Statistiques appliquées aux sciences

SCI 1018

Analyse de fréquences et test du khi carré (χ^2)

Marc J. Mazerolle

Institut de recherche sur les forêts, Université du Québec en Abitibi-Témiscamingue

Table des matières

1 Introduction

Dans la leçon précédente, nous avons illustré des variantes du test t de Student pour comparer les données de deux groupes indépendants ainsi que pour des données appariées. Nous avons présenté les suppositions sous-jacentes à ces analyses ainsi que plusieurs diagnostics formels et informels permettant de vérifier le respect de ces suppositions. Dans tous les cas, les variables réponses étaient numériques avec des valeurs décimales, comme la hauteur, le temps et la masse. Pour ce genre de données, la décimale représente une valeur observable : une hauteur de 193.24 cm a un sens. Toutefois, les fréquences, comme celles provenant du décompte d'événements d'intérêt prennent uniquement des valeurs entières (discrètes). Par exemple, si nous étudions le nombre de plants avec infestation de parasite ou le nombre de graines viables après hybridation, nous ne pouvons pas observer 12.23 événements. Nous en observerons plutôt 12 ou 13. Ce type de données fera l'objet de cette leçon.

2 Analyse de fréquences : le test du khi carré (χ^2)

Les entiers constituent un type de données qui mène à des analyses particulières. Dans certains cas, on récolte des données de fréquences (ou nombre d'occurrences) dans différentes catégories (p. ex., le sexe) et on désire savoir si la proportion des fréquences diffère selon les catégories. Lorsqu'elles proviennent d'un dispositif complètement aléatoire, les fréquences peuvent être analysées avec le test du khi carré (χ^2, *chi-squared test*). Ce genre de test est parfois appelé **test d'ajustement** ou **test d'adéquation** (*goodness-of-fit test*), car il compare les fréquences observées aux fréquences prédites conformes à H_0.

Exemple 5.1 On fait une étude d'observation pour déterminer s'il y a autant d'étudiants inscrits à la TÉLUQ que d'étudiantes (ratio 1 : 1). Pour ce faire, on décide *a priori* d'échantillonner aléatoirement 300 personnes sur la liste des personnes inscrites à la TÉLUQ. On note ensuite le sexe (homme ou femme) de chaque individu sélectionné dans l'échantillon. La variable sexe définit les deux catégories. On utilise souvent le terme **variable catégorique** pour désigner ce type de variable. Nous dressons un tableau avec les résultats suivants (tableau 1) :

Tableau 1 – Hommes et femmes dans un échantillon de 300 individus à la TÉLUQ.

Hommes	Femmes	TOTAL
124	176	300

2.1 La distribution du khi carré (χ^2)

Le test du khi carré permet de comparer des fréquences observées par rapport aux fréquences prédites selon une hypothèse nulle que l'on veut tester. Ce test du khi carré porte le nom d'une distribution statistique continue, celle du khi carré (χ^2). Cette distribution, tout comme la distribution normale et celle du t de Student, est définie par une fonction de densité. La fonction de densité du khi carré est

$$f(x|\nu) = \frac{\left(\frac{1}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)} x^{\left(\frac{\nu}{2}-1\right)} e^{-\frac{1}{2}x}$$

Malgré son apparence un peu intimidante, cette fonction de densité comporte un seul paramètre, ν, qui correspond aux degrés de liberté. Le symbole Γ représente la fonction gamma, qui comporte une intégrale (dans `R`, elle s'obtient avec `gamma( )`). La fonction de densité du khi carré est définie pour les valeurs > 0. En mots, cette fonction nous donne la densité de probabilité associée à $X = x$ pour un nombre de degrés de liberté donné. La forme de la distribution du khi carré dépend uniquement des degrés de liberté (fig. 1).

Dans `R`, on peut obtenir la densité de probabilité du khi carré associée à une certaine valeur de x à l'aide de `dchisq( )`. La probabilité cumulative s'obtient à l'aide de `pchisq( )` et la fonction quantile à l'aide de `qchisq( )`. Les valeurs obtenues par le test du khi carré sont comparées à cette distribution afin d'évaluer s'il est fréquent d'observer de telles valeurs avec un nombre de degrés de liberté donné lorsque H_0 est vraie.

