Association de variables qualitatives

Frédéric Bertrand[1]

[1]IRMA, Université de Strasbourg
Strasbourg, France

Master 1
2019

Sommaire

Sommaire

Sommaire

Exemple

Nombre de souris développant une tumeur au poumon après exposition à la fumée de cigarettes

Groupe	Tumeur présente	Tumeur absente	Total
Contrôle	19	13	32
Traitement	21	2	23

d'après Essenbergs, Science, 1952.

Question : Existe-t-il une association entre le développement de la maladie et l'apparition du cancer ?

Remarque

Lorsque nous disposons de deux variables qualitatives X et Y, les moyennes et les variances n'existent plus.

Par conséquent, les coefficients comme le coefficient de corrélation linéaire ou les rapports définis dans les autres chapitres n'ont plus lieu d'exister.

Il ne reste donc qu'un seul élément exploitable : **la loi conjointe du couple** (X, Y)**.**

Trois questions naturelles

À partir de cette information, trois questions semblent naturelles et pertinentes :

Q1. Les variables X et Y sont-elles indépendantes ?

Q2. Les distributions conditionnelles de Y sachant X (respectivement X sachant Y) sont-elles homogènes ?

Q3. La distribution du couple (X, Y) est-elle « proche » d'une distribution théorique ?

Indépendance entre deux variables

Une méthode pour répondre à la première question :

« Les variables X et Y sont-elles indépendantes ? »

consiste :

- à construire le tableau de contingence associé aux variables X et Y sous l'hypothèse d'indépendance, lequel est obtenu en effectuant le produit des fréquences marginales.

Indépendance entre deux variables (suite)

- Puis comparer la distribution empirique, c'est-à-dire celle contenue dans le tableau de contingence, avec la distribution théorique, c'est-à-dire celle obtenue par calcul. L'interprétation résultant de la comparaison de ces deux distributions est alors la suivante :
  1. Si les deux distributions sont identiques, les variables X et Y sont indépendantes.
  2. Si les deux distributions sont différentes, les variables X et Y ne sont pas indépendantes (elles peuvent être liées, corrélées ou non-corrélées).

Remarque

Pour autant, il est très rare en pratique, même dans le cas de variables réellement indépendantes, d'observer une égalité des distributions théoriques et empiriques et cela pour deux raisons :

1. du fait que nous observons un échantillon et non pas la population entière
2. à cause des erreurs de mesure.

Test du Khi-deux

Indépendance

Contexte du test

Le test du χ^2 d'indépendance sert à étudier la liaison entre deux caractères qualitatifs X et Y.

Nous considérons donc le tableau ci-après où correspond au nombre d'individus observés ayant la modalité i pour X et la modalité j pour Y.

Tableau de données

X \ Y	Modalité 1	$\cdots$	Modalité J	Totaux
Modalité 1	$m_{1,1}$	$\cdots$	$m_{1,J}$	$m_{1,\bullet}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
Modalité i	$m_{i,1}$	$\cdots$	$m_{i,J}$	$m_{i,\bullet}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
Modalité I	$m_{I,1}$	$\cdots$	$m_{I,J}$	$m_{I,\bullet}$
Totaux	$m_{\bullet,1}$	$\cdots$	$m_{\bullet,J}$	$m_{\bullet,\bullet} = n$

Contexte du test (suite)

La notation $m_{i,\bullet}$ correspond à $\sum_{j=1}^{J} m_{i,j}$ et la notation $m_{\bullet,j}$ correspond à $\sum_{i=1}^{I} m_{i,j}$. Le principe du test consiste à comparer les effectifs tels que nous les avons, à la répartition que nous aurions si les variables étaient indépendantes. Dans ce cas, en considérant que les marges

$$(m_{1,\bullet}, \ldots, m_{i,\bullet}, \ldots, m_{I,\bullet}, m_{\bullet,1}, \ldots, m_{\bullet,j}, \ldots, m_{\bullet,J})$$

sont fixées, nous pouvons calculer cette répartition théorique dans chacun des échantillons. Nous avons alors :

$$c_{i,j} = \frac{m_{i,\bullet} m_{\bullet,j}}{m_{\bullet,\bullet}}.$$

Hypothèses du test

$\mathcal{H}_0$: Les variables X et Y sont indépendantes

contre

$\mathcal{H}_1$: Les variables X et Y ne sont pas indépendantes.

