# Tout ce que vous n'avez jamais voulu savoir sur le $\chi^2$ sans jamais avoir eu envie de le demander

Julien Barnier
Groupe de Recherche sur la Socialisation
CNRS – UMR 5040
jbarnier@ens-lsh.fr

15 avril 2008

# Table des matières

# Partie 1

# Introduction

## 1.1 À propos de ce document

Ce document a pour ambition d'essayer de présenter les principes du test statistique dit « test du $\chi^2$ », autant que possible de manière pas trop rébarbative.

On insistera très peu sur le mode de calcul effectif (tous les logiciels de statistiques actuels s'en chargent bien mieux que nous) et beaucoup plus sur les concepts sur lesquels le test repose.

La version de référence de ce document se situe à l'adresse :

http://perso.ens-lsh.fr/jbarnier/pdf/khi2.pdf

Le code source LaTeX est également téléchargeable (sans les illustrations) à l'adresse :

http://perso.ens-lsh.fr/jbarnier/pdf/khi2.tex



## 1.2 Mode d'emploi

À l'image de son titre, ce document est long. Très long. Trop long.

La lecture intégrale de ce document pourrait avoir des conséquences en termes d'équilibre psychique et d'exacerbation de sentiments agressifs à l'égard de son prochain que nous ne saurions évaluer de manière parfaitement rigoureuse. Le principe de précaution nous dicte donc de prévoir des modes de lecture alternatifs.

Voici donc un plan rapide de ce qui suit afin que ceux qui le souhaitent n'aient pas à supporter la lecture de l'ensemble :

- la partie 2 présente l'hypothèse d'indépendance, qui est au cœur du test du $\chi^2$. La partie 3 présente la manière dont cette hypothèse d'indépendance se traduit par le calcul d'un tableau d'effectifs théoriques ;
- la partie 4 présente les différentes étapes de calcul du $\chi^2$ d'un tableau et les résultats qu'on peut en tirer ;
- la partie 5 se penche sur l'interprétation qui peut être faite des résultats du $\chi^2$, et notamment sur les facteurs qui influencent la valeur du test ;
- la partie 6 aborde les limites liées au test et qu'il faut prendre en compte dans l'interprétation ;
- la partie 7 indique des subtilités ou des compléments au test. Elle peut être joyeusement ignorée en cas de première lecture.

Enfin, la partie 8 se veut un récapitulatif des différents points importants à retenir. Chacun d'entre eux est accompagné du numéro de la page correspondant si on souhaite un peu plus de détail. Cette partie peut être utilisée comme « porte d'entrée » pour le reste du document si on ne souhaite pas une lecture linéaire intégrale.

## 1.3 Le test du quoi ?

Première interrogation : comment ça se prononce ?

Le $\chi$ n'est pas un X mais bien une lettre grecque dont le petit nom est *khi*, lequel se prononce *« qui »*. Et le $^2$, qui pourrait se prononcer *« au carré »*, se prononce plutôt tout simplement *« deux »*.

Moralité, si vous souhaitez briller dans un congrès international de statistiques, dites *« test du qui-deux »* plutôt que *« test du x-au-carré »* [1].

## 1.4 Et sinon, ça sert à quoi ?

En une phrase, le test du $\chi^2$ permet de déterminer la probabilité que les lignes et les colonnes d'un tableau croisé sont indépendantes [2].

Dit autrement, il permet d'évaluer si la répartition des effectifs dans une table de contingence est significativement différente de celle de la table calculée sous l'hypothèse d'indépendance des deux variables croisées.

Comme tout cela est absolument incompréhensible, nous allons commencer par définir les concepts de base, et en premier lieu le terme d'*indépendance*.

1. Quoi que l'expression *« qui-carré »* semble également tout à fait acceptable, d'autant que la version anglaise est *« chi squared »*.
2. Note pour les puristes : nous n'abordons dans ce document que le test du $\chi^2$ de contingence, c'est-à-dire celui qui teste l'indépendance des lignes et des colonnes d'un tableau croisé. On ne parlera pas des autres applications de la statistique du $\chi^2$, notamment pour tester l'adéquation à une loi ou à une répartition donnée.

# Partie 2

# L'hypothèse d'indépendance

## 2.1 Petits rappels

Une variable qualitative est une variable qui mesure une donnée pouvant être découpée en un nombre restreint de modalités, par exemple :

- le genre de l'enquêté : homme, femme ;
- la couleur de son arrosoir : vert, rouge, bleu, noir… ;
- son âge en classes de cinq ans : 21-25 ans, 26-30 ans, 31-35 ans… ;
- le dernier livre qu'il a lu : *Tractatus logico-philosophicus*, *Oui-oui et la voiture jaune*…

Une table de contingence, ou tableau croisé, est un tableau qui indique les effectifs du croisement entre deux variables qualitatives.

Un petit exemple, croisant l'âge et le dernier livre lu par la personne interrogée :

| | 0 à 10 ans | 11 à 70 ans | 71 ans et plus |
|---|---|---|---|
| *Tractatus Logico-philosophicus* | 1 | 15 | 2 |
| *Oui-oui et la voiture jaune* | 854 | 2 | 621 |

Sur ce genre de tableau, on peut regarder quelle est la répartition âges des lecteurs de chaque ouvrage. Pour cela on calcule les *pourcentages en ligne*, c'est à dire qu'on divise les effectifs de chaque case par l'effectif total de la ligne du tableau à laquelle elle appartient. Ce qui nous donne ici :

| | 0 à 10 ans | 11 à 70 ans | 71 ans et plus | *Total* |
|---|---|---|---|---|
| *Tractatus Logico-philosophicus* | 5,6 % | 83,3 % | 11,1 % | *100 %* |
| *Oui-oui et la voiture jaune* | 57,8 % | 0,1 % | 42,0 % | *100 %* |

La lecture de ce tableau donnerait *« 5,6 % de ceux dont le dernier livre lu est le Tractatus Logico-philosophicus ont entre 0 et 10 ans »*.

On peut aussi regarder la répartition de la lecture des livres en fonction de l'âge. Dans ce cas on calcule les *pourcentages colonnes*, c'est à dire qu'on divise les effectifs de chaque case par l'effectif total de la ligne du tableau à laquelle elle appartient. Ce qui nous donne ici :

| | 0 à 10 ans | 11 à 70 ans | 71 ans et plus |
|---|---|---|---|
| *Tractatus Logico-philosophicus* | 0,1 % | 88,2 % | 0,3 % |
| *Oui-oui et la voiture jaune* | 99,9 % | 11,8 % | 99,7 % |
| *Total* | *100 %* | *100 %* | *100 %* |

Ce qui pourrait se lire : *« 11,8 % des 11 à 70 ans ont lu comme dernier livre Oui-oui et la voiture jaune »*.

Plutôt que de « pourcentages lignes » et de « pourcentages colonnes », on parle également parfois de « profils lignes » et « profils colonnes ».

## 2.2 L'indépendance des lignes et des colonnes

L'objectif du test du $\chi^2$ est de déterminer si les lignes et les colonnes d'un tableau croisé (c'est à dire les deux variables étudiées) sont indépendantes. Par indépendantes, on veut dire que le fait d'appartenir à une modalité de la première variable n'a pas d'influence sur la modalité d'appartenance de la deuxième variable.

Prenons tout de suite un petit exemple avec les deux tableaux suivants, qui croisent le genre et le plat préféré :

| | Homme | Femme |
|---|---|---|
| Choucroute garnie | 10 | 10 |
| Brocolis vapeur | 10 | 10 |

| | Homme | Femme |
|---|---|---|
| Choucroute garnie | 0 | 20 |
| Brocolis vapeur | 20 | 0 |

Dans le tableau de gauche, les effectifs se répartissent de manière totalement uniforme : le fait d'être un homme ou une femme ne semble avoir aucune influence sur le plat préféré. On ne peut donc pas parler d'un lien entre les deux variables : elles sont indépendantes.

Dans le tableau de droite, inversement, on constate que le fait d'être un homme ou une femme conditionne totalement le fait de préférer la choucroute ou les brocolis. On a donc un lien extrêmement fort entre les deux variables : elles ne sont absolument pas indépendantes.

Ces deux tableaux présentent cependant une version quelque peu radicale de l'indépendance [1]. Pour obtenir quelque chose d'un peu moins caricatural, on peut repartir de la définition donnée plus haut en la reformulant : dire que les lignes et les colonnes d'un tableau sont indépendantes, c'est dire que la modalité d'appartenance en colonne n'a pas d'influence sur la modalité d'appartenance en ligne.

Ceci signifie donc que la répartition des effectifs du tableau entre les différentes lignes est la même quelle que soit la colonne. Dit autrement, cela signifie que les pourcentages colonnes du tableau sont identiques pour toutes les colonnes.

On comprendra sans doute mieux en regardant le tableau suivant :

| | Homme | Femme |
|---|---|---|
| Choucroute garnie | 20 % | 20 % |
| Brocolis vapeur | 80 % | 80 % |
| *Total* | 100 % | 100 % |

Avec une telle répartition il est assez naturel d'en déduire que la préférence culinaire est indépendante du sexe.

1. Si nous osions, nous parlerions même de vision à tendance indépendantiste.

Comme les lignes et colonnes d'un tableau sont parfaitement interchangeables, le raisonnement vaut aussi dans l'autre sens, c'est à dire que l'indépendance entre les lignes et les colonnes d'un tableau croisé signifie que les pourcentages lignes de ce tableau sont les mêmes pour toutes les lignes.

## 2.3 En résumé

Il n'y a qu'une seule chose à retenir : dire que les variables d'un tableau croisé sont indépendantes revient à dire les trois choses suivantes.

1. le fait d'appartenir à l'une des modalités de la première variable n'a aucune influence sur la modalité d'appartenance de la seconde ;
2. les pourcentages lignes du tableau croisé sont les mêmes pour toutes les lignes ;
3. les pourcentages colonnes du tableau croisé sont les mêmes pour toutes les colonnes.

# Partie 3

# Calculer l'indépendance

## 3.1 Le biais d'échantillonnage

Les exemples précédents utilisés pour illustrer ce qu'est l'hypothèse d'indépendance restent théoriques. En effet, nous ne rencontrerons jamais lors du traitement d'une vraie enquête des tableaux où les pourcentages lignes et colonnes sont tous *exactement* les mêmes et où les deux variables croisées sont *parfaitement* indépendantes :

- d'une part car un lien entre deux variables ne se traduit jamais en sciences sociales par du « tout ou rien ». On pourra toujours trouver une personne sans diplôme grande lectrice de Proust ou un spécialiste en droit constitutionnel collectionneur de nains de jardins ;
- d'autre part car les résultats obtenus sont en partie liés aux personnes interrogées. On nomme ce type de variations *biais d'échantillonnage*.

