Tables de contingence, graphe en mosaïques, test du χ^2

Mame Diarra Fall
IDP- Université d'Orléeans

2022-2022

Chapitre 1

Introduction

1.1 Cas de deux caractères qualitatifs

On considère un couple de caractère qualitatifs (X, Y) ; À chaque individu l, on fait correspondre sa réponse (X_l, Y_l) avec $X_l \in E = \{x_1, \ldots, x_K\}$ et $Y_l \in F = \{y_1, \ldots, y_L\}$. On note n_{ij} l'effectif du couple (x_i, y_j) (nombre de couples (X_l, Y_l) égaux à (x_i, y_j)) :

$$n_{ij} = \sum_{l=1}^{n} \mathbb{I}_{\{X_l = x_i, Y_l = y_j\}}, \quad i = 1, \ldots, K, \quad j = 1, \ldots, L.$$

1.1.1 Table de contingence

Un tableau de contingence (ou tableau croisé) contient les fréquences d'association entre les modalités de deux caractères qualitatifs. Il s'agit d'une table $K \times L$ dans laquelle à l'intersection de la ligne i et de la colonne j figure f_{ij} (ou n_{ij}).

Exemple 1 : On dispose d'un échantillon de $n = 592$ élèves sur lesquels on relève la couleur des yeux, variable EYE à 4 modalités (Brown=brun, Blue=Bleu, Hazel=noisette et Green=vert) et celle des cheveux, variable HAIR à 4 modalités (Black=noir, Brown=brun, Red=roux et Blond=blond). Le tableau de contingence répartit l'effectif total (592) selon les croisements deux à deux des modalités des deux variables. Ces données sont déjà dans R. On y accède de la façon suivante :

```
> data(HairEyeColor)
```

Question 1 : Regrouper les deux modalités dans un tableau `tab`.
Indication : On pourra utiliser la fonction **apply**.

Modalités :

Les entêtes de ligne sont les *modalités* de la variable HAIR. Les entêtes de colonne sont les *modalités* de la variable EYE.

Visualisation : On peut commencer par visualiser les données du tableau de contingence (fréquences) :

```
> spineplot(tab)
```

Vocabulaire : On désigne par i l'indice du numéro de ligne, $i = 1, \ldots, K$ et par j celui du numéro de colonne, $j = 1, \ldots, L$.

- n_{ij} : effectif de la case (i, j) du tableau,
- $n_{i\bullet}$: effectif marginal de la ligne i du tableau, défini par $n_{i\bullet} = \sum_{j=1}^{L} n_{ij}$, où le $\bullet$ remplace l'indice pour lequel on somme.
- $n_{\bullet j}$: effectif marginal de la colonne j du tableau, défini par $n_{\bullet j} = \sum_{i=1}^{K} n_{ij}$,
- $n_{\bullet\bullet}$: effectif total du tableau, défini par $n_{\bullet\bullet} = \sum_{i=1}^{K} \sum_{j=1}^{L} n_{ij} = n$.

On rajoute les totaux des lignes et colonnes (**marges**) :

```
> addmargins(tab)
```

Fréquences et probabilités. Soit f_{ij} la fréquence du couple (x_i, y_j) :

$$f_{ij} = \frac{n_{ij}}{n}, \quad i = 1, \ldots, K, \quad j = 1, \ldots, L.$$

On vérifie que $\sum_{i=1}^{K} \sum_{j=1}^{L} f_{ij} = 1$, et donc le vecteur $(f_{ij}, 1 \leq i \leq K, 1 \leq j \leq L)$ s'interprète comme la **loi de probabilité (empirique)** du couple (X, Y) (loi **conjointe**).

À cette loi empirique, on associe la **loi marginale** empirique de X, donnée par les K fréquences $(f_{1\bullet}, \ldots, f_{K\bullet})$, et la loi marginale empirique de Y, déterminée par les L fréquences $(f_{\bullet 1}, \ldots, f_{\bullet L})$:

$$f_{i\bullet} = \sum_{j=1}^{L} f_{ij}, \quad f_{\bullet j} = \sum_{i=1}^{K} f_{ij}, \text{ avec } \sum_{i=1}^{K} f_{i\bullet} = \sum_{j=1}^{L} f_{\bullet j} = 1.$$

Pour avoir le tableau de contingence des fréquences à partir de celui des effectifs (tableau précédent), on divise chacune des cases du tableau précédent par $n = 592$. Puis, on calcule les distributions marginales (sommes).

Question 2 : Créer sous R un tableau `tabfreq` contenant le tableau des fréquences (arrondi à 3 décimales). La fonction **round** permet d'arrondir.
Afficher les marges de `tabfreq`.