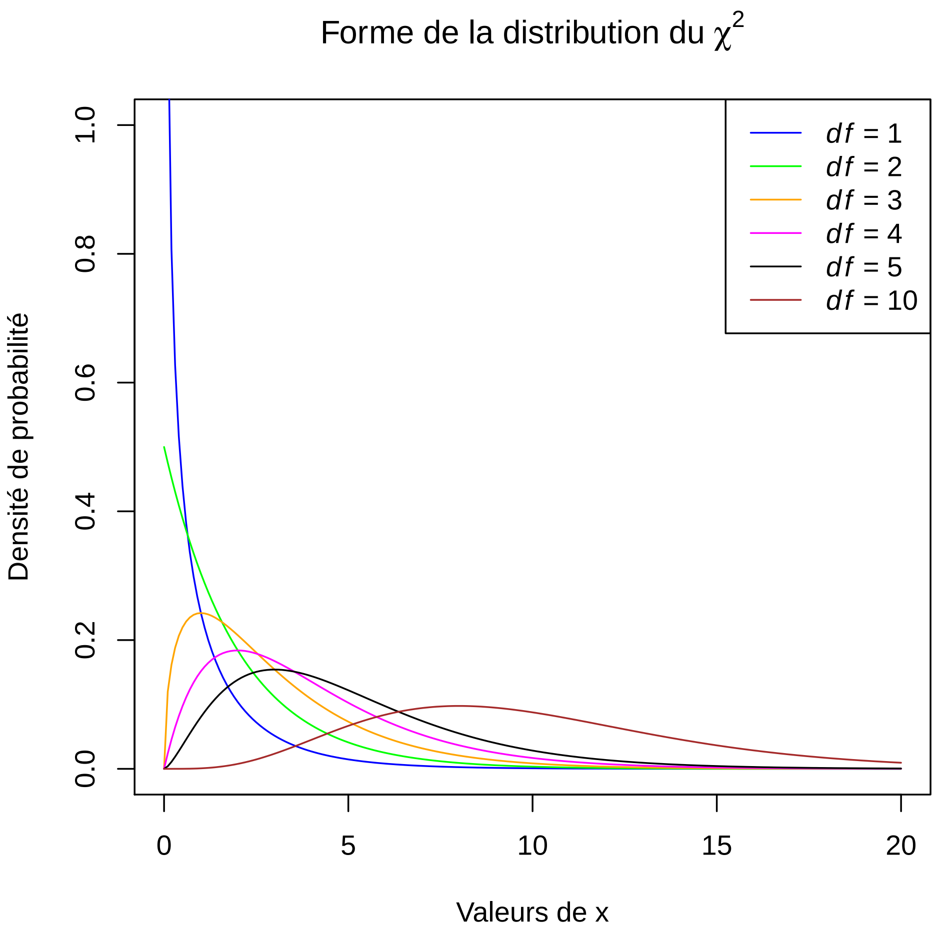

FIGURE 1 – Allure de la distribution du khi carré avec différents degrés de liberté.

2.2 Le test du khi carré (χ^2)

Le test du khi carré est basé sur une comparaison des fréquences observées par rapport à des fréquences attendues si l'hypothèse nulle est vraie. On calcule facilement la statistique du khi carré à l'aide de :

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - \hat{f}_i)^2}{\hat{f}_i}$$

où k correspond au nombre de catégories, f_i donne la **fréquence observée** dans la catégorie i et $\hat{f}_i$ donne la **fréquence théorique** pour la catégorie i si H_0 est vraie. On compare la valeur obtenue du khi carré à la distribution théorique du khi carré correspondant à $k-1$ degrés de liberté ($P(\chi^2_{\alpha,k-1} \geq \chi^2_{observé})$). Si cette valeur est plus grande que ce que l'on devrait observer lorsque H_0 est vraie pour un seuil α et des degrés de liberté donnés, on rejette H_0. Une des propriétés de la statistique du khi carré pour un degré de liberté et un seuil α donné ($\chi^2_{\alpha,1}$) est qu'il correspond au carré du z bilatéral[1], et au carré du t bilatéral pour un même α et pour un degré de liberté qui tend vers l'infini :

$$\chi^2_{\alpha,1} = z^2_{\alpha(2)} = t^2_{\alpha(2),\infty}$$

Exemple 5.2 Reprenons notre exemple sur l'étude d'observation des personnes inscrites à la TÉLUQ. On veut maintenant déterminer si le ratio étudiants : étudiantes est de 1 : 1. Nous pouvons énoncer les hypothèses statistiques suivantes :

H_0 : Le ratio étudiants : étudiantes est de 1 : 1.