Pour mieux comprendre ce qu'est ce biais, reprenons notre exemple gastronomique précédent. Imaginons que nous avons une population de 1000 personnes, 500 hommes et 500 femmes. On sait par ailleurs d'une part que le sexe n'a aucune influence sur le fait de préférer les brocolis ou la choucroute, et d'autre part qu'il y a autant de personnes qui apprécient les deux plats. Si nous interrogeons tout le monde, nous obtenons donc le tableau suivant :

| | Homme | Femme |
|---|---|---|
| Choucroute | 250 | 250 |
| Brocolis | 250 | 250 |

Seulement voilà, interroger tout le monde prend du temps et coûte des sous. On choisit donc en général de n'interroger qu'une partie des gens, disons 100 personnes. Si on « choisit » ces 100 personnes de manière totalement aléatoire, on peut s'attendre à trouver le tableau suivant :

| | Homme | Femme |
|---|---|---|
| Choucroute | 25 | 25 |
| Brocolis | 25 | 25 |

Mais en pratique, il suffit que Charles-Emmanuel, qui était malade parce qu'il avait mangé trop de brocolis, ne puisse pas répondre au questionnaire et qu'il soit remplacé au pied levé par Jean-Kevin qui est un fan de choucroute pour que vous obteniez le résultat suivant :

| | Homme | Femme |
|---|---|---|
| Choucroute | 26 | 25 |
| Brocolis | 24 | 25 |

Et en pratique, vous risquez surtout d'obtenir quelque chose qui va ressembler à l'un des tableaux suivants :

| | Homme | Femme |
|---|---|---|
| Choucroute garnie | 27 | 26 |
| Brocolis vapeur | 23 | 24 |

| | Homme | Femme |
|---|---|---|
| Choucroute garnie | 28 | 22 |
| Brocolis vapeur | 24 | 26 |

La question qui se pose, dès lors, est de savoir à partir de quand on peut dire que les variations observées sont dues au hasard, et à partir de quand on peut estimer qu'elles sont dues à un lien entre les deux variables. C'est tout l'objet du test du $\chi^2$.

Mais avant d'en arriver là nous devons regarder d'un peu plus près ce que signifie l'indépendance entre deux variables qualitatives dans un tableau croisé.

## 3.2 Contraintes sur les marges du tableau

Imaginons maintenant un nouvel exemple. À partir d'une population de 120 personnes, nous souhaitons étudier le lien entre la couleur des cheveux (bruns, blonds, roux) et la couleur des nœils (marrons ou bleus) [1]. La question posée est de savoir à quoi ressemblerait notre tableau dans le cas où couleur des cheveux et couleur des nœils seraient totalement indépendants [2].

Intuitivement, et c'est ce que nous avons fait jusque ici, on pense au tableau théorique suivant :

| | Bruns | Blonds | Roux |
|---|---|---|---|
| Marrons | 20 | 20 | 20 |
| Bleus | 20 | 20 | 20 |

Tab. 3.1 – Tableau des effectifs théoriques (faux)

Même effectif dans toutes les cases, et effectif total de 120 correspondant à notre population. Comment pourrait-on trouver une plus belle marque d'indépendance ?

Certes. Mais cette répartition théorique s'appuie sur une hypothèse très forte : *elle suppose d'une part qu'il y a autant de bruns, de blonds et de roux dans notre population, et d'autre part qu'il y a autant de personnes aux yeux marrons que de personnes aux yeux bleus.* Or cette hypothèse est très probablement fausse. Imaginons que notre étude se passe en Suède. On observerait alors dans notre population de 120 personnes les répartitions de couleurs des cheveux et des nœils suivantes :

| Bruns | Blonds | Roux | Total |
|---|---|---|---|
| 12 | 90 | 18 | 120 |

| Marrons | Bleus | Total |
|---|---|---|
| 30 | 90 | 120 |

Tab. 3.2 – Répartition des couleurs des cheveux et des nœils dans la population

Rajoutons maintenant à notre tableau 3.1 les totaux en ligne et en colonnes :

1. Les données qui suivent sont totalement imaginaires et fantaisistes, mais vous l'aurez sans doute déjà deviné...
2. Dans ce qui suit, on nommera ce tableau sous hypothèse d'indépendance *tableau théorique*, mais il faudrait en fait lire *tableau de répartition théorique sous l'hypothèse d'indépendance des lignes et des colonnes.*

| | Bruns | Blonds | Roux | *Total* |
|---|---|---|---|---|
| Marrons | 20 | 20 | 20 | *60* |
| Bleus | 20 | 20 | 20 | *60* |
| *Total* | *40* | *40* | *40* | *120* |

TAB. 3.3 – Tableau des effectifs théoriques (toujours faux)

On voit tout de suite que quelque chose ne colle pas : si on a bien 120 personnes en tout, on a 60 personnes aux yeux marrons et 60 aux yeux bleus, alors que notre population en compte respectivement 30 et 90. Même chose pour la couleur des cheveux. Cette répartition avec 20 personnes dans chaque case est donc tout simplement impossible.

Petit point de vocabulaire : on appelle les totaux en lignes et en colonnes du tableau 3.3 les *marges* du tableau croisé. Et on nomme les répartitions des couleurs des cheveux et des nœils indiquées tableau 3.2 les *tris à plat* de ces variables.

En un mot, on vient de rajouter une contrainte forte sur notre tableau théorique de répartition sous l'hypothèse d'indépendance : les marges de ce tableau doivent correspondre aux tris à plat des variables correspondantes dans notre population. Dans ce qui suit, on nommera cette contrainte *contrainte sur les marges* du tableau de répartition théorique.

## 3.3 Calculs des effectifs théoriques

Bon, c'est bien gentil tout ça, de nous rajouter des contraintes supplémentaires, mais concrètement, il va ressembler à quoi notre tableau théorique ?

Pour comprendre, nous allons d'abord transformer la répartition des différentes couleurs de cheveux et de nœils du tableau 3.2 en pourcentages, ce qui donne le résultat suivant :

| Bruns | Blonds | Roux | Total |
|---|---|---|---|
| 10 % | 75 % | 15 % | 100 % |

| Marrons | Bleus | Total |
|---|---|---|
| 25 % | 75 % | 100 % |

TAB. 3.4 – Répartition des couleurs des cheveux et des nœils dans la population, en pourcentages

**Avertissement** les trois paragraphes qui suivent peuvent être un peu pénibles à comprendre. Si la lecture des précédentes sections vous a déjà plongé dans un état de léthargie avancé, il est temps d'aller prendre un café ou un jus de carottes. Sinon, n'hésitez pas à relire plusieurs fois les passages incompréhensibles.

On se pose la question suivante : sachant que dans une population nous avons 10 % de bruns et 25 % de personnes aux yeux marrons, sous l'hypothèse d'indépendance des couleurs de cheveux et de nœils, quelle proportion d'individus devrait avoir les cheveux bruns *et* les yeux marrons ?

Pour répondre à cette question, on peut penser au fait que *l'hypothèse d'indépendance signifie que la proportion de personnes aux yeux marrons est la même quelle que soit la couleur des cheveux.* Elle est donc de 25 % pour les personnes ayant les cheveux bruns. Cela signifie qu'un quart des 10 % de personnes aux cheveux bruns ont les yeux marrons, ou encore que 2,5 % [3] de la population totale a à la fois les cheveux bruns et les yeux marrons.

3. 2,5 étant un quart de 10.

**Pourcentages théoriques** De manière générale, la règle est la suivante : le pourcentage théorique, sous l'hypothèse d'indépendance, des individus ayant la couleur de cheveux $x$ et la couleur des nœils $y$ est égal au produit entre le pourcentage d'individus ayant la couleur de cheveux $x$ et le pourcentage d'individus ayant la couleur des nœils $y$.

Pour reprendre un exemple, sachant qu'on a 75 % de blonds et 25 % de personnes aux yeux bleus, la proportion de personnes blondes aux yeux bleus dans notre population totale sous l'hypothèse d'indépendance vaut :

$$\frac{75}{100} \times \frac{25}{100} = \frac{18,75}{100}, \text{ soit } 18,75\%$$

Avec cette règle on peut désormais calculer le tableau des pourcentages théoriques sous l'hypothèse d'indépendance :

| | Bruns | Blonds | Roux | *Total* |
|---|---|---|---|---|
| Marrons | 2,5 % | 18,75 % | 3,75 % | *25 %* |
| Bleus | 7,5 % | 56,25 % | 11,25 % | *75 %* |
| *Total* | *10 %* | *75 %* | *15 %* | *100 %* |

Tab. 3.5 – Tableau des pourcentages théoriques (exacts)

Et maintenant que nous avons nos pourcentages théoriques, il est très facile de passer aux effectifs : il suffit de multiplier, dans chaque case, le pourcentage théorique par l'effectif total du tableau. Ainsi, pour les bruns aux yeux marrons, on obtient un effectif théorique de $2,5\% \times 120$, c'est à dire 3 personnes. On fait de même pour toutes les cases du tableau et on obtient :

| | Bruns | Blonds | Roux | *Total* |
|---|---|---|---|---|
| Marrons | 3 | 22,5 | 4,5 | *30* |
| Bleus | 9 | 67,5 | 13,5 | *90* |
| *Total* | *12* | *90* | *18* | *120* |

Tab. 3.6 – Tableau des effectifs théoriques (exacts)

Petite surprise : le tableau contient des nombres à virgule ! En effet, comme il s'agit d'effectifs *théoriques*, il ne s'agit pas forcément de nombres entiers.

Par contre, on remarquera que les marges de notre tableau correspondent bien aux tris à plat de nos variables indiquées tableau 3.2, ce qui est plutôt rassurant puisque c'est quand même pour ça que nous avons souffert depuis quelques pages.

## 3.4 En résumé

Pour faire notre test du $\chi^2$, nous avons besoin de déterminer à quoi ressemblerait notre tableau si les deux variables croisées étaient totalement indépendantes. Le calcul de ce tableau s'effectue en deux temps :

1. on calcule le tableau des pourcentages théoriques, en multipliant pour chaque case la proportion observée dans la population des deux modalités correspondantes ;
2. puis, le tableau des effectifs théoriques se calcule en multipliant le tableau des pourcentages théoriques par l'effectif total.

En pratique, il est important de comprendre le principe, et notamment l'existence de la contrainte sur les marges. Le mode de calcul importe peu puisqu'il sera toujours réalisé par un logiciel dédié.

# Partie 4

# Calcul du $\chi^2$ d'un tableau

## 4.1 Observons les écarts

Prenons maintenant un autre exemple, toujours plus passionnant. Lors d'une enquête à grande échelle réalisée en partenariat avec l'INSEE, l'INED et l'INSERM, on a demandé à 200 personnes leur profession et on a croisé cette information avec une variable indiquant s'ils possèdent ou non une brouette. Le résultat est le suivant :

| | Sociologue | Banquier | Archéologue | *Total* |
|---|---|---|---|---|
| Avec brouette | 37 | 36 | 12 | *85* |
| Sans brouette | 65 | 43 | 7 | *115* |
| *Total* | *102* | *79* | *19* | *200* |

Tab. 4.1 – Effectifs observés

Nous savons désormais calculer le tableau des pourcentages théoriques sous l'hypothèse d'indépendance entre les deux variables :

| | Sociologue | Banquier | Archéologue | *Total* |
|---|---|---|---|---|
| Avec brouette | 21,7 | 16,8 | 4,0 | *42,5* |
| Sans brouette | 29,3 | 22,7 | 5,5 | *57,5* |
| *Total* | *51,0* | *39,5* | *9,5* | *100* |

Tab. 4.2 – Pourcentages théoriques (en pourcentages, arrondis)

Et nous savons aussi en déduire rapidement les effectifs théoriques correspondant :

| | Sociologue | Banquier | Archéologue | *Total* |
|---|---|---|---|---|
| Avec brouette | 43,4 | 33,6 | 8,0 | *85* |
| Sans brouette | 58,7 | 45,4 | 10,9 | *115* |
| *Total* | *102* | *79* | *19* | *200* |

Tab. 4.3 – Effectifs théoriques (arrondis)

Intuitivement, il semble assez logique maintenant de comparer les effectifs observés avec les effectifs théoriques. On peut donc calculer les écarts entre les deux pour chaque case du tableau en soustrayant le tableau 4.3 du tableau 4.1 :

| | Sociologue | Banquier | Archéologue | *Total* |
|---|---|---|---|---|
| Avec brouette | -6,4 | 2,4 | 3,9 | *0* |
| Sans brouette | 6,4 | -2,4 | -3,9 | *0* |
| *Total* | *0* | *0* | *0* | *0* |

Tab. 4.4 – Écarts entre effectifs observés et effectifs théoriques (arrondis)

La première chose que l'on remarque est que la somme des écarts vaut 0 pour chaque ligne et chaque colonne du tableau. Pourquoi ? Tout simplement parce que nous l'avons bien cherché !