Conditions d'application du test

Considérons un échantillon formé de réalisations indépendantes du couple de variables aléatoires (X, Y) et de taille $n = m_{\bullet,\bullet}$. Les **effectifs théoriques $c_{i,j}$** et l'**effectif total de l'échantillon $m_{\bullet,\bullet}$** doivent vérifier les inégalités :

$$c_{i,j} \geqslant 5 \quad \text{et} \quad m_{\bullet,\bullet} \geqslant 50.$$

Statistique du test

Si l'hypothèse nulle $\mathcal{H}_0$ est vraie et lorsque les conditions d'application du test sont remplies,

$\chi^2(obs) = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(m_{i,j} - c_{i,j})^2}{c_{i,j}}$ est une réalisation d'une variable aléatoire qui suit approximativement la loi du Khi-deux à $(I-1)(J-1)$ degrés de liberté.

Règle de décision et conclusion du test

Pour un seuil fixé α, les tables de la loi du Khi-deux à $(I-1)(J-1)$ degrés de liberté nous fournissent une valeur critique c_α telle que $\mathbb{P}_{\mathcal{H}_0}\left(\chi^2((I-1)(J-1)) \leqslant c_\alpha\right) = 1-\alpha$.
Alors nous décidons :

$$\begin{cases} \text{si } \chi^2(obs) \geqslant c_\alpha & \mathcal{H}_1 \text{ est vraie}, \\ \text{si } \chi^2(obs) < c_\alpha & \mathcal{H}_0 \text{ est vraie}. \end{cases}$$

Remarque : Dans le cas où nous ne pouvons pas rejeter l'hypothèse nulle $\mathcal{H}_0$ et par conséquent nous l'acceptons, nous devrions calculer le risque de seconde espèce β du test.

Remarques

Ce test pose plusieurs difficultés :

1. Ce test, tel qu'il est exposé, ne peut pas être appliqué à des échantillons appariés.
2. S'il y a plus de deux modalités, nous pouvons essayer d'en regrouper si cela est possible, c'est-à-dire si cela a un sens.
3. Comment faire les regroupements en classe lorsque cela s'avère nécessaire, par exemple si les variables étudiées sont continues ?

Remarques (suite)

4. **Les conditions d'application sont très contraignantes**.

   Un effectif total n **supérieur à 50** et des fréquences d'apparition toutes supérieures à 5.

   Dans le livre de J. Bouyer, [1], ainsi que dans celui de G. Pupion et P.-C. Pupion, [3], il est indiqué que le test est encore utilisable si les effectifs théoriques sont tous supéreurs à 3.

   J. Bouyer, [1], évoque même la possibilité de se contenter du fait qu'il y ait moins de 20 % des cellules pour lesquelles les effectifs théoriques soient inférieurs à 5.

Remarques (suite)

(Suite) Néanmoins tous les auteurs s'entendent pour dire que si dans une telle situation vous obteniez des valeurs proches de la significativité, il est impératif de compléter l'étude par l'utilisation de certains des autres tests présentés ici.

Remarques (suite)

5. Lorsque les conditions ne sont pas remplies, il existe des corrections, par exemple celle de Yates ou les tests exacts de Fisher et de Fisher-Freeman-Halton présentées ci-après.
6. La nécessité de fondre plusieurs modalités en une seule pour que les conditions d'applications, mentionnées au paragraphe ci-dessus, soient remplies modifie les variables sur lesquelles porte le test.