1.1.2 Étude des profils

On peut aussi étudier les **distributions conditionnelles** empiriques, qu'en analyse de données on appelle les **profils-ligne** et **profils-colonnes**.

- Le $i^{\text{ème}}$**profil-ligne** est la répartition de Y lorsque X vaut x_i, qui s'interprète comme la loi de probabilité empirique $(Y|X = x_i)$. La probabilité conditionnelle empirique $\mathbb{P}(Y = y_j|X = x_i)$ se détermine en calculant les fréquences de réponses $\{Y = y_j\}$ parmi les individus ayant répondu $\{X = x_i\}$. Autrement dit, il s'agit de l'emploi de la formule de Bayes :

$$f_{j|i} = \mathbb{P}(Y = y_j|X = x_i) = \frac{\mathbb{P}(Y = y_j, X = x_i)}{\mathbb{P}(X = x_i)} = \frac{f_{ij}}{f_{i\bullet}} = \frac{\frac{n_{ij}}{n}}{\frac{n_{i\bullet}}{n}} = \frac{n_{ij}}{n_{i\bullet}}, \quad j = 1, \ldots, L.$$

  La fraction $\frac{f_{ij}}{f_{i\bullet}}$ représente la **fréquence conditionnelle** de la valeur y_j sachant $X = x_i$, notée $f_{j|i}$. Elle est donc donnée par la fréquence du couple (X_i, Y_j), divisée par la fréquence marginale de X.

- De façon analogue, on peut définir la fréquence conditionnelle de la valeur x_i sachant $Y = y_j$, notée $f_{i|j} = \frac{f_{ij}}{f_{\bullet j}}$:

$$f_{i|j} = \mathbb{P}(X = x_i | Y = y_j) = \frac{\mathbb{P}(Y = y_j, X = x_i)}{\mathbb{P}(Y = y_j)} = \frac{f_{ij}}{f_{\bullet j}} = \frac{\frac{n_{ij}}{n}}{\frac{n_{\bullet j}}{n}} = \frac{n_{ij}}{n_{\bullet j}}, \quad i = 1, \ldots, K.$$

  C'est le $j^{\text{ème}}$ **profil-colonne**.

Pour déterminer les profils-lignes et les profils-colonnes sous R, on utilise la commande **prop.table**.

Question 3 : créer deux tableaux `profLignes` et `profCols` contenant respectivement les profils-lignes et les profils-colonnes de `tab`.

On vérifie que les sommes en lignes de `profLignes` sont toutes égales à 1 (aux erreurs d'arrondi après) :

```
> rowSums(profLignes)
```

Idem pour les sommes en colonnes de `profCols` :

```
> colSums(profCols)
```

Question 4 : en vous basant sur ces tableaux, répondre aux questions suivantes :

1. Quelle est la probabilité qu'un élève blond, pris au hasard, ait les yeux bleus ?
2. Quelle est la probabilité qu'un élève aux yeux bleus, pris au hasard, soit blond ?

Liaison entre deux variables qualitatives

On reprend les mêmes notations que pour les tableaux de contingence. Soit donc n_{ij} l'effectif figurant à l'intersection de la ligne i et de la colonne j, $n_{i\bullet}$ et $n_{\bullet j}$ les effectifs marginaux, n l'effectif total. Si la variable X est **indépendante** de Y, alors

$$\mathbb{P}(X = x_i, Y = y_j) = \mathbb{P}(X = x_i)\mathbb{P}(Y = y_j).$$

Soit donc :

$$f_{ij} = f_{i\bullet} f_{\bullet j} = \frac{n_{i\bullet}}{n} \frac{n_{\bullet j}}{n}.$$

Cette situation correspond à des probabilités conditionnelles qui seraient égales quelque soit la strate (sous-population), c'est-à-dire des **profils identiques**. En effet, si X et Y sont *indépendantes* (la connaissance de l'une d'entre elles n'apporte rien à la connaissance de l'autre), alors pour chaque $1 \leq i \leq K$, $1 \leq j \leq L$, les relations suivantes sont simultanément vérifiées :

$$f_{i|j} = f_{i\bullet} \quad \text{et} \quad f_{j|i} = f_{\bullet j}.$$

(car $f_{i|j} := \frac{f_{ij}}{f_{\bullet j}} \overset{ind}{=} \frac{f_{i\bullet} f_{\bullet j}}{f_{\bullet j}} = f_{i\bullet}$).

En d'autres termes, les variables aléatoires X et Y sont indépendantes si et seulement si la distribution conditionnelle de Y sachant X (respectivement de X sachant Y) est égale à

la distribution marginale de Y (respectivement de X). Dans ce cas, on dit qu'il n y a **pas d'association** entre les deux variables.