En effet, la contrainte sur les marges que nous avons définie dans la section précédente pour le calcul des effectifs théoriques disait que les sommes en lignes et en colonnes des effectifs observés devaient être les mêmes que celles des effectifs théoriques. Ceci implique donc que la somme des écarts doit être égale à 0 pour chaque ligne, chaque colonne, et donc pour la totalité du tableau.

Pour bien comprendre, prenons la deuxième colonne de notre tableau. Dans la première case, nous avons ajouté 2,4 aux effectifs observés pour passer aux théoriques. Comme nous voulons avoir le même total au bout du compte, on a guère le choix sur ce qu'on peut faire dans la deuxième case : Si on a rajouté 2,4 dans la première, on est obligé d'enlever la même chose dans la deuxième. Et la somme du tout vaut forcément 0.

## 4.2 Variations à l'échelle d'une cellule

**Avertissement :** cette section a tendance à s'éloigner du $\chi^2$ proprement dit, elle est de plus d'une lecture plutôt ardue. Son intérêt étant davantage pédagogique que pratique, elle peut être allègrement ignorée en cas de première lecture ou de début de mal de crâne. On passera alors directement à la section suivante, page 17.

Bien, nous avons désormais notre tableau d'écart. Il est très joli. Mais, au fond, il ne nous dit pas grand-chose. Essayons de comprendre ce que signifie la première ligne : ce qu'elle nous dit, c'est que nous avons 6,4 sociologues à brouette de moins que ce à quoi on aurait dû s'attendre avec l'hypothèse d'indépendance. Par contre, nous avons 2,4 banquiers et 3,9 archéologues à brouette de plus. C'est intéressant, mais concrètement, c'est beaucoup ou c'est pas beaucoup ?

Essayons de reformuler la question. 6,4 sociologues à brouette en moins, est-ce que c'est dû à la variation due au biais d'échantillonnage ou au fait qu'il y a un lien entre les deux variables ?

Reformulons encore : si on recommençait notre enquête plusieurs fois, est-ce qu'on obtiendrait souvent un écart de 6,4 ? Ou est-ce que l'écart varierait beaucoup d'une enquête à l'autre ?

L'idéal pour cela serait de pouvoir disposer d'une population correspondant à notre questionnement et d'interroger un échantillon aléatoire tiré à plusieurs reprises dans cette population pour voir quels résultats on obtient. C'est très difficile à faire en pratique, mais c'est très facile à simuler avec un ordinateur.

Pour cela, nous nous plaçons sous l'hypothèse d'indépendance. On imagine que nous disposons d'une population très vaste parmi laquelle nous savons que la proportion de sociologues à brouettes est exactement de 21,7 %, c'est-à-dire la fréquence théorique que nous avons calculée sous hypothèse d'indépendance.

On choisit 200 personnes au hasard dans cette population et on note le nombre de sociologues à brouette parmi ces 200 personnes. Ensuite on recommence : on choisit à nouveau 200 personnes

et on note sur la même feuille le nombre de sociologues avec brouette. Et on recommence. Et on recommence.

On obtient une liste de chiffres qui pourrait ressembler à ça :

50 48 44 49 46 51 53 44 42 44 36 34 42 41 58 45 37 35 38 39

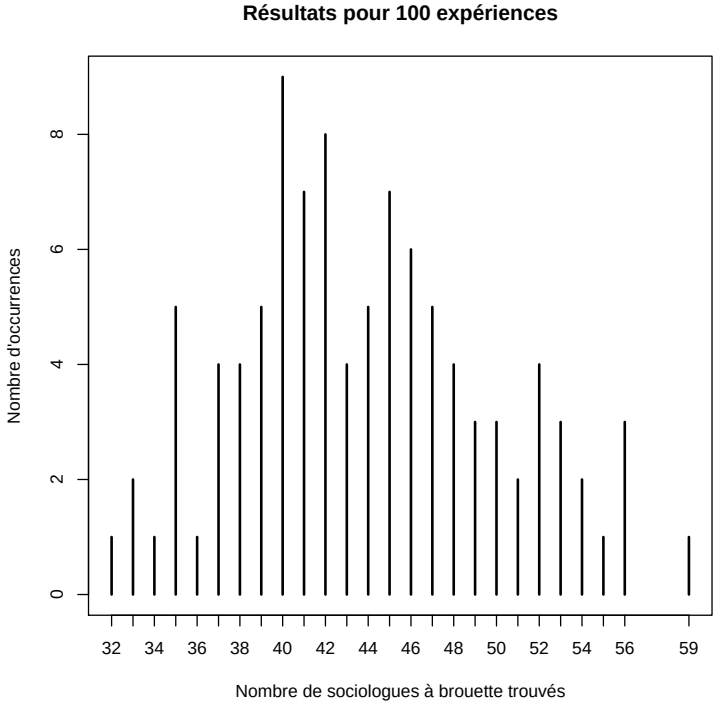
Qu'avons nous fait exactement ? En notant le nombre de sociologues à brouettes parmi les 200 personnes, nous n'avons rien fait d'autre que de noter l'effectif de la case du tableau croisé correspondant aux sociologues possédant une brouette. Et en utilisant une fréquence de 21,7 % de sociologues à brouettes, nous nous sommes mis dans les conditions exactes d'expérience exigées par l'hypothèse d'indépendance entre les variables. Nous avons donc simulé par ordinateur, et à plusieurs reprises, une réalisation de notre enquête sous l'hypothèse d'indépendance.

Maintenant on va oublier les tableaux (pas pour longtemps rassurez-vous) et on va faire des dessins.

Imaginons que nous reproduisons l'expérience 100 fois. On se retrouve avec une série de 100 nombres ressemblant à celle indiquée précédemment. On va maintenant compter le nombre de fois où on retrouve chaque nombre, c'est à dire le nombre de fois où on a trouvé 42 sociologues à brouettes, le nombre de fois où on a trouvé 43 sociologues à brouettes, etc. On obtient un tableau qui ressemble à ça :

| | | | | | |
|---|---|---|---|---|---|
| Nombre de sociologues à brouette | ... | 41 | 42 | 43 | ... |
| Nombre d'occurrences | ... | 10 | 9 | 12 | ... |

Enfin, on transforme ce tableau en graphique pour avoir une idée de la répartition de l'ensemble des nombres trouvés. Ce qui donnerait quelque chose comme la figure suivante :




Ce que nous dit la figure, c'est qu'on a trouvé au minimum 32 et au maximum 59 sociologues à brouettes parmi nos 100 simulations d'enquêtes, et que le nombre de sociologues à brouette le plus fréquemment observé est de 40.

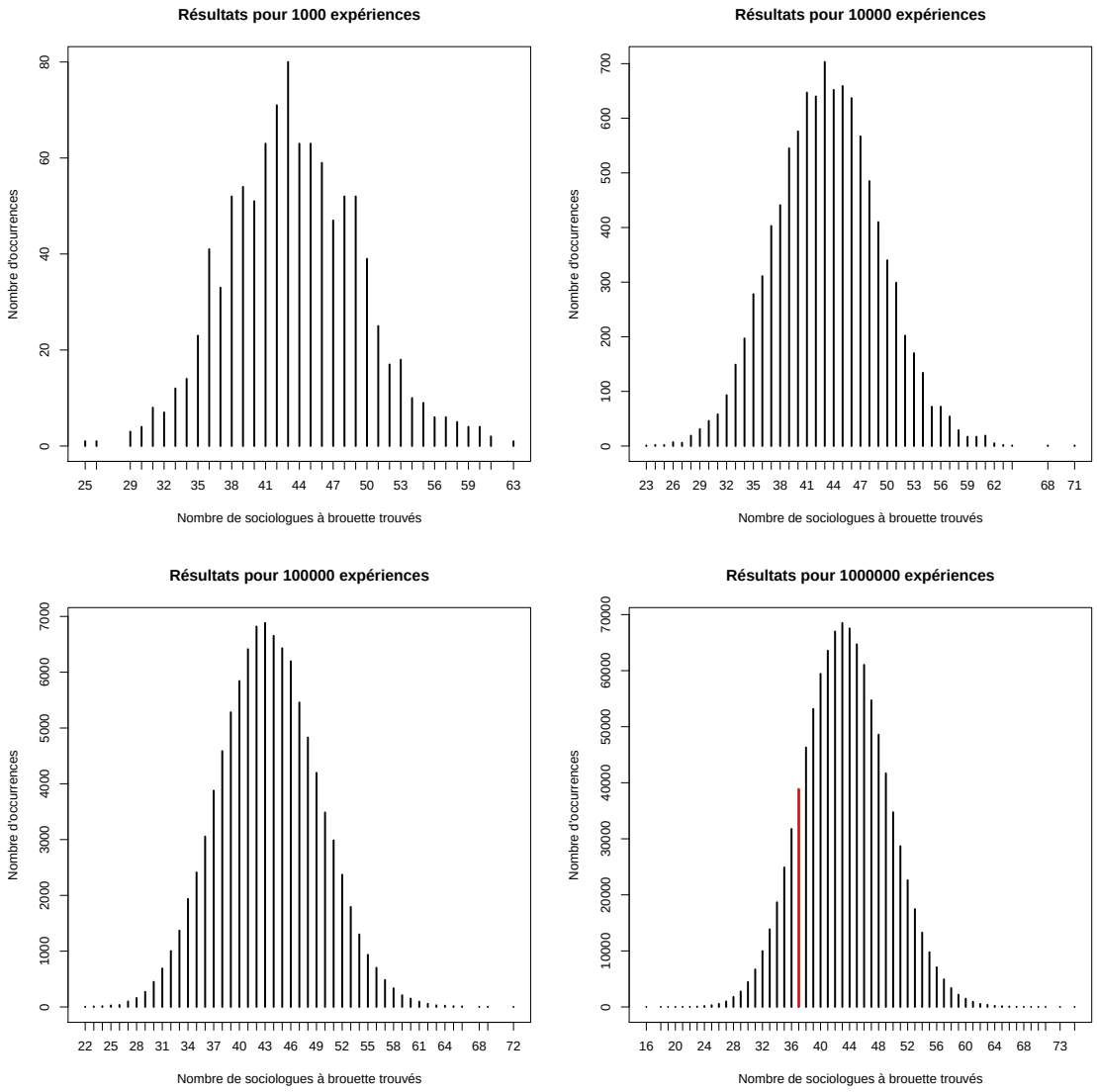



FIG. 4.1 – Simulation du tirage de sociologues à brouette

L'avantage d'une simulation par ordinateur c'est qu'on peut en faire facilement autant qu'on veut. On vient d'en faire 100, on va maintenant en faire 1 000, 10 000, 100 000 et 1 000 000. Les résultats sont indiqués figure 4.1.