Remarques (suite)

8. Ce test ne tient pas compte de l'éventuelle présence d'un ordre sur les lignes ou les colonnes du tableau de contingence.

   Si l'on peut ordonner les modalités de l'un des deux facteurs, on préférera utiliser un test de Kruskal-Wallis et si l'on peut ordonner les modalités des deux facteurs on utilisera un test de Jonckheere-Terpstra.

Remarques (suite)

(Suite) On pourra alors, si l'on rejette l'hypothèse nulle $\mathcal{H}_0$: « Le facteur X n'a pas d'effet sur la réponse Y », étudier les raisons à l'origine de la non-indépendance à l'aide d'un test post-hoc.

Le cas d'un tableau à deux lignes et k colonnes ou à h lignes et deux colonnes peut également être étudié à l'aide d'un test de Mann-Whitney.

8. Pour des modélisations plus complexes et, par exemple, l'étude de corrélations partielles, nous pourrions utiliser un **modèle log-linéaire**.

Correction de Yates

Lorsque l'on étudie l'indépendance de deux variables et que certaines des fréquences attendues sous l'hypothèse nulle $\mathcal{H}_0$: « X et Y sont indépendantes » sont inférieures à 5, on peut corriger la statistique du test pour prendre en compte cette situation.

Attention il faut néanmoins que toutes les fréquences attendues soient supérieures à 3.

La correction de Yates est une correction de continuité qui consiste à utiliser la statistique de test modifiée de la manière suivante.

Statistique modifiée de Yates

$$\chi^2_n = \sum_{i=1}^{h}\sum_{j=1}^{k} \frac{\left(\left|N_{i,j} - \frac{N_{i,\bullet} \times N_{\bullet,j}}{n}\right| - \frac{1}{2}\right)^2}{\frac{N_{i,\bullet} \times N_{\bullet,j}}{n}}$$

où $N_{i,j}$ est le nombre aléatoire de couple (X_i, Y_j) qui dans un échantillon indépendant identiquement distribué de taille n appartient à la classe $C_{i,j}$, $N_{i,\bullet} = \sum_{j=1}^{k} N_{i,j}$ et $N_{\bullet,j} = \sum_{i=1}^{h} N_{i,j}$.

Remarques

Attention, on n'utilisera la correction que si l'une des fréquences attendues est strictement inférieure à 5.

Si au contraire elles sont toutes supérieures ou égales à 5, on montre que la correction de Yates ne modifie que peu la valeur de la réalisation de la statistique du test.

Comme le signale J Bouyer, [1], ce point de vue, c'est-à-dire le fait de réserver cette correction à de petits échantillons, n'est pas partagé par tous les auteurs.

Remarques (suite)

Si l'un des effectifs théoriques est inférieur à 3, on n'a pas d'autre solution que d'appliquer le test exact de Fisher décrit ci-après.

Plus généralement, P. Dagnélie, [2], conseille l'utilisation systématique du test exact de Fisher lorsque l'effectif total n est inférieur ou égal à 40.

Étude des résidus

Lorsque **l'hypothèse d'indépendance est vérifiée**, les termes dont les carrés sont les contributions à la valeur χ^2_n :

$$\frac{N_{i,j} - \frac{N_{i,\bullet} \times N_{\bullet,j}}{n}}{\sqrt{\frac{N_{i,\bullet} \times N_{\bullet,j}}{n}}}$$

sont approximativement des variables aléatoires qui suivent des lois normales centrées réduites.