L'**effectif théorique**, ou **effectif attendu**, d'une case, est l'effectif qui serait obtenu sous l'hypothèse d'indépendance. Il s'obtient en multipliant la fréquence théorique $\frac{n_{i\bullet}}{n}\frac{n_{\bullet j}}{n}$ (qui serait observée sous l'hypothèse d'indépendance, en conservant les effectifs marginaux observés) par l'effectif total n, soit

$$\frac{n_{i\bullet}n_{\bullet j}}{n}.$$

Exemple : l'effectif théorique d'avoir les yeux Bleus et les cheveux blonds, sous l'hypothèse d'indépendance, est donné par produit des effectifs des marges divisé par l'effectif total : soit $215 \times 127/592 = 46.12331$.

Pour obtenir les effectifs théoriques sous R :

```
> reschi2=chisq.test(tab) # résultat du test du chi2
> reschi2$expected
```

1.1.3 Représentation graphique

Le diagramme en mosaïque est une visualisation en 2D du tableau de contingence. Les effectifs du tableau sont représentés par des mosaïques dont la surface est proportionnelle à l'effectif de la cellule du tableau. Cette représentation en surface montre non seulement l'effectif mais la manière dont il se décompose en terme de produit.

La fonction **mosaicplot** permet de tracer ce diagramme :

```
>mosaicplot(tab, main="Diagramme en mosaique du tableau de contingence")
```

Pour l'interpréter, on s'intéresse maintenant aux effectifs qu'on devrait avoir si les deux variables étaient indépendantes.

```
par(mfrow=c(1,2))
mosaicplot(reschi2$expected, main="Situation d'independance", col="orange3")
mosaicplot(tab, main="Situation observee")
```

Avec cette représentation visuelle, les notions abstraites de dépendance/indépendance prennent uns sens concret.

1.2 Test d'indépendance du χ^2

On observe deux caractères qualitatifs X et Y et l'on souhaite savoir si l'effet constaté, via par exemple un histogramme des profils, de l'une des variables sur l'autre (dépendance de la répartition empirique de l'une des variables aux modalités de conditionnement de l'autre) est significatif ou pas.

On le formalise tel quel : soient $X \in \{x_1, \ldots, x_K\}$ et $Y \in \{y_1, \ldots, y_L\}$ deux v.a. finies. On observe n couples de "réponses" $((X_1, Y_1), \ldots, (X_n, Y_n))$ et on souhaite tester l'hypothèse

$$H_0 : "X \text{ et } Y \text{ sont indépendantes } " \text{ contre } H_1 "\text{c'est faux}".$$

La loi du couple (X, Y) est un $K \times L$ uplet que l'on note

$$p = (p_{ij},\, i = 1, \ldots, K,\, j = 1, \ldots, L), \quad p_{ij} = \mathbb{P}(X = x_i, Y = y_j),$$

et les lois marginales de X et Y sont les suites $(p_{i\bullet})_{1 \leq i \leq K}$ et $(p_{\bullet,j})_{1 \leq j \leq L}$ données par

$$p_{i\bullet} = \mathbb{P}(X = x_i) = \sum_{j=1}^{L} p_{ij}$$
$$p_{\bullet j} = \mathbb{P}(Y = y_j) = \sum_{i=1}^{K} p_{ij}.$$

Sous l'hypothèse nulle, la loi du couple est la loi p^0 produit des marginales :

$$p^0_{ij} = p_{i\bullet} p_{\bullet j}, \quad \forall i = 1, \ldots, K,\, j = 1, \ldots, L.$$

Intuitivement, on souhaite procéder comme pour un test paramétrique usuel (e.g., le test de Student), c'est-à-dire estimer la loi du couple par la loi empirique $(\hat{p})$:

$$\hat{p}_{ij} = \frac{n_{ij}}{n} = f_{ij},\, i = 1, \ldots, K; j = 1, \ldots, L$$

et calculer une distance entre $\hat{p}$ et la loi sous H_0. La "distance" adaptée entre deux probabilités discrètes sur le même espace est la **distance du** χ^2 (qui n'est pas mathématiquement une distance car non symétrique, on parle de **dissimilarité**). Si cette distance $\chi^2(p^0, \hat{p})$ est "trop grande", alors on rejette H_0. La différence avec la situation précédente vient de ce que la loi p^0 sous H_0 n'est pas entièrement connue : à la différence par exemple du test de Student, elle n'est pas totalement spécifiée par l'hypothèse nulle. Il faut donc l'estimer elle aussi. Il faut donc l'estimer elle aussi, ce que l'on fera en estimant les marginales empiriques :

$$\widehat{p^0}_{ij} = \frac{n_{i\bullet}}{n} \frac{n_{\bullet j}}{n} = f_{i\bullet} f_{\bullet j}.$$