Que constate-t-on ? d'abord la forme de la répartition semble se stabiliser avec le nombre de tirages, pour atteindre une forme qui rappelera sans doute quelque chose à ceux qui ont subi quelques cours de statistiques durant leurs études. En gros, plus on fait d'expériences et plus on observe que les résultats ressemblent à la fonction de densité d'une loi normale (ou courbe de Gauss). Le maximum semble être atteint pour la valeur 43. Or, on remarquera que les effectifs théoriques que nous avons calculé s'élèvent justement à 43,4. C'est normal, car les effectifs théoriques sont ceux qu'on a la plus grande probabilité de trouver sous l'hypothèse d'indépendance.

Soit, voilà une bien jolie courbe. Mais cela ne répond toujours pas à notre question de savoir si l'écart que nous avons observé est « important » ou non.

Pour cela nous pouvons regarder où se trouve l'effectif observé dans notre « vraie » enquête, c'est-à-dire 37, dans le dernier graphique de la figure 4.1. Pour éviter la survenue d'une presbytie

trop précoce, nous avons pris la peine de surligner la barre du graphique incriminée en rouge.

Le nombre de fois où on a trouvé 37 s'élève en fait à 38 806. Si on ramène à notre million d'expériences cela signifie qu'on a 3,9 chances sur 100 de trouver un tel résultat sous l'hypothèse d'indépendance des deux variables. En pratique, la probabilité associée à la seule valeur 37 nous intéresse en fait assez peu : ce qui nous intéresse c'est de savoir si 37 est une valeur « significativement petite » ou pas. Donc ce qu'on cherche, ce n'est pas la probabilité d'obtenir exactement 37, mais plutôt celle d'obtenir 37 *ou moins*.

Ici, on obtient une valeur inférieure ou égale à 37 dans 155 360 cas sur un million, soit une probabilité de 15,5 chances sur 100. Ça n'est pas énorme, mais pas non plus négligeable.

Reformulons ce que nous venons de dire : si obtient 37 en valeur observée, il y a 15,5 chances sur 100 que cette valeur soit due au hasard, c'est-à-dire au biais d'échantillonnage.

Reformulons encore : si on observe un effectif de 37 et qu'on affirme qu'il y a un lien entre le fait d'être sociologue et le fait d'avoir une brouette, on a 15,5 chances sur 100 de se tromper. Est-ce que c'est beaucoup ou pas ? La statistique n'a pas de réponse à cette question. Par convention, elle fixe cependant un « seuil de significativité » qui est en général à 5 chances d'erreur sur 100 (c'est le fameux « significatif au seuil de 5 % »). Ce n'est qu'une convention, mais à défaut d'être mathématique elle a pour elle le fait que presque tout le monde l'utilise.

Qu'avons nous fait ici ? Nous avons montré qu'on peut, par simulation, arriver à calculer la probabilité d'obtenir un effectif observé au plus égal à une certaine valeur sous l'hypothèse d'indépendance. La statistique ne nous permet pas de dire si une valeur observée est significativement plus petite ou significativement plus grande *en soi*, mais elle permet d'estimer une probabilité d'observer cette valeur dans le cas où les deux variables sont indépendantes.

## 4.3 $\chi^2$ partiels et $\chi^2$ du tableau

Nous venons donc de voir comment, par simulation, on pouvait essayer de déterminer si les variations observées à l'échelle d'une cellule ont peu ou beaucoup de chances d'être dues au hasard, ou plus précisément au biais d'échantillonnage. Il nous reste à voir la même chose, mais cette fois au niveau du tableau tout entier.

Intuitivement [1], pour passer de la case du tableau au tableau tout entier, on aurait envie de faire la somme de tous les écarts observés dans chaque case pour obtenir une sorte d'écart global à l'indépendance à l'échelle du tableau. Et bien c'est une excellente idée que vous avez là, et je vous en félicite, mais comme d'habitude il y a encore une ou deux subtilités dont il va falloir tenir compte.

Tout d'abord, si on essaie immédiatement de faire la somme des écarts du tableau 4.4 page 14, on obtient tout aussi immédiatement... 0 ! Si cela ne vous semble pas logique, c'est que vous n'avez pas lu assez attentivement le paragraphe causant des contraintes sur les marges, page 14. C'est donc l'occasion de vous resservir un café ou un jus de tomates et de reprendre la lecture de ce passionnant passage.

Faire la somme, c'est donc une bonne idée, mais il faut tenir compte du fait que certains écarts sont positifs et d'autres négatifs et que tout ça finit par s'annuler. On pourrait s'en sortir en faisant la somme de la valeur absolue de chaque écart (c'est-à-dire en transformant les écarts négatifs en écart positif), mais les statisticiens, souvent d'humeur un peu chafouine, préfère utiliser le carré des écarts, ce qui revient à peu près au même dans la mesure où le carré d'un nombre est toujours positif.

Il reste une deuxième subtilité à prendre en compte, que nous comprendrons mieux en regardant

1. En fait ce n'est pas intuitif du tout, mais l'expression *intuitivement* permet à l'auteur d'éviter de fournir de nouvelles explications laborieuses tout en donnant l'impression que pour lui tout ça c'est quand même vachement simple et naturel.

directement le tableau 4.4. Si nous regardons la case des sociologues sans brouette, nous constatons un écart de 6,4. Si on regarde celle des archéologues avec brouette, on obtient un écart de 3,9. Spontanément on pourrait vouloir comparer les deux valeurs en affirmant que l'écart est plus grand chez les sociologues sans brouette que chez les archéologues avec brouette. Mais il faut tenir compte d'une chose : les effectifs théoriques ne sont pas du tout les mêmes dans les deux cases, puisque nous avions 58,7 sociologues sans brouette attendus contre 8 archéologues avec brouette. Or, un écart de 6 sur une valeur de référence de 58 semble tout de suite moins importante qu'un écart de presque 4 sur une valeur de référence qui vaut 8...

En additionnant les écarts de toutes les cases sans tenir compte des effectifs de référence auxquels ces écarts se rapportent, on risque donc de mélanger des choux, des carottes, des pommes de terre et des betteraves. Tout ça peut faire une très bonne soupe (surtout si on enlève les betteraves), mais du point de vue mathématique le mélange est assez indigeste.

Pour éviter de boire le potage, on va donc effectuer une opération assez courante en statistiques, et qu'on nomme *standardisation*, ce qui signifie qu'on va tout rapporter à une même échelle, ce qui va permettre de pouvoir travailler sur des choses comparables entre elles. En pratique, on va diviser la valeur des écarts par celle des effectifs théoriques correspondant.

**Récapitulons** Nous avons notre tableau d'effectifs observés, notre tableau d'effectifs théoriques. Nous pouvons à partir de là calculer les écarts entre les deux, mais pour raisonner à l'échelle du tableau entier nous devons rendre les écarts comparables en tenant compte d'une part de leur signe (en les élevant au carré) et d'autre part du fait qu'ils ne se rapportent pas aux mêmes effectifs de départ (en les divisant par les effectifs théoriques). On va donc calculer un nouveau tableau dont les cases contiennent la valeur suivante :

$$\frac{(\text{Effectif observé} - \text{Effectif théorique})^2}{\text{Effectif théorique}}$$

Cette valeur est appelée le $\chi^2$ *partiel* de la case du tableau. Dans notre exemple, on obtient le tableau suivant :

| | Sociologue | Banquier | Archéologue |
|---|---|---|---|
| Avec brouette | 0,93 | 0,18 | 1,91 |
| Sans brouette | 0,68 | 0,12 | 1,41 |

TAB. 4.5 – $\chi^2$ partiels (arrondis)

*Alléluia !* Nous avons enfin de beaux écarts bien positifs et bien standardisés, que nous allons pouvoir additionner tous ensemble dans la joie et l'allégresse. Ce faisant, nous obtenons la fort jolie valeur de 5,2402, qui n'est rien d'autre que la valeur du $\chi^2$ pour notre tableau croisé.

Passée l'euphorie bien compréhensible due à la beauté de ce résultat arraché à grand renfort d'empilements successifs de subtilités statistiques et de verres de jus d'artichaut vides dans l'évier de la cuisine, nous devons néanmoins nous rendre à l'évidence : 5,2402, c'est magnifique, mais nous sommes encore et toujours confrontés à la même question : c'est beaucoup ou c'est pas beaucoup ?

Avant de répondre, nous allons devoir tenir compte d'une dernière subtilité statistique. Ne vous inquiétez pas si ce genre de phrase commence à générer chez vous une certaine lassitude. Mais regardez là-bas au fond, ne voyez vous pas une faible lueur apparaître dans l'obscurité ? Le bout du tunnel n'est pas loin, et vous devriez l'atteindre encore plus facilement en reprenant un grand verre de nectar d'avocat.

## 4.4 Les degrés de liberté

La dernière chose dont nous devons tenir compte pour obtenir le résultat définitif de notre test porte le doux nom de *degré de liberté*. L'appellation ne manque pas de charme, mais la notion qu'elle recouvre n'est pas forcément la plus intuitive qui soit [2].

En fait, la notion de degrés de libertés dans le cas du test du $\chi^2$ d'indépendance d'un tableau croisé signifie que la valeur calculée du $\chi^2$ pour ce tableau doit être rapportée au nombre de colonnes et de lignes du tableau en question.

Pour tenter de comprendre, reprenons une célèbre enquête menée auprès de 100 professeurs agrégés, 50 en lettre modernes et 50 en lettres classiques, auxquels on a demandé leur style musical préféré. On fait l'hypothèse que les deux variables sont indépendantes. On aurait alors obtenu, par exemple, le tableau suivant :

| | Lettres classiques | Lettres modernes | *Total* |
|---|---|---|---|
| Hip-hop | 20 | 20 | *40* |
| Métal | 30 | 30 | *60* |
| *Total* | 50 | 50 | *100* |

Imaginons maintenant que l'enquête ait distingué des sous-genres musicaux à l'intérieur des catégories *Hip-hop* et *Métal* :

| | Lettres classiques | Lettres modernes | *Total* |
|---|---|---|---|
| Urban Street Gangsta Rap | 5 | 5 | *10* |
| Funky Groovy Soul | 15 | 15 | *30* |
| Industrial Death Metal | 10 | 10 | *20* |
| Gothic Hard Rock | 20 | 20 | *40* |
| *Total* | 50 | 50 | *100* |

Maintenant, imaginons qu'un premier agrégé de lettres classiques n'ait pas entendu la sonnerie du téléphone au moment où notre enquêteur l'appelait car il écoutait le dernier *Dr. X and the freakin' street boyz* à plein volume pendant qu'il travaillait sur une nouvelle traduction de l'*Ancien testament*. Et que du coup c'est un autre agrégé de lettres classiques qui a été enquêté, car celui-ci avait coupé le son de *Sexy groovy funky girlz* pour pouvoir écouter les commentaires du match Lorient - Valenciennes.