Étude des résidus (suite)

On a montré qu'une meilleure approximation de la loi normale centrée réduite est obtenue si l'on considère les valeurs ci-dessus et que l'on les divise par des estimations des écarts types correspondants :

$$\sqrt{\left(1-\frac{N_{i,\bullet}}{n}\right)\left(1-\frac{N_{\bullet,j}}{n}\right)}.$$

Étude des résidus (suite)

On obtient finalement les **écarts réduits** :

$$\frac{N_{i,j} - \frac{N_{i,\bullet} \times N_{\bullet,j}}{n}}{\sqrt{\left(1 - \frac{N_{i,\bullet}}{n}\right)\left(1 - \frac{N_{\bullet,j}}{n}\right)\frac{N_{i,\bullet} \times N_{\bullet,j}}{n}}}$$

On peut alors étudier les écarts réduits comme on le ferait pour des résidus obtenus après une régression par exemple : étude de la normalité, via un diagramme quantile-quantile par exemple, identification d'éventuelles valeurs aberrantes ou non représentatives, voir le cours sur ce sujet.

Retour à l'exemple

Nombre de souris développant une tumeur au poumon après exposition à la fumée de cigarettes

Groupe	Tumeur présente	Tumeur absente	Total
Contrôle	19	13	32
Traitement	21	2	23

d'après Essenbergs, Science, 1952.

Question : Existe-t-il une corrélation entre le développement de la maladie et l'apparition du cancer ?

Réponse

Pour **tester** l'existence de ce lien, il est possible de procéder à un test du χ^2 :

```
Les dénombrements attendus sont notés sous les
dénombrements observés
        Succès  Echec  Total
     1      21      2     23
         16,73   6,27

     2      19     13     32
         23,27   8,73

 Total      40     15     55
Khi deux = 1,091 + 2,910 + 0,784 + 2,092 = 6,878
DL = 1, c = 3,841
```

Sommaire

Tests de Mac-Nemar et d'homogénéité marginale

Test de Mac-Nemar

Contexte du test

Ce test s'applique à des tableaux de contingence à deux lignes et à deux colonnes qui dénombrent les résultats de deux tests obtenus sur les mêmes individus. Chacun des tests peut prendre la valeur A ou B et, de ce fait, ces tableaux sont donc de la forme suivante :

X/Y	A	B	Totaux
A	$n_{1,1}$	$n_{1,2}$	$n_{1,\bullet}$
B	$n_{2,1}$	$n_{2,2}$	$n_{2,\bullet}$
Totaux	$n_{\bullet,1}$	$n_{\bullet,2}$	n

Hypothèses testées

L'hypothèse nulle $\mathcal{H}_0$ que nous cherchons à tester est que les totaux marginaux de chaque réponse sont les mêmes pour chacun des deux tests :

$$n_{1,1} + n_{1,2} = n_{1,1} + n_{2,1} \quad \text{et} \quad n_{2,1} + n_{2,2} = n_{1,2} + n_{2,2}.$$

Ainsi le jeu d'hypothèses auquel le test de Mac-Nemar permet de s'intéresser est :

$$\mathcal{H}_0 : n_{1,2} = n_{2,1}$$

contre

$$\mathcal{H}_1 : n_{1,2} \neq n_{2,1}.$$

Conditions d'application du test

Les conditions sont les suivantes :

$$n_{1,2} + n_{2,1} \geqslant 20 \quad \text{et} \quad n_{1,2} \quad \text{et} \quad n_{2,1} \quad \text{assez grands.}$$

Statistique du test

La valeur de la statistique du test de Mac-Nemar, avec correction de continuité de Yates, observée sur l'échantillon est donnée par la formule suivante :

$$McN(obs) = \frac{\left(\left|n_{1,2} - n_{2,1}\right| - 1\right)^2}{n_{1,2} + n_{2,1}}.$$

Remarque

Pour obtenir la statistique du test sans correction de continuité, nous remplaçons le numérateur $\left(\left|n_{1,2}-n_{2,1}\right|-1\right)^2$ par $\left(n_{1,2}-n_{2,1}\right)^2$.
Si l'hypothèse nulle $\mathcal{H}_0$ est vérifiée et lorsque les conditions d'application sont remplies, la statistique *McN* suit approximativement la loi du Khi-deux $\chi^2(1)$.

