Explore data TD-TP : Analyse bivariée - Durée : 3h

Qualitatif $\times$ Qualitatif

Exercice 1. Niveau de diplôme en fonction du sexe

Le tableau suivant donne la répartition des salariés d'une entreprise selon le niveau de formation X et le sexe Y.

X | Y	F	H	Total: $n_{i\bullet}$
Bac+3	45	49	94
Bac+5	16	11	27
Bac+8	4	6	10
Total: $n_{\bullet j}$	65	66	131

1) Quels calculs est-il possible de faire avec ce type de tableau ?

2) Que représentent $n_{3\bullet}$ et $n_{\bullet 1}$? Que représente $f_{ij} = \frac{n_{ij}}{n_{\bullet\bullet}}$? Calculer f_{31} et interpréter les résultats.

3) Donner les distributions marginales des effectifs et des fréquences du caractère X (resp. Y)

4) Que représente $f_{i|j}$? Calculer $f_{3|1}$ et interpréter le résultat. Faire une phrase interprétant les fréquences suivantes : $f_{31}, f_{3|1}, f_{3\bullet}$ et $f_{\bullet 1}$.

5) Établir le tableau des profils lignes et des profils colonnes. Que peut-on en déduire ?

6) Comparer le tableau des effectifs théoriques avec celui des effectifs observés. Calculer la distance du chi-deux. Peut-on conclure sur l'indépendance de X et Y ?

Réponse :

1. Ce tableau croise deux caractères qualitatifs et les seuls calculs possibles qu'on peut faire sont les calculs des effectifs et /ou des fréquences. Les indicateurs numériques (moyenne, médiane, ...) sont mesurer selon une échelle de mesure numérique. Donc ce type calcule n'a pas de sens si (même si les modalités sont codées 0, 1, 2, ...)

2. $n_{3\bullet} = 10$: représente les effectifs des bac+8 tous sexes confondus.
$n_{\bullet 1} = 65$: représente les effectifs des femmes tous niveaux de formation confondus
f_{ij}: est la fréquence conjointe des modalités i et j.
$f_{31} = \frac{4}{131} = 0,03$: parmi les salariés il y a 3% de femmes ayant Bac+8.

3.
Distribution marginale des effectifs du caractère X:

X	$n_{i\bullet}$
x_1	94
x_2	27
x_3	10

Répartition des effectifs des salariés en fonction de leur niveau de formation

Distribution marginale des effectifs du caractère Y :

y	$n_{\bullet j}$
y_1	65
y_2	66

répartition des salariés de l'entreprise selon leur sexe.

4.

- $f_{i|j} = \frac{n_{ij}}{n_{\bullet j}}$:

représente la fréquence de la modalité i parmi la sous-population définie par la modalité j

- $f_{i\bullet}$ donne est la fréquence de la modalité i de X.

- $f_{i|j}$ et $f_{i\bullet}$ donnent une information sur le même phénomène mais dans deux populations différentes.

$f_{3\bullet} == \frac{10}{131} = 0,076$. Il y a $7,6\%$ des salariés qui ont un bac+8

$f_{\bullet 1} = 0,496$. Il y a $49,6\%$ de femmes dans l'entreprise

$f_{3|1} = 4/65 = 0.061 = 6,1\%$ parmi les femmes salariées ayant Bac+8.

Tableau des profils lignes- distribution conditionnelle: $f_{Y=j|X=i} = \frac{n_{ij}}{n_{i\bullet}}$

Il représente la distribution en proportion selon les lignes, c'est-à-dire par niveau de formation. A chaque ligen est associé son poids (qui donne l'importance de la ligne dans l'échantillon global).

Y| X	F	H	Total	Poids des profils lines: $\frac{n_{i\bullet}}{n} = f_{i\bullet}$
Bac+3	$\frac{45}{94} = 0,478$	$\frac{49}{94} = 0,5216$	1	$(94/131) * 100 \simeq 17.75\%$
Bac+5	$\frac{16}{27} = 0,592$	$\frac{11}{27} = 0,407$	1	$(27/131) * 100 \simeq 20.61$
Bac+8	$\frac{4}{10} = 0,4$	$\frac{6}{10} = 0,6$	1	$(10/131) * 100 \simeq 7.63$
Fre.Marg: $f_{\bullet j} = \frac{n_{\bullet j}}{n}$	$\frac{65}{131}$ =0,496	$\frac{66}{131} = 0,503$	1	

- On remarque le fort poids du niveau Bac+5 et Bac+3 (respectivement).