Dans le cas de notre deuxième enquête, ceci a une conséquence claire : l'effectif de la case *Lettres classiques - Urban Street Gangsta Rap* perd un enquêté, au profit de la case *Lettres classiques - Funky Groovy Soul*. Mais dans le cas de notre première enquête, cet événement n'a aucune influence : dans les deux cas on reste dans la case *Lettres classiques - Hip-hop*.

Moralité ? Plus il y a de cases dans le tableau, plus les données sont susceptibles de varier aléatoirement et donc plus elles sont sensibles au biais d'échantillonnage.

**Version mathématique** D'un point de vue mathématique, cette notion de « plus grande sensibilité au biais d'échantillonnage » est fortement liée aux contraintes sur les marges.

Pour essayer de comprendre, regardons le premier tableau : de par les contraintes sur les marges, je sais quels doivent être mes totaux en lignes et en colonnes. Maintenant fixons l'effectif de la première case du tableau (20 dans l'exemple donné). Comme je sais que le total de la première ligne vaut 40, j'en déduis immédiatement la valeur de la deuxième case de la première ligne. Et

2. L'auteur l'affirme d'autant plus facilement qu'elle est loin de l'être pour lui-même et que ça fait un moment qu'il se demande comment il va bien pouvoir essayer d'expliquer ce machin.

comme je connais aussi les totaux en colonne, je peux aussi en déduire les valeurs des cases de la deuxième ligne. En fait, dès que je connais la valeur d'une des cases, je connais celles de l'ensemble du tableau. On peut donc considérer que toute la variabilité possible du tableau est contenue dans une seule case.

Regardons maintenant le deuxième tableau. Si je fixe la première case, je peux calculer l'effectif de la deuxième case de la première ligne, mais pas plus. En fait, pour pouvoir reconstruire l'ensemble du tableau, j'ai besoin de connaître les effectifs de trois cases.

De manière plus générale, le nombre de cases d'un tableau pouvant varier « librement » dans un tableau avec contraintes sur les marges est toujours égal à :

$$(\text{Nombre de lignes} - 1) \times (\text{Nombre de colonnes} - 1)$$

Et c'est précisément avec cette formule qu'on calcule le nombre de *degrés de liberté* d'un tableau [3].

## 4.5 Le calcul final

Bien, nous avons désormais d'un côté la valeur du $\chi^2$ pour notre tableau, et de l'autre son nombre de degrés de libertés.

Rappelez-vous ce que nous avions fait dans la section 4.2 page 14 : nous avions réussi à calculer, pour une cellule de tableau, la probabilité d'obtenir un effectif donné sous l'hypothèse d'indépendance. Ce calcul avait été obtenu en faisant toute une série de simulations informatiques. On pourrait procéder de la même manière à l'échelle de l'ensemble du tableau, mais on se heurte vite à deux obstacles :

1. C'est plus compliqué.
2. Les ordinateurs n'existaient pas quand le test du $\chi^2$ a été inventé.

La statistique va donc nous permettre de déterminer directement le même résultat qu'à l'échelle de la cellule, mais sans avoir à effectuer de simulations [4] et en utilisant des raisonnements mathématiques. Elle va ainsi nous permettre de déterminer immédiatement quelle est la probabilité d'obtenir le $\chi^2$ observé sur notre tableau compte tenu du nombre de degrés de libertés et sous l'hypothèse d'indépendance [5].

Pour être un peu plus concret, reprenons notre exemple des sociologues à brouettes. À partir du tableau 4.5 page 18, nous avions déduit que la valeur de notre $\chi^2$ était de 5,2402. Du fait que le tableau en question a 2 lignes et 3 colonnes, nous en déduisons que son nombre de degrés de libertés vaut $(2-1) \times (3-1) = 2$. Et ce que notre logiciel favori va nous indiquer [6], c'est que la probabilité d'observer un tel résultat compte tenu de l'hypothèse d'indépendance s'élève à 0,0728. C'est le fameux $p$.

Comment interpréter ce $p$ ? Plusieurs formulations sont possibles, toutes signifient la même chose :

– la probabilité d'observer une valeur du $\chi^2$ de 5,2402 avec deux degrés de liberté s'élève à 0,0728 ;

3. Les logiciels qui appliquent le test du $\chi^2$ indiquent en général le nombre de degrés de liberté du tableau. En général la notation utilisée est *ddl* pour les logiciels francophones, et *df* pour les anglophones.

4. Les ordinateurs et les algorithmes actuels rendent cependant possibles l'utilisation de simulation, ce qui est peut être très utile dans certains cas. On en reparlera dans le cas où les effectifs théoriques sont considérés comme trop faibles, voir section 6.1 page 30.

5. Plus précisément, ce que nous dit la statistique c'est que la valeur du $\chi^2$ calculé pour un tableau donné sous l'hypothèse d'indépendance des lignes et des colonnes tend vers une loi du $\chi^2$ au nombre de degrés de libertés correspondant à celui du tableau.

6. Auparavant les statisticiens, qui devaient connaître des week-end longs et pluvieux plus fréquemment que la moyenne, s'amusaient à rechercher ces informations dans des tables...

– la probabilité d'obtenir le tableau croisé observé sous l'hypothèse d'indépendance des deux variables est d'environ sept chances sur cent ;
– la probabilité que les lignes et les colonnes du tableau sont indépendantes est d'environ sept chances sur cent ;
– si j'affirme à partir du tableau croisé observé que la profession exercée a une influence sur le fait d'avoir ou non une brouette, j'ai 7 % de chances d'avoir tort, et 93 % de chances d'avoir raison.

Sept chances sur cent de me tromper, c'est beaucoup ou pas ? Là la statistique n'a plus de règle mathématique à fournir. En général, le seuil à partir duquel on considère le résultat comme réellement significatif, c'est à dire le niveau « acceptable » de la probabilité de se tromper, est fixé par convention et habitude à 5 %. Dans le tableau que nous observons depuis maintenant un bon moment, nous sommes donc à la limite : si on se contente d'appliquer mécaniquement le traditionnel seuil de 5 %, alors on dira qu'il n'y a pas de lien statistiquement significatif entre la profession et le fait de posséder une brouette. Mais on peut s'accorder davantage de souplesse et prendre en compte des résultats jusqu'à 10 % ou même un peu plus...

## 4.6 En résumé

La section qui précède a été longue et fastidieuse. Les détails du calcul ne sont là que pour comprendre la démarche et faciliter l'interprétation, les calculs eux-mêmes étant mis en œuvre par un logiciel approprié.

1. Le $\chi^2$ d'un tableau représente l'écart entre la répartition observée dans ce tableau et celle qu'on observerait si les lignes et les colonnes de ce tableau étaient indépendantes, c'est-à-dire, si le fait d'appartenir à une modalité d'une des deux variables croisées n'avait aucun influence sur la modalité d'appartenance de la deuxième variable.
2. Le nombre de degrés de libertés dépend du nombre de lignes et de colonnes d'un tableau.
3. Avec les deux valeurs précédentes, on peut estimer la probabilité $p$ d'obtenir le tableau observé dans le cas où lignes et colonnes sont indépendantes. $p$ représente le nombre de chances que j'ai de me tromper si j'affirme que les deux variables croisées ne sont pas indépendantes.
4. Le seuil de significativité pour le $p$ est par convention fixé à 5 %, ou 0,05, ou 5 chances sur cent. Si le $p$ est supérieur à ce seuil, c'est-à-dire si on a plus de 5 chances sur 100 de se tromper en disant l'inverse, alors on considère que les deux variables sont indépendantes. Sinon, on considère qu'il y a un lien entre les deux.

Nous allons maintenant enfin pouvoir sortir de cette partie théorique aussi distrayante que l'observation d'un escargot par temps sec pour aborder des exemples plus concrets d'utilisation du test et d'interprétation des résultats.

# Partie 5

# Interprétation

## 5.1 Résumé des épisodes précédents

Pour ceux qui n'auraient pas voulu lire les sections précédentes, ceux qui auraient craqué en cours de route, ou ceux qui auraient ressenti le besoin de se reposer un moment avant d'attaquer la suite en faisant deux ou trois semaines de stage de méditation dans un monastère bouddhiste, voici un récapitulatif des idées à bien assimiler pour comprendre ce qui suit.

Le test du $\chi^2$ vise à tester l'hypothèse d'indépendance des lignes et des colonnes d'un tableau croisé. Cette hypothèse signifie que :

1. Le fait d'appartenir à l'une des modalités de la première variable n'a aucune influence sur la modalité d'appartenance de la seconde.
2. Les pourcentages lignes du tableau croisé sont les mêmes pour toutes les lignes.
3. Les pourcentages colonnes du tableau croisé sont les mêmes pour toutes les colonnes.

Le test du $\chi^2$ se base sur la valeur du $\chi^2$ du tableau, qui est une mesure de l'écart entre le tableau observé et le tableau qu'on aurait obtenu si les variables étaient parfaitement indépendantes, et sur le nombre de degrés de liberté du tableau, qui dépend du nombre de lignes et de colonnes.

A partir de ces deux données, le test donne une valeur $p$ qui est la probabilité que les variables soient indépendantes compte tenu du tableau observé, ou encore le nombre de chances de se tromper si on dit que les deux variables ne sont pas indépendantes.

## 5.2 Valeur du $p$

Le tableau suivant est, pour une fois, tirée de données réelles, en l'occurrence celles de l'enquête *Histoire de vie* réalisée en 2003 par l'INSEE [1]. Il croise le fait d'avoir été élevé par sa mère seule jusqu'à 18 ans par la catégorie socio-professionnelle du père en 6 postes.

| | Agriculteur | Indépendant | Cadre | Intermédiaire | Employé | Ouvrier |
|---|---|---|---|---|---|---|
| Élevé par sa mère seule | 22 | 50 | 60 | 57 | 50 | 161 |
| Autre | 990 | 801 | 572 | 800 | 690 | 2861 |

Tab. 5.1 – Croisement de la CS du père avec le fait d'avoir été élevé seul par sa mère

Le $\chi^2$ vaut 44,63, le nombre de degrés de libertés est 5, $p$ vaut 0,00000001726.

1. Dans ces exemples on s'est contenté des données brutes et on n'a pas utilisé la pondération donnée par l'INSEE.

On peut donc rejeter l'hypothèse d'indépendance sans crainte, puisqu'on n'a qu'une chance sur plus de 57 000 000 de se tromper [2]. La catégorie sociale d'appartenance du père a une influence sur le fait d'avoir été ou non élevé par sa mère seule.

Le tableau qui suit croise le fait de pratiquer ou non le football et le sentiment d'appartenir ou non à une classe sociale :

| | Pratique le football | Ne pratique pas le football |
|---|---|---|
| Sentiment d'appartenance | 93 | 3921 |
| Pas de sentiment d'appartenance | 92 | 4165 |
| Ne sait pas | 1 | 131 |

Le $\chi^2$ vaut 1,5448, le nombre de degrés de libertés est 2, $p$ vaut 0,4619.