Règle de décision et conclusion du test

Pour un seuil fixé α, les tables de la loi du χ^2, à 1 degré de liberté, nous fournissent une valeur critique c_α telle que $\mathbb{P}_{\mathcal{H}_0}\left(\chi^2_1 \leqslant c_\alpha\right) = 1 - \alpha$. Alors nous décidons :

$$\begin{cases} \text{si } \chi^2(obs) \geqslant c_\alpha & \mathcal{H}_1 \text{ est vraie}, \\ \text{si } \chi^2(obs) < c_\alpha & \mathcal{H}_0 \text{ est vraie}. \end{cases}$$

Dans le cas où nous ne pouvons pas rejeter l'hypothèse nulle $\mathcal{H}_0$ et par conséquent nous l'acceptons, nous devons calculer le risque de seconde espèce β du test.

Règle de décision et conclusion du test (suite)

Si nous utilisons un logiciel de statistique celui-ci nous fournit une *p*-valeur. Alors nous décidons :

$$\begin{cases} \textit{si } p\text{-valeur} \leqslant \alpha & \mathcal{H}_1 \textit{ est vraie.} \\ \textit{si } p\text{-valeur} > \alpha & \mathcal{H}_0 \textit{ est vraie.} \end{cases}$$

Remarques

1. Ce test permet de comparer deux proportions dans le cas où les deux échantillons sont appariés.
2. Le **test d'homogénéité marginale** est une extension du test de Mac-Nemar au cas des variables qualitatives à plus de deux modalités. Il permet de comparer des proportions, au nombre de deux ou plus, lorsque les échantillons, au nombre de deux ou plus, sont dépendants.

Sommaire

Cadre d'application

Dans deux populations indépendantes, un caractère qualitatif X pouvant prendre la modalité A est observé.

Les fréquences d'apparition de A dans les deux populations sont les nombres inconnus $\pi_{A,1}$ et $\pi_{A,2}$.

Soit n_i le nombre de personnes présentes dans l'échantillon i, $n_{A,i}$ le nombre de personnes de l'échantillon i qui présentent la modalité A et $f_{A,i}$ la fréquence associée.

Posons $n_\bullet = n_1 + n_2$ la somme des deux effectifs des deux échantillons et $p = (n_{A,1} + n_{A,2})/n_\bullet$.

Les tests unilatéraux se déduisent facilement de la procédure que nous allons introduire.

Hypothèses testées

Nous souhaitons choisir entre les deux hypothèses suivantes :

$$\mathcal{H}_0 : \pi_{A,1} = \pi_{A,2}$$

contre

$$\mathcal{H}_1 : \pi_{A,1} \neq \pi_{A,2}.$$

Conditions d'application du test

Les deux échantillons sont indépendants et formés de réalisations indépendantes du caractère X. Les effectifs n_1 et n_2 peuvent ne pas être égaux.

Statistique du test

La variable aléatoire $n_{A,1}$ suit la loi hypergéométrique $\mathcal{H}\left(n, n_1, \frac{n_{A,1}+n_{A,2}}{n_\bullet}\right)$.

Règle de décision et conclusion du test

Les valeurs critiques du test, $c_{\alpha/2}$ et $c_{1-\alpha/2}$, sont lues dans une table de la loi hypergéométrique.

Si la valeur de la statistique calculée sur l'échantillon, notée $n_{A,1}(obs)$, n'appartient pas à l'intervalle $]c_{\alpha/2}; c_{1-\alpha/2}[$, alors le test est significatif. Nous rejetons $\mathcal{H}_0$ et nous décidons que $\mathcal{H}_1$ est vraie avec un risque de première espèce α.

Si la valeur de la statistique calculée sur l'échantillon, notée $n_{A,1}(obs)$, appartient à l'intervalle $]c_{\alpha/2}; c_{1-\alpha/2}[$, alors le test n'est pas significatif. Nous conservons $\mathcal{H}_0$ avec un risque de deuxième espèce β.