H_a : Le ratio étudiants : étudiantes n'est pas de 1 : 1.

$\alpha = 0.05$

On peut appliquer le test du khi carré pour tester ces hypothèses :

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - \hat{f}_i)^2}{\hat{f}_i}$$

On connaît déjà les f_i qui sont les fréquences récoltées dans chaque catégorie

1. Rappel : le z est l'écart normal de la distribution normale centrée réduite.

(tableau 1). Pour ce qui est des fréquences théoriques ($\hat{f}_i$), on doit les obtenir selon l'hypothèse nulle que l'on a spécifiée. Dans notre exemple, puisque H_0 s'intéresse à un ratio 1 : 1, la fréquence théorique dans chaque catégorie s'obtient en divisant le total des fréquences par le nombre de catégories (c.-à-d., $\hat{f}_i = \frac{1}{2}(300) = 150$). On peut ensuite résoudre :

$$\begin{aligned} \chi^2 &= \frac{(124-150)^2}{150} + \frac{(176-150)^2}{150} \\ \chi^2 &= 9.013 \end{aligned}$$

Puisque nous avons deux catégories, le test du khi carré a $k-1 = 2-1 = 1$ degrés de liberté. On observe que la valeur de $P(\chi^2_{\alpha,k-1} \geq 9.013) = 0.0027$:

```
> ##on détermine la probabilité cumulative
> 1 - pchisq(q = 9.013, df = 1, lower.tail = TRUE)
[1] 0.0026807
> ##même chose avec lower.tail = FALSE
> pchisq(q = 9.013, df = 1, lower.tail = FALSE)
[1] 0.0026807
```

Dans `R`, on peut exécuter rapidement le test du khi carré avec la fonction `chisq.test( )`. On doit fournir les fréquences à l'argument `x`, suivi du type d'hypothèse nulle que l'on veut tester à l'aide de l'argument `p`. Ici, puisque H_0 spécifie qu'on devrait avoir autant de fréquences dans les deux groupes, on indique `p = c(0.5, 0.5)` pour avoir la moitié du total des fréquences dans chaque groupe. On peut écrire :

```
> ##on crée un vecteur de fréquences
> ratio <- c(124, 176)
> ##on peut ajouter les étiquettes (optionnel)
```

```
> names(ratio) <- c("Hommes", "Femmes")
> ratio
Hommes Femmes
   124    176
> ##on effectue le khi carré
> out <- chisq.test(x = ratio, p = c(0.5, 0.5))
> out

        Chi-squared test for given probabilities


data:  ratio
X-squared = 9.0133, df = 1, p-value = 0.00268
```

On rejette H_0 car $P \leq \alpha$ (0.05) et on conclut que le ratio étudiants : étudiantes n'est pas de 1 : 1 à la TÉLUQ.

2.3 Conditions d'utilisation

2.3.1 Utilisation des fréquences

Certaines conditions sont nécessaires à l'application correcte du test du khi carré. Ce dernier utilise des fréquences (des entiers) comme données brutes. Le test permet de déterminer si les proportions des fréquences observées sont conformes à l'hypothèse nulle. On ne doit jamais utiliser des proportions (0.2, 0.5) pour réaliser l'analyse.

2.3.2 Indépendance des observations

Le test du khi carré suppose que les données sont récoltées selon un dispositif complètement aléatoire et qu'il n'y a pas de structure dans les données telle que la non-indépendance imposée par l'échantillonnage. On peut illustrer la non-indépendance par un bref exemple où on s'intéresse à l'occurrence de moisissures dans des silos à grains maintenus à l'une de deux différentes températures dans différentes fermes. Dans chacune des fermes, certains silos sont maintenus à 15°C alors que d'autres sont maintenus à 20°C. Si on échantillonne plusieurs silos dans une même ferme, il ne serait pas approprié d'utiliser un khi carré pour comparer les fréquences de moisissures aux deux températures parce que les silos d'une même ferme ne sont pas indépendants. Pour une parfaite indépendance, il faudrait utiliser un seul silo dans chaque ferme étudiée.