L'hypothèse d'indépendance entre les deux variables ne peut donc a priori pas être rejetée, on ne peut pas établir de lien entre les deux variables.

## 5.3 Le test du $\chi^2$ est symétrique

Comme on a déjà eu l'occasion de le souligner [3], les lignes et les colonnes d'un tableau croisé sont interchangeables. Vous pouvez donc échanger vos deux variables, le résultat du test sera toujours exactement le même.

Ceci signifie notamment que le tableau n'a pas en lui-même de *sens de lecture* : c'est notre connaissance de l'objet étudié qui nous fait dire « le sexe a une influence sur le fait de préférer la choucroute ou les brocolis » et non pas « le fait de préférer la choucroute ou les brocolis a une influence sur le sexe ».

Ce que le $\chi^2$ nous dit, c'est « les deux variables sont dépendantes ». Ce qu'il ne nous dit pas, c'est « la variable Y est dépendante de la variable X ». Le fait de considérer une variable comme ayant une influence sur une autre relève de l'interprétation et de l'analyse. Cela se traduit en général apr le choix d'utiliser les pourcentages lignes ou les pourcentages colonnes dans la lecture du tableau.

Si on reprend l'exemple du tableau 5.1, l'interprétation va naturellement dans le sens d'une influence de la catégorie sociale du père sur le fait d'avoir été élevé seul par sa mère, et non l'inverse. Ceci se traduit par l'utilisation de pourcentages colonnes pour l'analyse du tableau :

| | Agriculteur | Indépendant | Cadre | Intermédiaire | Employé | Ouvrier | *Ensemble* |
|---|---|---|---|---|---|---|---|
| Élevé par sa mère seule | 2,2 % | 5,9 % | 9,5 % | 6,7 % | 6,8 % | 5,3 % | *5,6 %* |
| Autre | 97,8 % | 94,1 % | 90,5 % | 93,3 % | 93,2 % | 94,7 % | *94,4 %* |
| *Total* | 100,0 % | 100,0 % | 100,0 % | 100,0 % | 100,0 % | 100,0 % | *100,0 %* |

C'est grâce aux pourcentages colonnes qu'on peut approfondir l'analyse du tableau au-delà de la seule existence ou non d'une dépendance entre les variables. Ils nous permettent en effet, par exemple, de constater que seuls 2,2 % des enquêtés dont le père est agriculteur ont été élevé seuls par leur mère, contre 9,5 % de ceux dont le père est cadre, la moyenne pour l'ensemble des enquêtés étant de 5,6 % [4].

2. $\frac{1}{0{,}00000001726} = 57\,937\,428$, soit quatre fois moins de chances que de gagner au Loto.
3. Mais vous aurez remarqué que ce document ne recule pas devant une certaine dose de répétitions, mais si celle-ci frise parfois le radotage.
4. Cette analyse sera grandement facilitée et statistiquement validée par l'utilisation des *résidus*, voir section 5.7 page 26.

## 5.4 Le test du $\chi^2$ dépend du découpage en modalités

Dans ce qui précède on a pu dire indifféremment que le test du $\chi^2$ portait sur l'indépendance des lignes et des colonnes d'un tableau croisé, ou bien sur les deux variables d'un tableau croisé. En fait, la première formulation est plus rigoureuse, car la deuxième tend à masquer le fait que la manière dont chacune des deux variables est découpée en modalités joue un rôle considérable dans la valeur finale du test.

Il semble parfois contre-intuitif d'imaginer que la manière dont on code, découpe ou regroupe une variable en classes ou en modalités puisse influencer le sa dépendance ou son indépendance vis-à-vis d'autres variables. Si on tient compte de la manière dont le $\chi^2$ est calculé, cette influence s'explique cependant assez bien :

- si on regroupe des modalités existantes ou si on en crée de nouvelles, les dimensions du tableau changent, et donc le degré de liberté qui lui est associé également. Ceci influence donc la valeur finale du $p$ ;
- mais surtout, selon la manière dont on regroupe ou éclate ces modalités, on peut masquer des écarts à l'indépendance ou au contraire en faire apparaître de nouveaux.

Prenons un exemple à nouveau tiré de l'enquête *Histoire de vie* en croisant l'âge (découpé en classes) et la variable indiquant si les types d'émission préférés à la télévision sont les séries et les feuilletons. Commençons par un découpage en âges assez fin (ici on donne les pourcentages colonnes) :

| | 25 et moins | 26-35 | 36-45 | 46-55 | 56-65 | 66 et plus | *Ensemble* |
|---|---|---|---|---|---|---|---|
| Oui | 20,4 % | 9,8 % | 7,5 % | 7,5 % | 8,1 % | 12,5 % | *10,2 %* |
| Non | 79,6 % | 90,2 % | 92,5 % | 92,5 % | 91,9 % | 87,5 % | *89,8 %* |
| *Total* | 100,0 % | 100,0 % | 100,0 % | 100,0 % | 100,0 % | 100,0 % | *100,0 %* |

Le $\chi^2$ est extrêmement significatif ($p$ quasiment égal à zéro). On constate que les séries et les feuilletons sont préférés à la fois par les plus jeunes et par les plus âgés [5].

Imaginons maintenant que la question qui nous intéressait au départ était de différencier les moins de 55 ans des plus de 55 ans. Nous aurions alors obtenu le tableau suivant :

| | 55 et moins | 56 et plus | *Ensemble* |
|---|---|---|---|
| Oui | 10,0 % | 10,5 % | *10,2 %* |
| Non | 90,0 % | 89,5 % | *89,8 %* |
| *Total* | 100,0 % | 100,0 % | *100,0 %* |

Avec un $\chi^2$ plus du tout significatif, puisque le $p$ vaut désormais 0,49 ! En regroupant les classes d'âge, on a regroupé des catégories où la préférence pour les séries était sur-représentée et d'autres où elle ne l'était pas du tout. Au final, on a construit deux populations homogènes en regroupant des populations hétérogènes mais opposées.

De manière générale, il est donc préférable de partir avec des découpages en classes les plus détaillés possibles, pour pouvoir éventuellement ensuite pouvoir regrouper entre elles des modalités ayant des profils semblables (identifiés par leurs pourcentages lignes ou colonnes). Dans notre exemple, on aurait pu regrouper les tranches d'âge de 36 à 65 ans pour mieux faire ressortir l'opposition entre les âges intermédiaires et les âges « extrêmes ».

5. Phénomènes bien connus en sociologie des médias et identifiés respectivement sous les noms d'*effet Prison break* et d'*effet Derrick*.

## 5.5 Le test du $\chi^2$ dépend des effectifs

Dans une étude évidemment très sérieuse réalisée par le ministère de la Santé, on a voulu étudier le lien entre le degré de calvitie et le fait d'avoir ou non attrapé un rhume dans les six derniers mois. On a interrogé un premier échantillon en obtenant les résultats suivants :

| | A eu un rhume | N'a pas eu de rhume |
|---|---|---|
| Totalement chauve | 7 | 5 |
| Partiellement chauve | 4 | 8 |
| Porte une perruque | 9 | 12 |

Si on fait les pourcentages lignes, on obtient le tableau suivant :

| | A eu un rhume | N'a pas eu de rhume | *Total* |
|---|---|---|---|
| Totalement chauve | 58,3 % | 41,7 % | *100 %* |
| Partiellement chauve | 33,3 % | 66,7 % | *100 %* |
| Porte une perruque | 42,9 % | 57,1 % | *100 %* |
| *Ensemble* | *44,4 %* | *55,6 %* | *100 %* |

Le $\chi^2$ de notre tableau n'est pas du tout significatif, avec un $p$ de 0,459. Fort déçu, le ministère a décidé de renouveler l'enquête mais en accordant une rallonge budgétaire qui a permis d'interroger dix fois plus de personnes, avec les résultats suivants :

| | A eu un rhume | N'a pas eu de rhume |
|---|---|---|
| Totalement chauve | 70 | 50 |
| Partiellement chauve | 40 | 80 |
| Porte une perruque | 90 | 120 |

Si on calcule les pourcentages lignes de ce nouveau tableau, on obtient exactement les mêmes que précédemment, car les effectifs de chaque case ont tous été multipliés par 10.

Par contre, le $\chi^2$ de ce nouveau tableau est lui devenu très significatif, avec un $p$ inférieur à 0,001.

Que s'est-il passé ? On vient tout simplement d'observer le fait que plus les effectifs de notre tableau augmentent, plus les écarts à l'indépendance observés ont de chances d'être significatifs. Si j'interroge dix personnes et que j'obtiens six fois *oui* et quatre fois *non*, je ne peux rien dire. Mais si j'en interroge 10 000 et que j'obtiens 6 000 *oui* et 4 000 *non*, là je peux en conclure quelque chose.

Le $\chi^2$ est donc extrêmement sensible aux effectifs : plus ceux-ci sont élevés, plus le risque de se tromper en rejetant l'hypothèse d'indépendance est faible, et donc plus la valeur du $p$ est petite. Un $\chi^2$ non significatif peut donc signifier soit qu'il y a indépendance entre les lignes et les colonnes du tableau (dans le cas où les pourcentages lignes ou colonnes sont très proches les uns des autres), soit qu'il n'y a pas indépendance mais que les effectifs dont je dispose ne me permettent pas d'en être sûr statistiquement (dans le cas où les pourcentages lignes ou colonnes sont sensiblement différents).

## 5.6 Le test du $\chi^2$ ne mesure pas l'intensité de la dépendance

En fait, ceci découle directement de la section précédente et de la sensibilité du $\chi^2$ aux effectifs. Prenons les deux tableaux suivants :

| | Rouge | Vert |
|---|---|---|
| Rond | 10 | 20 |
| Carré | 20 | 10 |

| | Rouge | Vert |
|---|---|---|
| Rond | 100 | 200 |
| Carré | 200 | 100 |

Si on veut parler de la force de la dépendance entre les deux variables, on ne peut pas différencier ces deux tableaux : la répartition des effectifs entre les cases est la même, les pourcentages lignes et colonnes sont identiques. Pourtant si dans le premier cas on a bien un $\chi^2$ significatif d'une valeur de 5,4 avec un $p$ de 0,02, dans le second le test devient extrêmement significatif avec un $\chi^2$ de 65,34 et un $p$ quasiment égal à zéro.

Le raisonnement ici est exactement le même que dans la section précédente : pour une même répartition dans mon tableau, j'ai d'autant plus de chances d'être significativement éloigné de l'indépendance que mes effectifs sont importants.

Ce qu'on peut en conclure ici c'est que les valeurs du $\chi^2$ et du $p$ ne doivent pas être utilisées comme indicateurs de la force du lien de dépendance entre les variables du tableau croisé. On ne peut donc pas comparer les résultats du test du $\chi^2$ pour deux tableaux différents en en concluant que la dépendance entre les variables serait plus forte pour l'un que pour l'autre [6].

## 5.7 Les résidus

Les résidus sont une aide à l'interprétation extrêmement utile pour l'analyse d'un tableau croisé. Pour le dire rapidement, le $\chi^2$ indique si les écarts à l'indépendance sont significatifs à l'échelle du tableau, les résidus, eux, donnent cette indication à l'échelle de chaque cellule. Leur résultat est en fait très proche de ce que nous avons effectué dans la section *Variations à l'échelle d'une cellule*, page 14.