Test exact de Barnard

Pour le test exact de Barnard, les hypothèses testées et les conditions d'application du test sont les mêmes que celles du test exact de Fisher. La statistique du test est trop complexe pour être décrite dans ce cours. Nous utilisons un logiciel de statistique qui nous fournit une *p*-valeur. Nous décidons alors :

$$\begin{cases} \textit{si } p\text{-valeur} \leqslant \alpha & \mathcal{H}_1 \textit{ est vraie}, \\ \textit{si } p\text{-valeur} > \alpha & \mathcal{H}_0 \textit{ est vraie}. \end{cases}$$

Introduction

L'extension du test exact de Fisher au cas où les deux variables étudiées ont un nombre fini quelconque de modalités, mais supérieur à deux, a été réalisée en premier par G. H. Freeman et J.H. Halton en 1951, c'est pourquoi ce test est aussi parfois appelé test de Freeman-Halton ou de Fisher-Freeman-Halton.

Contexte du test

Soit X et Y deux variables quantitatives discrètes ou qualitatives.

Nous nous donnons un échantillon aléatoire $((X_1, Y_1), \ldots, (X_n, Y_n))$ suivant la loi de (X, Y) ainsi qu'un échantillon $(\boldsymbol{x}, \boldsymbol{y})$ formé d'une réalisation de chaque (X_i, Y_i), $1 \leqslant i \leqslant n$.

Enfin nous notons $\boldsymbol{x}$ l'échantillon des réalisations de X et $\boldsymbol{y}$ l'échantillon des réalisations de Y.

Nous considérons le tableau des effectifs $n_{i,j}$, reproduit ci-après, de chacune des h modalités de X et des k modalités de Y apparaissant dans l'échantillon $(\boldsymbol{x}, \boldsymbol{y})$.

Tableau de données

X \ Y	1	$\cdots$	j	$\cdots$	k	Totaux
1	$n_{1,1}$	$\cdots$	$n_{1,j}$	$\cdots$	$n_{1,k}$	$n_{1,\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
i	$n_{i,1}$	$\cdots$	$n_{i,j}$	$\cdots$	$n_{i,k}$	$n_{i,\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
h	$n_{h,1}$	$\cdots$	$n_{h,j}$	$\cdots$	$n_{h,k}$	$n_{h,\bullet}$
Totaux	$n_{\bullet,1}$	$\cdots$	$n_{\bullet,j}$	$\cdots$	$n_{k,\bullet}$	$n_{\bullet,\bullet} = n$

Contexte du test (suite)

Dans ce tableau, les marges $(n_{1,\bullet}, \ldots, n_{h,\bullet}, n_{\bullet,1}, \cdots, n_{\bullet,k})$ n'apportent pas d'information sur l'éventuelle dépendance de X et de Y.

En effet, elles n'indiquent que la répartition des effectifs entre les h modalités de X, indépendamment de la valeur de Y, et la répartition des effectifs entre les k modalités de Y, indépendamment de la valeur de X.

Ce sont les valeurs prises par $n_{1,1}, n_{1,2}, \ldots, n_{1,k}, n_{2,1}, \ldots, n_{2,k}, \ldots, n_{i,j}, \ldots, n_{h,k-1}$ et $n_{h,k}$ qui servent pour étudier la dépendance de X et de Y.

Famille de tableaux ayant les mêmes marges

L'idée du test exact de Fisher est de considérer l'ensemble Γ des tableaux ayant les mêmes marges $(n_{1,\bullet}, \ldots, n_{h,\bullet}, n_{\bullet,1}, \ldots, n_{\bullet,k})$ que le tableau des effectifs observés ci-dessus :

$$\Gamma_{(n_{\bullet,1},\ldots,n_{\bullet,k})}^{(n_{1,\bullet},\ldots,n_{h,\bullet})} = \left\{ \begin{array}{|c|c|c|c|c|} \hline n_{1,1} & \cdots & n_{1,j} & \cdots & n_{1,k} \\ \hline \vdots & & \vdots & & \vdots \\ \hline n_{i,1} & \cdots & n_{i,j} & \cdots & n_{i,k} \\ \hline \vdots & & \vdots & & \vdots \\ \hline n_{h,1} & \cdots & n_{h,j} & \cdots & n_{h,k} \\ \hline \end{array} \right. ,$$