2.3.3 Faibles fréquences théoriques

De faibles fréquences théoriques $(\hat{f}_i)$ [2] entraînent un biais du χ^2. En effet, de trop faibles fréquences théoriques surestiment la valeur du χ^2 et, par conséquent, augmentent l'erreur de type I. Certains statisticiens suggèrent que le test du χ^2 est approprié uniquement pour les cas où aucune fréquence théorique n'est inférieure à 3 et où pas plus de 20 % des fréquences théoriques sont inférieures à 5. D'autres sont plus stricts et recommandent de n'avoir aucune fréquence théorique inférieure à 5. Afin d'éviter des problèmes, il est préférable de viser une bonne taille d'échantillon, notamment en s'assurant d'avoir au moins 5 fois plus d'observations que de cellules dans le tableau. En d'autres mots, si on a deux cellules comme dans le tableau de l'exemple 5.2 (homme et femme), on veut au moins un total de 10 individus échantillonnés aléatoirement dans la population d'intérêt.

Continuons avec un exemple sur la sélection d'habitat qui applique le χ^2 en écologie animale.

Exemple 5.3 On s'intéresse à la sélection d'habitat par des Orignaux (*Alces alces*). Pendant deux années, on suit 117 individus munis de colliers émetteurs et on note leur utilisation de différents habitats dans une aire d'étude. À l'aide d'un

2. Rappel : les fréquences théoriques sont les fréquences que l'on aurait dû observer si H_0 est vraie. En contraste, les fréquences observées sont les données brutes utilisées dans l'analyse.

système d'information géoréférencé, on calcule la proportion des quatre habitats qui constituent l'aire d'étude. On veut savoir si les Orignaux sélectionnent un habitat en particulier par rapport à la disponibilité de chaque habitat dans l'aire d'étude. En d'autres mots, on détermine si certains types d'habitats rares sont plus utilisés que des habitats plus communs, ce qui indiquera une réelle préférence d'habitat. Le tableau 2 présente ce jeu de données.

Tableau 2 – Sélection d'habitat par des Orignaux selon la disponibilité d'habitat dans une aire d'étude.

Habitat	Disponibilité	Fréquence
brulis intérieur	0.34	25.00
brulis bordure	0.10	22.00
hors brulis bordure	0.10	30.00
hors brulis loin	0.46	40.00

L'objectif est de déterminer si certains habitats sont surutilisés ou sous-utilisés par les orignaux. On émet les hypothèses statistiques suivantes :

H_0 : L'utilisation des habitats est proportionnelle à la disponibilité de chacun des habitats (0.340 : 0.101 : 0.104 : 0.455).

H_a : L'utilisation des habitats n'est pas proportionnelle à la disponibilité de chacun des habitats (certains habitats sont surutilisés ou sous-utilisés).

$\alpha = 0.05$

On obtient les fréquences théoriques ($\hat{f}_i$) à l'aide des proportions de disponibilité de chaque habitat :

$$\begin{aligned}
\hat{f}_1 &= 0.340 \cdot 117 = 39.78 \\
\hat{f}_2 &= 0.101 \cdot 117 = 11.82 \\
\hat{f}_3 &= 0.104 \cdot 117 = 12.17 \\
\hat{f}_4 &= 0.455 \cdot 117 = 53.23
\end{aligned}$$

Ce qui nous permet ensuite de calculer le χ^2 :

$$\begin{aligned}
\chi^2 &= \sum_{i=1}^{k} \frac{(f_i - \hat{f}_i)^2}{\hat{f}_i} \\
\chi^2 &= \frac{(25-39.78)^2}{39.78} + \frac{(22-11.82)^2}{11.82} + \frac{(30-12.17)^2}{12.17} + \frac{(40-53.23)^2}{53.23} \\
\chi^2 &= 43.689
\end{aligned}$$

Avec $k-1 = 4-1 = 3$ degrés de liberté, on trouve que $P(\chi^2_{0.05,3} \geq 43.689) < 0.0001$.