Dans cette section, nous avions tenté de voir comment on peut, par simulation, estimer si, à l'échelle d'une case, un écart entre un effectif observé et un effectif attendu était statistiquement significatif ou non. Les résidus permettent d'obtenir cette information pour toutes les cases et donc de déterminer dans quels sens vont les écarts et où ceux-ci sont significatifs.

D'un point de vue mathématique, il existe deux types de résidus : les *résidus de Pearson* et les *résidus de Pearson standardisés* (ou ajustés). La différence entre les deux a relativement peu d'importance, car leur interprétation est semblable. D'un point de vue calcul et à titre tout à fait indicatif, la formule pour les résidus de Pearson est la suivante :

$$\frac{\text{Effectifs observés} - \text{Effectifs théoriques}}{\sqrt{\text{Effectifs théoriques}}}$$

La formule des résidus est un tantinet plus complexe [7], mais l'interprétation est la même dans les deux cas.

Au final il n'y a que deux choses à retenir :

6. Pour être tout à fait rigoureux, on pourrait le faire mais seulement quand les deux tableaux ont les mêmes dimensions et les mêmes effectifs totaux. Mais dans tous les cas on préfère utiliser des indices calculés exprès pour, comme le $V$ de Cramer, que nous verrons section 7.1 page 33.

7. Pour plus d'informations, on pourra se reporter à (Agresti, 2002, p. 81).

– si un résidu est positif, c'est que les effectifs dans la case sont supérieurs à ceux attendus sous l'hypothèse d'indépendance. S'il est négatif, c'est que les effectifs observés sont inférieurs aux effectifs théoriques ;
– les résidus correspondant à des écarts statistiquement significatifs sont *grosso modo* ceux dont la valeur est supérieure à 2 ou inférieure à -2 [8].

Tout cela peut sembler compliqué, mais un exemple permettra de mieux comprendre de quoi il s'agit. Exemple réel cette fois, tiré toujours de l'enquête *Histoire de vie*, et pour lequel nous allons croiser la catégorie sociale et le sentiment d'appartenir ou non à une classe sociale :

| | Appartient | N'appartient pas | Ne sait pas |
|---|---|---|---|
| Agriculteur | 125 | 194 | 9 |
| Indépendant | 190 | 300 | 6 |
| Cadre | 588 | 433 | 9 |
| Intermédiaire | 842 | 694 | 10 |
| Employé | 1105 | 1227 | 38 |
| Ouvrier | 888 | 1024 | 45 |

Le $\chi^2$ est extrêmement significatif, avec un $p$ proche de zéro.

On peut regarder les pourcentages lignes :

| | Appartient | N'appartient pas | Ne sait pas |
|---|---|---|---|
| Agriculteur | 38,1 % | 59,1 % | 2,7 % |
| Indépendant | 38,3 % | 60,5 % | 1,2 % |
| Cadre | 57,1 % | 42,0 % | 0,9 % |
| Intermédiaire | 54,5 % | 44,9 % | 0,6 % |
| Employé | 46,6 % | 51,8 % | 1,6 % |
| Ouvrier | 45,4 % | 52,3 % | 2,3 % |
| *Ensemble* | 48,4 % | 50,1 % | 1,5 % |

Plus le nombre de cases est élevé, plus il devient difficile de lire le tableau. Regardons ce que valent les résidus (ici les résidus de Pearson) :

| | Appartient | N'appartient pas | Ne sait pas |
|---|---|---|---|
| Agriculteur | **-2,7** | **2,3** | 1,8 |
| Indépendant | **-3,2** | **3,3** | -0,6 |
| Cadre | **4,0** | **-3,7** | -1,7 |
| Intermédiaire | **3,4** | **-2,9** | **-2,8** |
| Employé | -1,2 | 1,1 | 0,4 |
| Ouvrier | -1,9 | 1,4 | **2,8** |

Les résidus permettent « d'orienter le regard » vers les cases où les écarts sont statistiquement significatifs. *A priori*, en regardant ce dernier tableau on peut se rendre compte que le sentiment d'appartenance à une classe sociale est moins fréquent que la moyenne chez les agriculteurs et les indépendants, tandis qu'il l'est plus chez les cadres et les professions intermédiaires. Par ailleurs, ceux-ci sont moins nombreux que la moyenne à ne pas savoir s'ils appartiennent ou non à une classe sociale, tandis que les ouvriers sont un peu plus nombreux que la moyenne à être dans ce cas.

8. Ceci étant dû au fait que les résidus tendent à suivre une loi normale centrée réduite.

Il y a cependant une chose importante à noter lorsqu'on utilise les résidus, c'est que ceux-ci mesurent la significativité de l'écart par rapport aux effectifs théoriques attendus de la case. Ils sont donc liés à ces derniers : un écart de 10 quand les effectifs théoriques étaient de 20 (c'est-à-dire un effectif observé de 30) sera sans doute significatif, tandis que le même écart de 10 quand les effectifs théoriques sont de 2 000 ne le sera pas.

Ainsi, de la même manière que pour le $\chi^2$, avoir un résidu très supérieur à 2 ne signifie pas que l'écart entre effectifs observés et effectifs théoriques est très élevé. Ceci signifie juste qu'il est très significativement différent de zéro. Dans notre exemple, si on regarde la case des ouvriers ne sachant pas s'ils appartiennent ou non à une classe sociale, on a un résidu supérieur à 2 avec un écart de « seulement » 0,8 points par rapport au profil moyen (2,3 % contre 1,5 %). Encore une fois, c'est en se rapportant aux pourcentages lignes ou colonnes qu'on peut voir si l'écart au profil moyen est élevé ou pas.

Résumons :

- les résidus indiquent dans quel case on a des sur-représentations (si leur valeur est supérieure à 2) ou des sous-représentations (si elle est inférieure à -2) statistiquement significatives ;
- ils orientent le regard vers les cases pour lesquelles on peut dire quelque chose, et montrent à l'inverse celles pour lesquelles l'écart au profil moyen n'est pas significatif ;
- en dernier lieu ce sont toujours les pourcentages lignes ou colonnes qui permettent de mesurer l'amplitude de cet écart.

Les résidus sont donc très utiles pour l'analyse d'un tableau dont le $\chi^2$ permet de rejeter l'hypothèse d'indépendance. Ils le seront d'autant plus que le tableau comporte un grand nombre de cases. Ils permettent de plus de valider statistiquement les écarts observés à l'échelle de la case [9].

**Représentation graphique** L'utilisation des résidus a un autre avantage, c'est de permettre la représentation graphique de tableaux croisés incluant les liens entre les différentes modalités, c'est à dire les cases dans lesquelles les effectifs observés sont significativement supérieurs ou inférieurs aux effectifs théoriques.

Prenons par exemple la figure 5.1 page ci-contre. Elle représente le tableau croisant, pour l'enquête *Histoire de vie*, la catégorie professionnelle de l'enquêté et la fréquence de ses visites à sa famille proche ou éloignée. Ce graphique contient une représentation visuelle de chaque case construite de la façon suivante :

- la largeur de chaque case est proportionnelle au pourcentage ligne correspondant. On a d'ailleurs indiqué dans chaque case la valeur de ce pourcentage ;
- la surface de la case est proportionnelle aux effectifs observés ;
- la couleur de la case dépend de la valeur du résidu de Pearson associé : bleu si le résidu est significativement positif, rouge s'il est significativement négatif, gris s'il n'est pas significatif.

La lecture de ce type de graphique n'est peut-être pas évidente de prime abord, mais une fois habitué elle permet de synthétiser de manière visuelle la quasi-totalité des informations nécessaires pour l'analyse.

Pour reprendre l'exemple de la figure 5.1, on peut ainsi voir immédiatement que les employés et les ouvriers ont plus fréquemment des visites familiales hebdomadaires, tandis que les cadres et les professions intermédiaires en ont moins souvent. On remarquera également que le pourcentage est très élevé chez les agriculteurs (49,4 %), mais qu'il n'est pas significatif, sans doute du fait d'effectifs trop faibles. On peut également remarquer que les cadres ont plus souvent des fréquences de visite intermédiaires (plusieurs fois par mois ou par an) tandis que les ouvriers ont plus souvent des fréquences de visite « extrêmes » (soit hebdomadaires, soit exceptionnelles ou inexistantes).

Ce type de graphique en mosaïque permet donc de faciliter l'analyse, là encore plus particulièrement dans le cas de tableaux croisés avec un nombre de cases élevé.

9. Il est dommage que certaines logiciels comme Modalisa ne proposent pas le calcul des résidus pour les tableaux croisés, même si dans ce cas l'utilisation du PEM (pourcentage de l'écart maximum) s'en rapproche (Cibois, 1993).

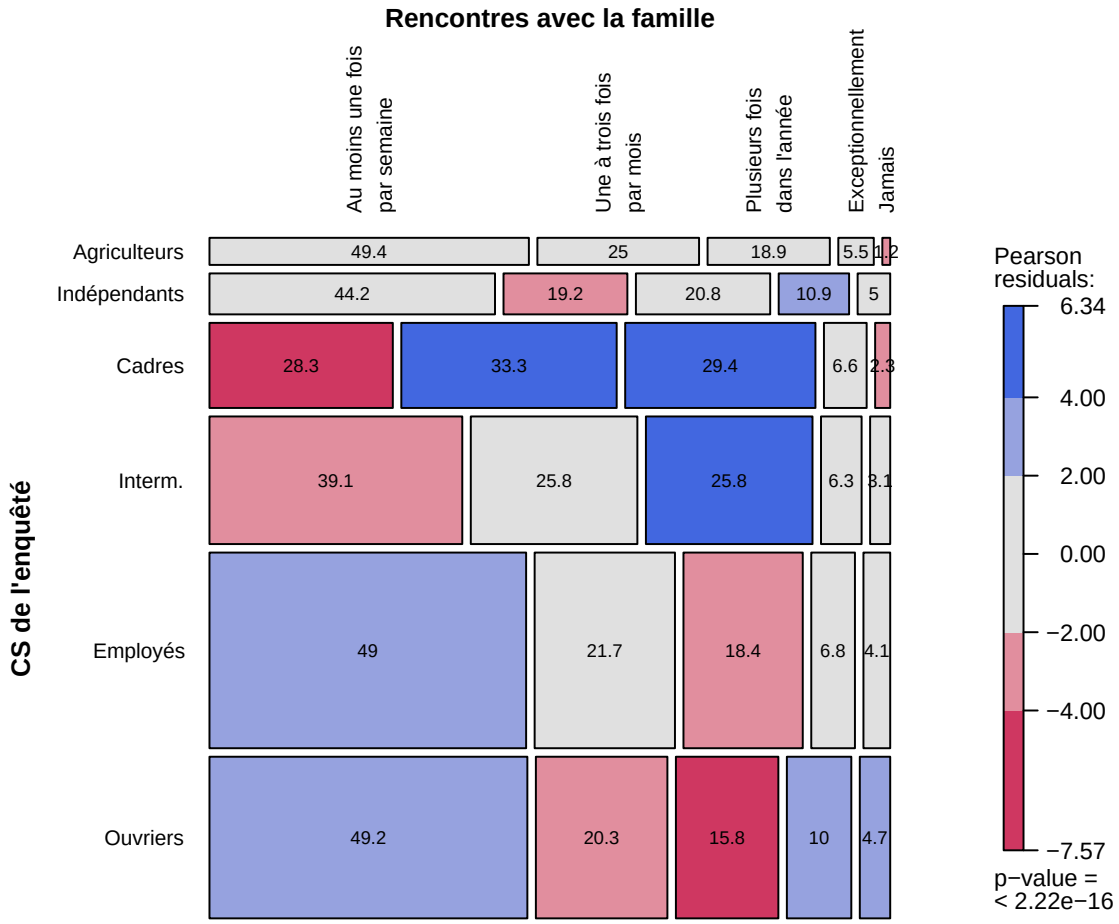



FIG. 5.1 – Graphique en mosaïque du croisement entre la CS de l'enquêté et la fréquence des visites dans la famille