- Profil ligne moyen : $(f_{\bullet 1}, f_{\bullet 2}) = (0.496, 0.503)$. Il y a à peu près 50% de femmes dans l'entreprise.

- Ce pourcentage descend à 40% chez les bac+8 et augmente à 60% chez les bac+5. Cela laisse penser que la répartition des femmes dépend du niveau de formation.

Tableau des profils colonnes ou distribution conditionnelle de X: $f_{X=i|Y=j} = \frac{n_{ij}}{n_{\bullet j}}$

L représente la distribution selon les colonnes, c'est-à-dire par sexe.

$X\|Y$	F	H	Freq. Marg: $f_{i\bullet} = \frac{n_{i\bullet}}{n}$
Bac+3	0.69	0.74	0.71
Bac+5	0.24	0.16	0.206
Bac+8	0.061	0.09	0.07

Concernant les profils colonnes, il y a beaucoup moins d'écart entre la répartition totale des niveaux de formation et la répartition de sexe.

Tableau des fréquences théoriques

Effectifs théoriques			
	F	H	Total ($n_{i\bullet}$)
Bac+3	46.6412214	47.3587786	94
Bac+5	13.3969466	13.6030534	27
Bac+8	4.96183206	5.03816794	10
Total ($n_{\bullet j}$)	65	66	131

Fréquences observées			
	F	H	Total ($f_{i\bullet}$)
Bac+3	0.34351145	0.374045802	0.717557252
Bac+5	0.122137405	0.083969466	0.20610687
Bac+8	0.030534351	0.045801527	0.076335878
Total	($f_{\bullet j}$)	0.496183206	0.503816794 1

Tableau des différences	F	H
Bac+3	0.057751652	0.056876627
Bac+5	0.505778474	0.498115164
Bac+8	0.186447446	0.183622484

chi-deux	1.488591847
V	0.106598764

$k = 2$ et $s = 3$ donc $k \neq s$ donc coef van Cramer

Exercice 2.

On a interrogé 410 personnes de l'île et on leur a proposé deux questions :

- Dans quelle ville habitez-vous ?

- Quel type de programme regardez-vous le plus souvent à la télévision dans la liste suivante : actualités, reportages, filmes, séries, dessins animés, variétés, sports, jeux ?

Les réponses sont données dans le tableau suivant :

Ville	Actu	Report	Films	Séries	dess.anim	Variétés	Sport	Jeux
V1	10	6	18	10	3	1	11	4
V2	5	6	5	7	-	4	6	2
V3	2	2	3	4	1	4	5	6
V4	2	5	8	9	1	1	5	5
V5	8	2	12	7	1	6	7	1
V6	1	2	4	1	-	5	5	3
V7	2	1	3	3	1	4	2	4
V8	5	4	4	6	3	3	2	2
V9	6	2	13	8	-	-	6	2
V10	3	1	2	2	2	4	4	6
V11	8	4	6	5	1	3	6	1
V12	5	3	9	7	2	3	8	3

Nous souhaitons savoir si le type de programme regardé diffère selon le type de la ville. Nous regroupons certains modalités entre elles pour définir deux variables X et Y ayant les modalités suivantes :

- x_1: Capitale et banlieue regroupant V1, V5, V9 et V12.
- x_2: Villes moyennes regroupant V2, V4, V8 et V11.
- x_3: Petites villes regroupant V3, V6, V7 et V10.
- y_1: Regroupant actualités et reportages.
- y_2: Regroupant films, séries et dessins animés.
- y_3: Regroupant variétés, sport et jeux

1. Dresser le tableau de contingence croisant les variables X et Y. Quel est le croisement le plus fréquent ?

2. Dresser le tableau des effectifs théoriques sous l'indépendance puis celui des contingences simples. Quelle est la contingence simple la plus importante ?

3. On donne $\chi^2 = 26.45$. calculer la contribution de chacune des deux contingences mentionnées ci-dessus. Commenter.

Réponse.

1. Le tableau croisant X et Y est le suivant:

	y_1	y_2	y_3	Total
x_1	42	90	52	184
x_2	39	55	40	134
x_3	14	26	52	92
Total	95	171	144	410

Le croisement des modalités x_1 et y_2 est le cas le plus fréquent.