Avec R, le tout s'obtient facilement à l'aide de :

```
> ##on crée une matrice pour le jeu de données
> select <- matrix(data = c(25, 22, 30, 40), nrow=1)
> ##on ajoute les noms des habitats
> colnames(select) <- c("brulis_interieur", "brulis_bordure",
                        "hors_brulis_bordure", "hors_brulis_loin")
> rownames(select) <- "Freq"
> ##on effectue le test du khi carré
> ##important de spécifier les bonnes valeurs
> ##pour H0
> out.select <- chisq.test(select, p = c(0.340, 0.101, 0.104, 0.455))
> out.select

        Chi-squared test for given probabilities


data:  select
X-squared = 43.689, df = 3, p-value = 1.757e-09
```

On conclut qu'il est peu probable d'observer une valeur de $\chi^2 \geq 43.689$ avec 3 degrés de liberté dans une population où l'utilisation de l'habitat par les Orignaux est proportionnelle à la disponibilité de l'habitat. Effectivement, on rejette H_0.

Certains types d'habitat sont sous-utilisés et d'autres surutilisés par rapport à leur disponibilité dans l'aire d'étude. De façon équivalente, on peut dire que l'utilisation de différents habitats par les Orignaux est indépendante de la disponibilité de cet habitat.

Afin de nous aider à interpréter les résultats, on peut examiner les catégories qui contribuent le plus au χ^2 :

```
> ##on vérifie les éléments présents dans l'objet
> names(out.select)
[1] "statistic" "parameter" "p.value"   "method"
[5] "data.name" "observed"  "expected"  "residuals"
[9] "stdres"
> ##les fréquences observées
> out.select$observed
[1] 25 22 30 40
> ##les fréquences théoriques
> out.select$expected
[1] 39.780 11.817 12.168 53.235
```

On constate que les deuxième et troisième catégories d'habitat sont beaucoup plus utilisées que ce qui est prédit selon leur disponibilité, alors que les premières et quatrième catégories sont moins utilisées que ce qui est prédit selon leur disponibilité.

2.3.4 Correction de continuité

Les calculs à partir de la formule du χ^2 approximent bien la distribution théorique du khi carré, sauf dans le cas où il n'y a qu'un seul degré de liberté ($df = 1$). Dans de tels cas, on peut utiliser la correction de continuité de Yates pour réduire le biais :

$$\chi^2_{corr} = \sum_{i=1}^{k} \frac{(|f_i - \hat{f}_i| - 0.5)^2}{\hat{f}_i}$$

Le prochain exemple illustre justement l'application de la continuité de correction pour un χ^2 avec 1 degré de liberté.

Exemple 5.4 Dans une expérience en génétique, on veut savoir si un échantillon constitué de 100 plantes provient d'une population où le ratio entre le nombre de fleurs jaunes et celui de fleurs vertes est de 3 : 1. Après avoir récolté 100 plantes, on remarque que 84 plantes ont des fleurs jaunes et 16 plantes ont des fleurs vertes. On émet les hypothèses statistiques suivantes :

H_0 : L'échantillon provient d'une population avec un ratio fleurs jaunes : fleurs vertes de 3 : 1.

H_a : L'échantillon ne provient pas d'une population avec un ratio fleurs jaunes : fleurs vertes de 3 : 1.

$\alpha = 0.05$

Puisqu'on veut tester un ratio 3 : 1, les fréquences théoriques conformes à l'hypothèse nulle seront obtenues en multipliant le nombre total de fréquences par 0.75 (3/4) et 0.25 (1/4) pour les fleurs jaunes et les fleurs vertes, respectivement. On peut réaliser l'analyse dans `R` :

```
> ##on crée une matrice avec les fréquences
> freq.fleurs <- matrix(data = c(84, 16), nrow = 1)
```

```
> ##on indique H0
> H0.prop <- c(0.75, 0.25)
> ##on effectue le test du khi carré
> khi2 <- chisq.test(x = freq.fleurs, p = H0.prop)
> khi2

        Chi-squared test for given probabilities

data:  freq.fleurs
X-squared = 4.32, df = 1, p-value = 0.03767
```