Famille de tableaux ayant les mêmes marges (suite)

$$\text{avec} \left\{ \begin{array}{rcl} n_{1,1} + \ldots + n_{1,k} & = & n_{1,\bullet} \\ \vdots & = & \vdots \\ n_{h,1} + \ldots + n_{h,k} & = & n_{h,\bullet} \\ n_{1,1} + \ldots + n_{h,1} & = & n_{\bullet,1} \\ \vdots & = & \vdots \\ n_{1,k} + \ldots + n_{h,k} & = & n_{\bullet,k} \end{array} \right\}.$$

Les marges $(n_{1,\bullet}, \ldots, n_{h,\bullet}, n_{\bullet,1}, \ldots, n_{\bullet,k})$ étant fixées, la connaissance de $hk - h - k + 1 = (h-1)(k-1)$ valeurs détermine celle de toutes les autres.

Hypothèses testées

Nous souhaitons choisir entre les deux hypothèses suivantes :

$$\mathcal{H}_0 : X \text{ et } Y \text{ sont indépendantes}$$

contre

$$\mathcal{H}_1 : X \text{ et } Y \text{ sont liées.}$$

Remarque : Ce test peut également être utilisé pour comparer k pourcentages sur l populations.

Statistique de test

Nous montrons que, lorsque l'hypothèse nulle $\mathcal{H}_0$ est vraie, la probabilité d'obtenir un tableau γ appartenant à Γ s'exprime à l'aide d'une loi hypergéométrique généralisée :

$$\begin{aligned}\mathbb{P}(\gamma) &= \frac{1}{n_{1,1}! \times \cdots \times n_{h,k}!(n_{1,1}+\cdots+n_{h,k})!} \\ &\quad \times \Big[(n_{1,1}+\cdots+n_{1,k})! \times \cdots (n_{h,1}+\cdots+n_{h,k})! \times \\ &\quad (n_{1,1}+\cdots+n_{h,1})! \times (n_{1,k}+\cdots+n_{h,k})!\Big] \\ &= \frac{\prod_{i=1}^{h} n_{i,\bullet}! \prod_{j=1}^{k} n_{\bullet,j}!}{\left(\prod_{i=1}^{h}\left(\prod_{j=1}^{k} n_{i,j}!\right)\right)\left(\sum_{i=1}^{h}\sum_{j=1}^{k} n_{i,j}\right)!}.\end{aligned}$$

Statistique de test (suite)

Le principe du test consiste à évaluer la probabilité de rencontrer une distribution aussi anormale ou plus anormale dans un tableau γ de Γ que celle que nous avons observée $\gamma(obs)$:

$$\begin{aligned} p_2 &= \sum_{D(\gamma)\geqslant D(\gamma(obs))} \mathbb{P}(\gamma) \\ &= \mathbb{P}(D(\gamma) \geqslant D(\gamma(obs))) \end{aligned}$$

où la fonction $\gamma \to D(\gamma)$ est définie pour tout $\gamma \in \Gamma$ par :

Statistique de test (suite)

$$D(\gamma) = -2 \ln \left(\frac{(2\pi)^{\frac{(h-1)(k-1)}{2}}}{N^{\frac{(hk-1)}{2}}} \prod_{i=1}^{h} {n_{i,\bullet}}^{\frac{h-1}{2}} \prod_{j=1}^{h} {n_{\bullet,j}}^{\frac{k-1}{2}} \mathbb{P}(\gamma) \right).$$

Règle de décision et conclusion du test

Nous définissons alors la région critique du test comme la partie $\Gamma^\star$ de l'ensemble de référence Γ :

$$\Gamma^\star = \{y \in \Gamma \text{ tels que } D(y) \geqslant D(x)\}.$$

Le test bilatéral de niveau α conduit alors au rejet de l'hypothèse nulle $\mathcal{H}_0$ quand la probabilité totale p_2 calculée par la méthode ci-dessus est inférieure ou égale à α.