Le tableau correspondant à l'indépendance est le suivant :

	y_1	y_2	y_3
x_1	42,6	76,7	64,6
x_2	31,0	55,9	47,1
x_3	21,3	38,4	32,3

on obtient les contingences simples en formant les différences avec les données réelles:

	y_1	y_2	y_3
x_1	-0,6	13,3	-12,6
x_2	8	-0,9	-7,1
x_3	-7,3	-12,4	19,7

La contingence simple la plus importante est celle concernant les modalités x_3 (petites villes) et y_3 (divertissements) avec une différence de $+19,7$ entre l'effectif observé (52) et l'effectif théorique sous l'hyothèse d'indépendance (32,3).

3. La contribution de la case la plus fréquente est :

$$\frac{(n_{12}-n_{12}^*)^2}{n_{12}^*} = \frac{(90-76,7=13,3)^2}{76,7} \simeq 2,3, \qquad \text{Rappel:} \quad n_{ij}^* = \frac{n_{i\bullet}n_{\bullet j}}{n}.$$

Ainsi, cette contingence contribue pour

$$\frac{2,3}{\text{valeur théorique de } \chi^2} = \frac{2,3}{26,45} \simeq 9\%$$

aux écarts à l'indépendance. Pour la contingence $x_3 \times y_3$ la contibution est :

$$\frac{(n_{33}-n_{33}^*)^2}{n_{33}^*} = \frac{(19,7)^2}{32,3} \simeq 12.$$

ce qui correspond à $12/26,45 \simeq 45\%$ des écarts à l'indépendance constatés.
Ainsi, si l'on doit considérer que le type d'émission préférée dépend du lieu de résidence, cela est "dû" pour 45% (selon la mesure donnée par le χ^2) aux habitants des petites villes qui préfèrent les émissions de divertissements aux actualités et aux filmes.

A la première question, nous remarquions que l'association $x_1 \times y_2$ avait la plus forte fréquence. Par contre, notons qu'elle ne contribue à la valeur de χ^2 qu'à hauteur de 9% : par conséquent, il

n'y a *a priori* pas de lien évident entre l'importance de la fréquence d'une case et sa contribution relative aux écarts à l'indépendance.

Exercice 3. Durée du Chômage en fonction de l'âge et le sexe.

Étudier le fichier DureeChomageData.xls

Exercice 4. Couleur des cheveux de garçons et de fille

Cet exemple historique est dû à Fisher. Il étudie la couleur des cheveux de garçons et de filles d'un discret écossais.

	Blond	Roux	Châtain	Brun	Noir de jais
Garçon	592	119	849	504	36
Fille	544	97	677	451	14

Nous souhaitons savoir si la couleur des cheveux est indépendante du sexe (avec une erreur de 5%). Si la couleur dépend du sexe, nous aimerions avoir une idée des couleurs qui sont les plus dépendantes du sexe.

Étapes

1. Saisir les données
2. Visualiser les données
3. Calculer les profils lignes et les profiles colonnes
4. Construire le test du χ^2
5. Calculer les constructions au χ^2.

Réponse.

1. Saisir les données.

On saisit le jeu de données dans une matrice en lui affectant des nomes des lignes (rownames) et de colonnes (colnames).

- tab <-**matrix**(c(592,544,119,97,849,677,504,451,36,14), ncol=5)
- **rownames**(tab)<-c("Garçon","Fille")
- **colnames**(tab)<-c("Bland","Roux","Châtain","Brun", "Noir de jais")

2. Visualiser les données.

On peut visualiser les données à l'aide d'un diagramme en barres (voir figure 1). On choisit de représenter les données par sexe sur une même fenêtre graphique.

- Pour cela, on utilise la fonction **par** et l'attribut qui gère le nombre de graphiques en lignes et colonnes (mfrow=c(2,1): 2×1 graphiques en 2 lignes et 1 colonnes).

- Le vecteur couleur permet d'attribuer les noms de couleur à chaque catégorie:

- **par**(mfrow=c(2,1))
- couleur<-c("Gold","OrangeRed","Goldenrod","Brown","Black")
- barplot(tab[1,],main="Garçon",col=couleur)
- barplot(tab[2,],main="Fille",col=couleur)

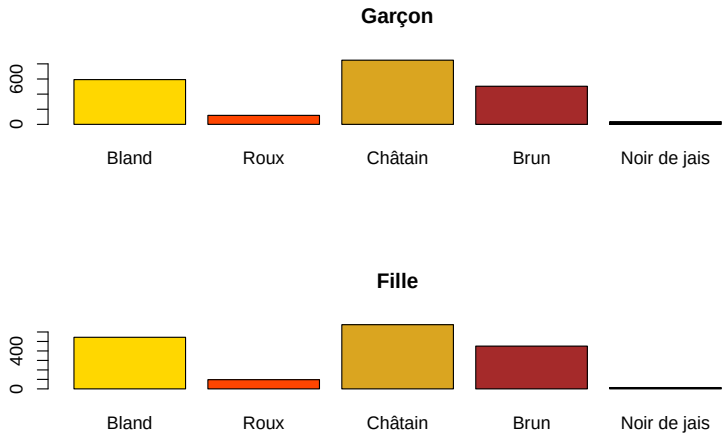

Figure 1: Distributions des couleurs de cheveux par sexe.