On remarque que le test a un degré de liberté. Il faut donc appliquer la correction de continuité de Yates à notre valeur comme suit :

```
> ##khi carré original
> khi.orig <- sum(((khi2$observed - khi2$expected)^2)/
                  khi2$expected)
> khi.orig
[1] 4.32
> ##khi carré corrigé pour continuité
> khi.corr <- sum(((abs(khi2$observed - khi2$expected) - 0.5)^2)/
                  khi2$expected)
> khi.corr
[1] 3.8533
```

Avec le χ^2 corrigé pour la continuité (3.853 comparé à 4.32 pour le χ^2 non corrigé), on rejette tout de même H_0 ($P = 0.0496$) et on conclut que l'échantillon ne provient pas d'une population où il y a un ratio de trois fleurs jaunes pour une fleur verte (3 : 1).

3 Tableau de contingence

Le **tableau de contingence** (*contingency table*) est un tableau à deux dimensions où on note, à la suite d'une étude d'observation ou d'une expérience, la fréquence d'occurrence d'un événement par rapport à deux variables catégoriques (p. ex., sexe, classes de taille). Le tableau de contingence se prête à un **test d'indépendance** entre les deux variables catégoriques. Autrement dit, le tableau de contingence permet de tester l'hypothèse que les fréquences d'occurrence d'un événement pour une variable sont indépendantes des fréquences pour les catégories de l'autre variable. Le test du χ^2 s'applique naturellement au tableau de contingence.

Exemple 5.5 On veut savoir si la couleur des yeux est indépendante de la couleur des cheveux chez des humains d'origine septentrionale. On sélectionne aléatoirement 114 individus au Canada et, pour chacun, on note la couleur des yeux et la couleur des cheveux. Les données sont illustrées dans le tableau 3. Notons ici qu'il y a deux variables catégoriques : la couleur des cheveux (deux niveaux : pâle, foncé) et la couleur des yeux (deux niveaux : bleu, brun). Nous reprendrons plus tard cet exemple.

Tableau 3 – Tableau de contingence selon la couleur des cheveux et des yeux.

	Pâle	Foncé
Bleu	38	11
Brun	14	51

3.1 Formulation des hypothèse nulles

La formulation des hypothèses nulles à partir d'un tableau de contingence diffère légèrement de ce que nous avons vu pour le test du khi carré d'ajustement. On peut formuler l'hypothèse nulle en termes d'indépendance d'une variable catégorique par rapport à une deuxième variable catégorique. L'hypothèse nulle peut aussi être formulée en termes de proportion d'un niveau d'une variable catégorique par rapport à une deuxième variable catégorique. Les deux types de formulation sont illustrés dans le prochain exemple.

Exemple 5.6 On continue l'exemple débuté plus haut. Nous pouvons formuler les hypothèses statistiques pour les données du tableau de contingence de deux manières différentes.

Formulation en termes d'indépendance :

H_0 (indépendance) : La couleur des yeux est indépendante de la couleur des cheveux.

H_a (non-indépendance) : La couleur des yeux n'est pas indépendante de la couleur des cheveux [3].

Formulation en termes de proportions :

H_0 : La proportion des individus avec des yeux bleus ne diffère pas selon la couleur des cheveux.

H_a : La proportion des individus avec des yeux bleus diffère selon la couleur des cheveux [4].

---

3. Attention : la non-indépendance n'implique pas la dépendance ou une relation de cause à effet. Il faut être prudent dans l'interprétation de la non-indépendance.
4. Nous verrons avec l'analyse de variance à deux critères que cette interprétation ressemble beaucoup au concept d'interaction.

3.2 Fréquences théoriques

L'obtention des fréquences théoriques est un peu plus compliquée pour les tableaux de contingence, mais suit la même logique que pour le χ^2 d'ajustement qui lui n'a qu'une seule variable catégorique. L'équation suivante nous donne les $\hat{f_{ij}}$ conformes à l'hypothèse nulle :

$$\hat{f_{ij}} = \frac{R_i \times C_j}{G} \; ,$$

où R_i correspond au total de la rangée i, C_j correspond au total de la colonne j et G est le grand total du tableau (somme de toutes les fréquences).