Sous des conditions qui sont vérifiées dans le cas présent, on montre alors que, sous H_0, la loi limite de la statistique

$$n\chi^2(\widehat{p^0}, \hat{p}) = n \sum_{i=1}^{K} \sum_{j=1}^{L} \frac{(\widehat{p}^0_{ij} - f_{ij})^2}{\widehat{p}^0_{ij}} = \sum_{i=1}^{K} \sum_{j=1}^{L} \frac{(n\widehat{p}^0_{ij} - n_{ij})^2}{n\widehat{p}^0_{ij}}$$

tend vers une loi du chi-deux dont les degrés de liberté sont le nombre de paramètres estimés pour la loi empirique $(KL - 1)$ diminués du nombre de paramètres estimés sous H_0. Ici, on estime $K - 1$ paramètres $p_{i\bullet}$ et $L - 1$ paramètres $p_{\bullet,j}$ puisque chaque suite somme à 1, soit $KL - 1 - (K-1) - (L-1) = (K-1)(L-1)$. On admet donc le résultat asymptotique suivant :

Théorème 1 *Si l'hypothèse nulle d'indépendance est satisfaite, alors*

$$D^2 = n\chi^2(\widehat{p^0}, \hat{p}) \xrightarrow{\mathcal{L}} \chi^2_{((K-1)(L-1))} \quad \textit{lorsque } n \to \infty.$$

On dit que $n\chi^2(\widehat{p^0}, \hat{p})$ *converge en loi vers la loi limite ci-dessus.*

Proposition 1 (Test avec région de rejet.) *Le test de* H_0 : "X *et* Y *sont indépendantes* " *contre* H_1 *"c'est faux", de niveau voisin de* $0 < \alpha < 1$*, conduit au rejet de* H_0 *si*

$$D^2 = \{n\chi^2(\widehat{p^0}, \hat{p}) > \chi^2_{(K-1)(L-1),1-\alpha}\},$$

où $\chi^2_{(K-1)(L-1),1-\alpha}$ *est le quantile d'ordre* $(1-\alpha)$ *de la loi chi-deux à* $(K-1)(L-1)$ *degrés de liberté.*

L'application de cette proposition produit la décision "rejet" ou "non rejet" de H_0 au niveau α. Mais cette décision seule est imprécise : on ne sait pas si on a rejeté H_0 "largement" ou "de justesse". Les logiciels de statistique préfèrent donner le résultat d'un test sous la forme de la **p-valeur** ou **probabilité critique** du test, plus petit niveau qui permette de rejeter H_0 avec l'observation obtenue pour la statistique de test à partir des données $D^2 = n\chi^2(\widehat{p^0}, \hat{p}) = d^2$. L'expression mathématique de la p-valeur dépend du test (de la loi de la statistique de test sous H_0 et de la forme de sa région de rejet). Pour le test du χ^2 d'indépendance, la p-valeur est :

$$p = \mathbb{P}(D^2 > d^2) \quad \text{où } D^2 \sim \chi^2_{(K-1)(L-1)}.$$

Elle représente le risque de se tromper en rejetant H_0 à tort (i.e. en affirmant que X et Y ne sont pas indépendantes).

Exemple : on applique le test du χ^2 aux données précédentes. La valeur réalisée pour la statistique de test D^2 est $d^2 = 138.29$, avec D^2 qui suit approximativement une loi χ^2_9. Le seuil de rejet au niveau $\alpha = 1\%$ est ici $\chi^2_{9,0.99} = 21.66599$. On le lit dans une table, ou on l'obtient par exemple sous R avec la commande

```
> qchisq(0.99, 9)
[1] 21.66599
```

Décision : On a $d^2 = 138.29 > \chi^2_9 = 21.66599$ donc on rejette (fortement) l'hypothèse d'indépendance H_0 (avec une probabilité de 1% de se tromper).

Décision avec la p-valeur :
La p-valeur du test est : $p = \mathbb{P}(D^2 > 138.29)$, où $D^2 \sim \chi^2_9$, $p = 1 - \mathbb{P}(D^2 < 138.29)$ $= 1 - pchisq(138.29, 9) = 0$.
Pour avoir la valeur exacte de la p-valeur :

```
 > summary(as.table(tab))
   # Rq:  as.table: pour convertir en tab de contingence
Number of cases in table: 592
Number of factors: 2
Test for independence of all factors:
Chisq = 138.29, df = 9, p-value = 2.325e-25
```

On obtient $p = 2.325e-25$, ce qui signifie que si on rejette H_0, la probabilité de se tromper (rejet de H_0 à tort) est de $2.325e-25\%$, ce qui est très peu !
En général, on rejette H_0 (l'hypothèse d'indépendance) lorsque la p-value est jugée trop faible, classiquement inférieure à 5%. La p-value représente en quelque sorte le "degré d'attachement à H_0".