3. Calculer les profils lignes et les profils colonnes:

Calculons d'abord les fréquences conjointes $\frac{n_{ij}}{n}$, on présente les résultats en multipliant par 100 en arrondissant à la première décimale avec la fonction **round**

a) **round**(100× tab/sum(tab),1)

```
>  round(100*tab/sum(tab),1)
        Bland Roux Châtain Brun Noir de jais
Garçon   15.2  3.1    21.9 13.0          0.9
Fille    14.0  2.5    17.4 11.6          0.4
```

b) Calculons les profils lignes $\frac{n_{ij}}{\sum_j n_{ij}}$ avec la fonction **prop.table** (on obtient les distributions des couleurs de cheveux par sexe). On précise que l'on calcule les pourcentages lignes avec margin=1; pour ce faire, le jeu de données doit être une matrice, sinon on le transforme en une matrice par :

tab<-as.matrix(tab)

round(100×prop.table(tab, margin=1),1)

```
> round(100*prop.table(tab, margin=1),1)
       Bland Roux Châtain Brun Noir de jais
Garçon  28.2  5.7    40.4 24.0          1.7
Fille   30.5  5.4    38.0 25.3          0.8
```

Ces profils lignes sont représentés au graphique 1. Il y a peu de différence. La liaison entre les deux variables n'est pas flagrante, d'où l'intérêt de recourir à un test.
De même, on calcule les profils colonnes $\frac{n_{ij}}{\sum_i n_{ij}}$ en précisant que l'on travail cette fois sur les colonnes (margin=2)

round(100×prop.table(tab, margin=2), 1)

```
> round(100*prop.table(tab, margin=2), 1)
       Bland Roux Châtain Brun Noir de jais
Garçon  52.1 55.1    55.6 52.8           72
Fille   47.9 44.9    44.4 47.2           28
```

Au regard de ces profils colonnes, on remarque des différences notamment pour la couleur de noirs de jais: 72% des personnes ayant des cheveux noirs de jais sont des garçons.

4. Construction du test χ^2.

Pour effectuer le test d'indépendance entre les variables sexe et couleur de cheveux, nous calculons la valeur de statistique χ^2_{obs} et déterminons la probabilité critique.

resultat<-**chisq.test**(tab)

```
> resultat
        Pearson's Chi-squared test
data:  tab
X-squared = 10.4674, df = 4, p-value = 0.03325
```

La probabilité critique, p-value, indique qu'une valeur du χ^2_{obs} aussi grande n'aurait qu'une probabilité d'environ 3% d'être observée dans un échantillon de cette taille si la couleur des cheveux était indépendante du sexe.

Nous rejetons au seuil de 5% l'hypothèse d'indépendance et concluons qu'au vu de ces données la couleur des cheveux dépend du sexe.

5. Calculer les contributions au χ^2.

On peut étudier plus en détail cette liaisons en calculant les contributions

$$\frac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}}$$

à la statistique χ^2_{obs}. Les racines carrés de ces contributions sont dans l'objet residuals. En divisant chaque terme par le total (i.e.la valeur de χ^2_{obs} contenue dans la composante stat), on obtient un pourcentage.

round(100×resultat$residuals^2/resultat$stat,1)

```
> round(100*resultat$residuals^2/resultat$stat,1)
       Bland Roux Châtain Brun Noir de jais
Garçon   7.8  0.4     6.5  2.9         28.4
Fille    9.2  0.5     7.7  3.4         33.4
```

Les combinaisons qui contribuent le plus à la non-indépendance des deux variables sont celles concernant la couleur noir de jais. Un retour aux données, via les résidus et notamment leurs signes, permet de dire que le nombre de garçons ayant des cheveux noir de jais est plus important qu'attendu (sous-entendu: plus important que si l'hypothèse d'indépendance était vraie) et le nombre de filles est plus faible qu'attendu:

round(resultat$residuals, 3)

```
> round(resultat$residuals, 3)
        Bland   Roux Châtain   Brun Noir de jais
Garçon -0.903  0.202   0.825 -0.549        1.723
Fille   0.979 -0.219  -0.896  0.596       -1.870
```

Exercice 5. Propriétés

L'objectif de cette exercice est de démontrer quelques propriétés vues dans le cours.