Exemple 5.7 On applique le calcul des fréquences théoriques aux données de l'échantillon de couleur des yeux et des cheveux. Le tableau 4 montre les fréquences observées avec les fréquences théoriques entre parenthèses. Pour illustrer, la fréquence théorique des individus avec des cheveux pâles et des yeux bleus s'obtient :

$$\begin{aligned} \hat{f_{ij}} &= \frac{R_i \times C_j}{G} \\ \hat{f_{ij}} &= \frac{49 \times 52}{114} \\ \hat{f_{ij}} &= 22.35 \end{aligned}$$

Avant d'aller plus loin, allons voir les détails du calcul du χ^2 d'indépendance.

Tableau 4 – Fréquences observées et théoriques (entre parenthèses) et totaux marginaux selon la couleur des cheveux et des yeux.

	Pâle	Foncé	
Bleu	38 (22.35)	11 (26.65)	49
Brun	14 (29.65)	51 (35.35)	65
	52	62	114

3.3 Calcul du χ^2 d'indépendance

Le calcul du χ^2 pour des données d'un tableau de contingence donne la somme des valeurs partielles pour toutes les cellules du tableau :

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(f_{ij} - \hat{f}_{ij})^2}{\hat{f}_{ij}}$$

où r correspond au nombre total de rangées et c au nombre total de colonnes du tableau de contingence, f_{ij} correspond à la fréquence observée pour la rangée i et la colonne j du tableau de contingence, alors que $\hat{f}_{ij}$ représente la fréquence théorique de la rangée i et colonne j du tableau de contingence. On calcule des degrés de liberté avec $df = (r-1)(c-1)$.

Exemple 5.8 En poursuivant l'exemple de la couleur des yeux et des cheveux, on peut calculer le χ^2 :

$$\begin{aligned}
\chi^2 &= \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(f_{ij} - \hat{f}_{ij})^2}{\hat{f}_{ij}} \\
\chi^2 &= \frac{(38-22.35)^2}{22.35} + \frac{(14-29.65)^2}{29.65} + \frac{(11-26.65)^2}{26.65} + \frac{(51-35.35)^2}{35.35} \\
\chi^2 &= 35.334
\end{aligned}$$

Avec cette valeur de χ^2, nous avons $(r-1)(c-1) = (2-1)(2-1) = 1$ degré de liberté. On peut aussi appliquer la correction de Yates pour la continuité :

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(|f_{ij} - \hat{f}_{ij}| - 0.5)^2}{\hat{f}_{ij}}$$

Ici, nous obtenons 33.112 avec la correction de Yates et une valeur de $P(\chi^2_{0.05,1} \geq$

33.112) < 0.0001. On rejette H_0 et on conclut que la couleur des cheveux n'est pas indépendante de la couleur des yeux chez des individus d'origine septentrionale résidant au Canada. À l'aide de R, on procède ainsi :

```
> ##on assemble les données dans une matrice
> freq.obs <- matrix(data = c(38, 11, 14, 51), nrow = 2,
                     ncol = 2, byrow = TRUE)
> ##on ajoute des étiquettes aux colonnes et rangées
> ##ceci est facultatif
> colnames(freq.obs) <- c("Pale", "Fonce")
> rownames(freq.obs) <- c("Bleu", "Brun")
> ##analyse sans correction pour continuité
> out.tab <- chisq.test(x = freq.obs, correct = FALSE)
> out.tab

        Pearson's Chi-squared test


data:  freq.obs
X-squared = 35.334, df = 1, p-value = 2.778e-09

> ##analyse avec correction pour continuité
> out.tab.corr <- chisq.test(x = freq.obs, correct = TRUE)
> out.tab.corr

        Pearson's Chi-squared test with Yates' continuity
        correction


data:  freq.obs
X-squared = 33.112, df = 1, p-value = 8.7e-09
```

Bien que `chisq.test( )` possède un argument `correct` qui prend la valeur `TRUE` par défaut, il n'applique la correction de continuité de Yates que sur les tableaux de contingence de 2 x 2. Nous n'aurions pas pu appliquer cette correction directement avec `chisq.test( )` sur le test du χ^2 d'ajustement pour les données de fleurs jaunes et vertes.