Règle de décision et conclusion du test (suite)

Nous remarquons que la *p*-valeur calculée pour le cas d'un test exact de Fisher sur un tableau de taille 2×2 n'est pas obtenue de la même manière qu'ici.

Ceci est dû au fait que nous pouvions déduire toutes les valeurs du tableaux en connaissant uniquement ses marges et une des valeurs des effectifs $n_{1,1}$, $n_{1,2}$, $n_{2,1}$ ou $n_{2,2}$ et ainsi ranger « naturellement » les tableaux dans l'ordre croissant ce qui permettait alors de définir facilement ce que nous entendions par une distribution aussi ou plus anormale que celle du tableau observée.

Il s'agit en fait d'un niveau de **signification maximal** α car la distribution hypergéométrique généralisée sur laquelle repose le test est discrète.

Remarque

- La dénomination « exact » du test vient du fait que nous ne faisons appel à aucune approximation pour calculer la *p*-valeur du test.
- Le niveau de signification du test est d'au plus α, il se peut qu'il soit beaucoup plus faible, ce qui fait de ce test un test conservatif.
- Lorsque les effectifs observés sont relativement importants, il devient difficile d'évaluer de manière exacte la *p*-valeur du test. Certains logiciels proposent alors de la calculer par simulation.

Exemple : Couleur des yeux et des cheveux

Nous cherchons à déterminer si les données suivantes permettent de mettre en évidence une association entre la couleur des yeux et la couleur des cheveux.

Yeux \ Cheveux	Brun	Châtain	Roux	Blond	Totaux
Marron	14	24	5	1	44
Noisette	3	11	3	2	19
Vert	1	6	3	3	13
Bleu	4	17	4	19	44
Totaux	22	58	15	25	120

Exemple (suite)

Le calcul des effectifs attendus montre que les conditions d'utilisation d'un test du Khi-deux ne sont pas remplies puisque **six** cellules ont des effectifs attendus **inférieurs ou égaux à 5** et **quatre** d'entre elles ont même des effectifs attendus **inférieurs ou égaux à 3** :

Yeux \ Cheveux	Brun	Châtain	Roux	Blond
Marron	8,07	21,27	5,50	9,17
Noisette	**3,48**	9,18	**2,38**	**3,96**
Vert	**2,38**	6,28	**1,62**	**2,71**
Bleu	8,07	21,27	5,50	9,17

Exemple (suite)

Il est néanmoins possible d'utiliser un test de exact de Fisher-Freeman-Halton obtenu ici avec le logiciel R en simulant la *p*-valeur à partir d'un échantillon de 500000 tableaux parmi tous ceux pour lesquels il faudrait calculer la fonction $D(\gamma)$.

En évaluant le dénominateur de la formule permettant de calculer $\mathbb{P}(\gamma)$, nous voyons qu'il existe un total de $8,061168 \times 10^{283}$ tableaux de ce type. Ceci explique pourquoi si nous réalisons plusieurs fois ce test la *p*-valeur pourra varier légèrement à chaque essai.

La *p*-valeur est égale à $0,000194 < 0,05$. Nous en déduisons donc une dépendance significative au seuil de $\alpha = 5\%$ entre la couleur des yeux et la couleur des cheveux. Le risque associé à cette décision est un risque de première espèce égal à 5%.


J. Bouyer.
Méthodes Statistiques.
Editions INSERM, 1996.


P. Dagnélie.
Statistique Théorique et Appliquée, volume 2.
De Boeck & Larcier, Bruxelles, 1998.


G. Pupion and P.-C. Pupion.
Tests non paramétriques.
Statistique mathématique et probabilité. Economica, Paris, 1998.