1. L'égalité des profils lignes entraîne l'égalités des profils colonnes

On rappelle que:

- le profil de la i-ème ligne est:

  l'ensemble des fréquences des modalités de Y dans la modalité x_i de X

  c'est-à-dire les $\frac{n_{ij}}{n_{i\bullet}}$ pour j variant de 1 à s.

- le profil de la j-ème colonne est:

  l'ensemble des fréquences des modalités de X dans la modalité y_j de Y

  c'est-à-dire les $\frac{n_{ij}}{n_{\bullet j}}$ pour i variant de 1 à r.

a) Vérifier que l'égalité des profils lignes dans un tableau de contingence se traduit ainsi:

$$\text{pour tout } j \text{ fixé et pour tout } i \text{ variant de 1 à } r, \quad \text{on a : } \quad \frac{n_{ij}}{n_{i\bullet}} = \frac{n_{\bullet j}}{n}$$

b) En déduire que l'on a alors l'égalité des profils colonnes.

2. La somme des contingences simples est nulle

Vérifier sur une ligne ou une colonne quelconque que la somme des contingences simples $n_{ij} - n^*_{ij}$ est nulle.

3. $\chi^2_{\max} = n \times \min(r-1, s-1)$

a) En développant le carré dans l'expression de χ^2, montrer que l'on a:

$$\chi^2 = n \times \left(\sum_{i=1}^{r} \sum_{j=1}^{s} \frac{n_{ij}^2}{n_{i\bullet} n_{\bullet j}} \right) - n.$$

b) En remarquant qu'une fréquence est toujours inférieure ou égale à 1, montrer que l'on a:

$$\chi^2 \leq n(s-1) \qquad \text{et} \quad \chi^2 \leq n(r-1)$$

c) en déduire l'inégalité

$$\chi^2 \leq n \min(s-1, r-1)$$

d) on considère un tableau de contingence tel que, sur chaque ligne, une seule case contient des effectifs non nul. En reprenant l'écriture du χ^2 obtenue en 3 a), montrer qu'alors

$$\chi^2 = n(s-1)$$

Réponse.

Rappel. On suppose que X possède r modalités $x_1, \ldots, x_r$ et Y s modalités $y_1, \ldots, y_s$. On note

$$n_{i\bullet} = \sum_{j=1}^{s} n_{ij} \quad \text{l'effectif marginal de la i-ème modalité de } X$$

$$n_{\bullet j} = \sum_{i=1}^{r} n_{ij} \quad \text{l'effectif marginal de la j-ème modalité de } Y$$

$$n = \sum_{i=1}^{r} n_{i\bullet} = \sum_{j=1}^{s} n_{i\bullet} = \sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij} \quad \text{le nombre total d'individus dans la population.}$$

- Les distributions conditionnelles de Y, sous-entendu les distributions de Y conditionnée par le fait d'être dans telle ou telle modalité de X sont appelées "les profils lignes"

- La fréquence de la modalité y_j de Y sachant que l'on est dans la modalité x_i de X est définie par:

$$f_{Y=y_j|X=x_i} = \frac{n_{ij}}{n_{i\bullet}}$$

1. a) le profil de la i-ème ligne est formé de :

$$\frac{n_{ij}}{n_{i\bullet}} \quad \forall\ 1 \leq j \leq s.$$

On a donc égalité des profils lignes si dans chaque ligne ainsi considérée, on retrouve les fréquences $\frac{n_{\bullet j}}{n}$ des modalités de Y observées dans l'ensemble de la population. Ainsi, on aura égalité des profils lignes si quelle que soit la colonne j fixé, on a, pour tout les i:

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

b) Le le profil de la j-ème colonne est formé de :

$$\frac{n_{ij}}{n_{\bullet j}} \quad \forall\ 1 \leq i \leq r.$$

or nous avons l'équivalence suivante :

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

ainsi, le i-ème élément du profil de la j-ème colonne ne dépend pas de j puisqu'il est égal à la fréquence de la i-ème modalité de X: $\frac{n_{i\bullet}}{n}$. Les profils colonnes sont par conséquent égaux.