---

4 Alternatives au test du χ^2

Il existe des alternatives au test du khi carré, selon le type de fréquences récoltées. Les prochaines lignes mentionnent quelques tests à titre indicatif. Le test G (*G-test*) s'applique dans les mêmes conditions que le test du khi carré. Le test exact de Fisher (*Fisher's exact test*, `fisher.test( )`) peut s'avérer utile, lorsqu'on ne peut pas rencontrer les suppositions du test du χ^2 ou du test G, notamment, des fréquences théoriques $(\hat{f}_{ij}) < 5$. Le test exact de Fisher est basé sur la distribution hypergéométrique et s'applique surtout sur des tableaux de contingence de 2 x 2. Le test binomial (*binomial test*) est une autre alternative au test du χ^2 qui permet de tester si les proportions diffèrent dans un tableau 2 x 2 (`prop.test`) ou si une seule proportion diffère d'une valeur spécifique (`binom.test( )`). Si les catégories sont ordinales (p. ex., petit, moyen, grand ; jeune, mature, vieux), on peut utiliser le test de Kolmogorov-Smirnov modifié pour les données discrètes. Le cas échéant, le test de Kolmogorov-Smirnov sera préférable et plus puissant que le χ^2 qui ne tient pas compte de l'ordre des catégories. Toutefois, ce dernier n'était pas disponible dans `R` pour les données discrètes ordinales lors de l'écriture de ce document, c'est-à-dire, `ks.test( )` est un test d'ajustement uniquement pour les données continues.

Dans d'autres cas, on peut modéliser les fréquences directement à l'aide de modèles log-linéaires (*log-linear models*) ou des probabilités à l'aide de régressions logistiques (*logistic regression*) en fonction d'une série de variables catégoriques ou numériques. Ces dernières approches sont des modèles linéaires généralisés que l'on peut réaliser à l'aide de `glm( )`.

5 Le paradoxe de Simpson

Lorsqu'on analyse des fréquences, il faut être conscient du **paradoxe de Simpson** (*Simpson's paradox*), qui se traduit par une inversion des relations dues à l'effet d'une variable qui n'a pas été considérée dans l'analyse. Voici une illustration de ce paradoxe. Considérons la survie de souris exposées à un traitement comparé à un témoin (tableau 5). D'après ce tableau, on serait porté à conclure que la survie des souris est indépendante du traitement (60/120 vs 200/400 = 0.5 vs 0.5). Toutefois, en stratifiant l'analyse par sexe, on obtient un résultat différent. Le tableau des mâles suggère que le traitement augmente la survie (tableau 6, 14/60 vs 60/70 = 0.23 vs 0.85). Le tableau des femelles, quant à lui, suggère que le traitement diminue la survie (tableau 7, 46/60 vs 140/330 = 0.77 vs 0.42). Dans cet exemple, les analyses réalisées séparément par sexe mènent à des conclusions différentes. Dans ce cas-ci, faire abstraction du sexe dans l'analyse mène au paradoxe de Simpson et à une conclusion erronée. Il faut songer attentivement avant de se lancer dans l'analyse de fréquences et il est important de considérer toutes les variables pouvant influencer les résultats. Un moyen de prévenir le paradoxe de Simpson est de mieux cibler la population statistique d'intérêt en échantillonnant une partie de la population bien précise (p. ex., âge, type d'habitat). Une

Tableau 5 – Effet d'un traitement sur la survie d'individus.

	Témoin	Traitement
Vivant	60	200
Mort	60	200

Tableau 6 – Effet d'un traitement sur la survie des mâles.

	Témoin	Traitement
Vivant	14	60
Mort	46	10

Tableau 7 – Effet d'un traitement sur la survie des femelles.

	Témoin	Traitement
Vivant	46	140
Mort	14	190

autre approche consiste à stratifier l'analyse selon une ou plusieurs variables (p. ex., le sexe, la taille).

Conclusion

Cette leçon a présenté les rudiments de l'analyse de fréquences, particulièrement le test du χ^2 d'ajustement et d'indépendance. Les conditions sous-jacentes à cette analyse ont été énoncées, ainsi que quelques diagnostics et mesures préventives lors de l'élaboration de la stratégie d'échantillonnage. Des approches alternatives ont aussi été brièvement présentées lorsque les conditions du test du χ^2 ne peuvent être respectées. Finalement, le paradoxe de Simpson a été illustré avec un exemple afin de montrer que l'analyse de fréquences sans tenir compte d'une variable additionnelle peut influencer les conclusions.

Index