises précédemment. En toute généralité, la puissance du test en θ est

$$\frac{\sum_{\binom{a\ b}{c\ d}\in R_\alpha} \frac{q^a}{a\,!\,(\ell-a)\,!\,(k-a)\,!\,(n-k-\ell+a)\,!}}{\sum_{a'=\max(0,k+\ell-n)}^{\min(k,\ell)} \frac{q^{a'}}{a'\,!\,(\ell-a')\,!\,(k-a')\,!\,(n-k-\ell+a')\,!}}$$

où $q = \pi_{1,1}\pi_{2,2}/(\pi_{1,2}\pi_{2,1})$ et il faudrait pouvoir évaluer le minimum de cette quantité sur Θ_1, c'est-à-dire pour $q \in [0,\infty]\backslash\{1\}$. Cette tâche semble à première vue être trop ardue pour être menée théoriquement, c'est-à-dire sans la donnée numérique de n, k, ℓ et α — et même dans cette situation, l'utilisation de méthodes numériques semble nécessaire.