2.

a) Considérons par exemple une ligne quelconque i. Les contingences simples sont les $n_{ij} - n^*_{ij}$ pour j variant de 1 à s. Il vient donc

$$\begin{aligned}
\sum_{j=1}^{s}(n_{ij} - n^*_{ij}) &= \sum_{j=1}^{s}(n_{ij} - \frac{n_{i\bullet}n_{\bullet j}}{n}) \\
&= \sum_{j=1}^{s} n_{ij} - \sum_{j=1}^{s} \frac{n_{i\bullet}n_{\bullet j}}{n} \\
&= n_{i\bullet} - n_{i\bullet} \sum_{j=1}^{s} \frac{n_{\bullet j}}{n} \\
&= n_{i\bullet} - n_{i\bullet} = 0.
\end{aligned}$$

3.

a) On développe le carré et on remplace dans l'expression du χ^2, on obtient :

$$\begin{aligned}
\chi^2 &= \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij}-n^*_{ij})^2}{n^*_{ij}} \\
&= \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n^2_{ij}+{n^*_{ij}}^2-2n_{ij}n^*_{ij})}{n^*_{ij}} \\
&= \sum_{i=1}^{r}\sum_{j=1}^{s}\left[\frac{n^2_{ij}}{n^*_{ij}}+n^*_{ij}-2n_{ij}\right] \\
&= \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n^2_{ij}}{n^*_{ij}}+\sum_{i=1}^{r}\sum_{j=1}^{s}n^*_{ij}-2\sum_{i=1}^{r}\sum_{j=1}^{s}n_{ij}
\end{aligned}$$

La dernière somme est la somme de tous les effectifs conjoints observés, elle est égale donc à n. La première peut s'écrire (en remplaçant n^*_{ij} par $\frac{n_{i\bullet}n_{\bullet j}}{n}$)

$$\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n^2_{ij}}{n^*_{ij}} = n\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n^2_{ij}}{n_{i\bullet}n_{\bullet j}}$$

La seconde correspond à la somme des effectifs conjoints théoriques. C'est encore l'effectif total n.

on en déduit donc l'expression demandée.

b) Remarquons d'abord que :

$$\frac{n^2_{ij}}{n_{i\bullet}n_{\bullet j}} = \frac{n_{ij}}{n_{i\bullet}}\times\frac{n_{ij}}{n_{\bullet j}}$$

Comme $\frac{n_{ij}}{n_{i\bullet}} \leq 1$, en utilisant le résultat de la question précédente, on obtient :

$$\chi^2 \leq n\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}}{n_{\bullet j}} - n.$$

Par ailleurs

$$\begin{aligned}
\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}}{n_{\bullet j}} &= \sum_{j=1}^{s}\sum_{i=1}^{r}\frac{n_{ij}}{n_{\bullet j}} \\
&= \sum_{j=1}^{s}\frac{\sum_{i=1}^{r}n_{ij}}{n_{\bullet j}} \\
&= \sum_{j=1}^{s}\frac{n_{\bullet j}}{n_{\bullet j}} \\
&= s
\end{aligned}$$

Ainsi, $\chi^2 \leq n(s-1)$ et on obtient de la même façon $\chi^2 \leq n(r-1)$. par conséquent,

$$\chi^2 \leq n\min(s-1, r-1).$$

c) Si un tableau n'a qu'une seule case non vide par ligne alors :

pour tout i, il existe un unique j (notons le j_i, pour mentionner qu'il dépend de la ligne i) tel que $n_{ij_i}/n_{i\bullet} = 1$ et $n_{ij} = 0$ pour tout $j \neq j_i$.

En reprenant l'écriture obtenue en début de question 3, il vient :

$$\begin{aligned}
\chi^2 &= n \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{n_{ij}^2}{n_{i\bullet} n_{\bullet j}} - n \\
&= n \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{n_{ij_i}^2}{n_{i\bullet} n_{\bullet j}} - n \\
&= n \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{n_{ij_i}}{n_{i\bullet}} \times \frac{n_{ij_i}}{n_{\bullet j}} - n \\
&= n \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{n_{ij_i}}{n_{\bullet j}} - n \\
&\quad \text{on intervertir les indices de sommation} \\
&= n \sum_{j=1}^{s} \sum_{i=1}^{r} \frac{n_{ij_i}}{n_{\bullet j}} - n \\
&= n \sum_{j=1}^{s} 1 \times \frac{n_{\bullet j}}{n_{\bullet j}} - n \\
&= n(s-1)
\end{aligned}$$