# Partie 6

# Limites

## 6.1 Fausse limite : quand les effectifs théoriques sont trop faibles

Commençons par un exemple. Soit le tableau croisé suivant, qui s'intéresse au fait de gagner ou non au Loto selon qu'on possède un trèfle à quatre feuilles, un fer à cheval ou aucun des deux :

| | Perdant | Gagnant |
|---|---|---|
| Trèfle | 220 | 7 |
| Fer | 200 | 1 |
| Aucun | 200 | 1 |

Le $\chi^2$ est significatif, avec un $p$ à 0,03. Cependant tout bon logiciel de statistique qui se respecte devrait vous gratifier d'un joli message d'avertissement vous annonçant amicalement que le résultat obtenu pourrait bien n'être pas plus valable que celui d'un thème astral réalisé par un docteur en sociologie.

Pourquoi donc ? Car en calculant votre $\chi^2$, vous avez enfreint le commandement suivant : *dans tout tableau croisé, jamais plus de 20 % d'effectifs théoriques inférieurs à 5 tu n'auras.*

Qu'est-ce que c'est encore que ça ? Pour comprendre l'origine de ce principe, il faut se rappeler que le résultat du test du $\chi^2$ (le $p$) est une *approximation*, qui en toute rigueur ne deviendrait parfaitement exacte que quand les effectifs de mon tableau seraient extrêmement élevés.

Plus précisément, on peut se rappeler que dans le calcul des $\chi^2$ partiels associés à chaque case, on a « standardisé » l'écart entre effectifs observés et effectifs théoriques de manière à ce qu'un écart de 15 dans une case où on attendait 6 ne soit pas considéré de la même manière qu'un écart de 15 dans une case où on en attendait 6 000.

Une conséquence de cette standardisation est qu'un poids important est accordé aux petites cases, même si en effectifs les écarts correspondants sont relativement faibles. Reprenons notre tableau et calculons respectivement les effectifs théoriques, les écarts entre effectifs observés et effectifs théoriques, et les résidus :

| | Perdant | Gagnant |
|---|---|---|
| Trèfle | 223,7 | 3,2 |
| Fer | 198,1 | 2,9 |
| Aucun | 198,1 | 2,9 |

*Effectifs théoriques*

| | Perdant | Gagnant |
|---|---|---|
| Trèfle | -3,8 | 3,8 |
| Fer | 1,9 | -1,9 |
| Aucun | 1,9 | -1,9 |

*Écarts*

| | Perdant | Gagnant |
|---|---|---|
| Trèfle | -0,3 | **2,1** |
| Fer | 0,1 | -1,1 |
| Aucun | 0,1 | -1,1 |

*Résidus*

Que constate-t-on ? Malgré la significativité du $\chi^2$, les écarts entre effectifs observés et effectifs théoriques sont plutôt faibles. Les résidus nous indiquent que la seule case où cet écart est significatif est la case « gagnant avec un trèfle », mais celle-ci a un effectif observé de 7 au lieu d'un effectif théorique attendu de 3,2, ce qui ne constitue pas forcément une variation très sensible.

On voit donc comment des variations sur des cases à faible effectif peuvent générer un $\chi^2$ globalement significatif à partir d'écarts pourtant assez minimes en termes d'effectifs. C'est pourquoi une règle assez courante (mais qui relève de la convention et non de la démonstration mathématique) veut que pour éviter ce genre de « perturbations », on ne doit pas avoir, dans un tableau croisé, plus de 20 % des cases avec un effectif théorique inférieur à 5. Dans le tableau qui nous intéresse, ce sont 3 cases sur 6 qui sont dans ce cas, soit 50 %, donc la condition de validité n'est pas remplie.

Bien, et qu'est-ce qu'on fait alors ? On abandonne notre étude, empli de frustration et d'amertume, et quelque peu angoissé à l'idée d'expliquer tout ça à notre directeur de thèse qui était déjà en train de cocher ses numéros, un trèfle à quatre feuilles dans chaque main ? Et bien non !

Comme nous l'avons évoqué précédemment, le fait d'utiliser une approximation mathématique pour évaluer le $p$ du test du $\chi^2$ n'est plus une obligation compte tenu de l'évolution des algorithmes et de la puissance de calcul des ordinateurs. Plutôt que de calculer le $p$ par cette approximation, on peut en effet procéder à une simulation, de la même manière que nous l'avons fait à l'échelle d'une case du tableau dans la section 4.2 [1].

Pour aller très vite, ce calcul du $p$ par simulation s'effectue en tirant au sort un grand nombre de tableaux (plusieurs milliers) dont les lignes et les colonnes sont indépendantes et ayant les mêmes dimensions et les mêmes marges que notre tableau d'intérêt. Pour chaque tableau, on calcule la valeur de son $\chi^2$. Une fois qu'on a tous ces $\chi^2$, on regarde quelle proportion d'entre eux sont supérieurs à celui de notre tableau : ce pourcentage n'est rien d'autre que la valeur du $p$ [2].

Le détail du calcul importe peu. Ce qu'il faut retenir c'est qu'on a là une méthode qui nous permet de calculer un $p$ pour n'importe quel tableau croisé, quels que soient les effectifs théoriques [3]. Si on applique tout ceci à notre exemple, on obtient un $p$ par simulation d'environ 0,025. Notre test demeure donc toujours significatif et nous allons pouvoir poursuivre notre enquête.

Il reste que les résidus nous ont indiqué que l'écart à l'indépendance dans notre tableau se jouait essentiellement sur une seule case, et avec des effectifs très faibles. Parfois cela rend le tableau inintéressant du point de vue de l'analyse. Dans notre cas, montrer que la possession d'un trèfle à quatre feuilles augmente significativement la probabilité de gagner au loto peut être un sujet d'intérêt central dans notre étude et pour notre directeur de thèse.

## 6.2 Vraie limite : les variables cachées

Partons d'un nouvel exemple réel tiré une fois de plus de l'enquête *Histoire de vie* en croisant le fait de tenir ou d'avoir tenu un journal intime, et celui d'avoir pratiqué le tricot, la broderie ou la couture au cours des douze derniers mois.

| | Tient ou a tenu un journal | N'a jamais tenu de journal |
|---|---|---|
| A pratiqué broderie, tricot ou couture | 348 | 1065 |
| N'a pas pratiqué | 1166 | 5824 |

1. Des logiciels comme Modalisa ne le proposent pas. R, lui, le permet à l'aide de l'option `simulate.p.value` de la fonction `chisq.test` (R Development Core Team, 2008).
2. Ceux, ô combien nombreux, que ces questions passionnent pourront se référer à (Chessel, 2005) pour plus de détails.
3. À l'exception des tableaux ayant un effectif théorique nul, mais ceci n'arrive que si l'une des marges du tableau est nulle, c'est donc fort peu probable.

Le $\chi^2$ de ce tableau est très significatif, avec un $p$ quasiment égal à zéro. Le fait de pratiquer la broderie aurait donc une influence sur le fait de tenir un journal intime (ou inversement).

Ce résultat est tout à fait passionnant, mais n'y aurait-il pas un petit biais ? On peut par exemple remarquer que les deux pratiques sont en général perçues comme plutôt « féminines ». Le sexe n'aurait-il donc pas un effet dans tout ça ?

Pour le savoir, la méthode la plus efficace est de recommencer notre test en séparant les hommes et les femmes. On effectue deux test du $\chi^2$ sur les deux tableaux suivants :

| | Journal | Pas de journal |
|---|---|---|
| Couture | 2 | 26 |
| Pas de couture | 286 | 3473 |

*Hommes*

| | Journal | Pas de journal |
|---|---|---|
| Couture | 346 | 1039 |
| Pas de couture | 880 | 2351 |

*Femmes*

Si on regarde les $\chi^2$, on constate qu'aucun des deux n'est significatif : le $p$ vaut 0,79 pour les hommes, et 0,12 pour les femmes. Que peut on en conclure ? Qu'*a priori* la répartition observée dans notre premier tableau n'était pas due à un effet d'une variable sur l'autre, mais au fait que les deux sont étroitement liées au sexe.

On a découvert là ce qu'on appelle l'existence d'une *variable cachée*. On observe une dépendance entre les variables $A$ et $B$, mais en fait cette dépendance provient uniquement du fait que toutes deux dépendent d'une troisième variable $C$. Le plus souvent, $C$ sera une des grandes variables socio-démographiques classiques, comme le sexe ou l'âge. Ainsi, les particularités observées pour la catégorie socio-professionnelle des employés sont assez souvent liées au fait qu'il s'agit d'une catégorie où les femmes sont largement sur-représentées.

La méthode pour vérifier l'existence d'une variable cachée est toujours la même : on applique à nouveau les tests sur des sous-populations à peu près homogènes par rapport à la variable suspectée. Dans le cas du sexe, on séparera les hommes et les femmes. Dans le cas de l'âge, on appliquera le test sur des tranches d'âge plus ou moins fines, etc.

# Partie 7

# Raffinements

Nous détaillons ici des améliorations du test du $\chi^2$ dont vous entendrez peut-être parler ou qui pourront vous être utiles.

## 7.1 Le $V$ de Cramer

Dans la section 5.6 page 26, nous avons montré en quoi le $\chi^2$ n'était pas une mesure du degré de dépendance entre les lignes et les colonnes d'un tableau. On a notamment souligné que du fait de sa sensibilité à la fois à l'effectif total et aux nombres de lignes et de colonnes, les résultats du test du $\chi^2$ et la valeur du $p$ ne peuvent en général pas être comparés d'un tableau à l'autre.

C'est justement pour remédier à ce problème que Monsieur Harald Cramér [1] a mis au point une statistique joliment prénommée $V$ et qui se calcule de la manière suivante :

$$V = \sqrt{\frac{\chi^2}{\text{Effectif total} \times \min(\text{nombre de lignes} - 1, \text{nombre de colonnes} - 1)}}$$

Cette formule compliquée s'applique de la manière suivante : étant donné un tableau, on calcule la valeur de son $\chi^2$, on la divise par l'effectif total lui-même multiplié par la plus petite dimension du tableau à laquelle on aura enlevé un. Puis on fait la racine carrée de tout ça.

Prenons un exemple de calcul sur le tableau suivant (il s'agit d'une copie éhontée du tableau 4.1 page 13) :

| | Sociologue | Banquier | Archéologue |
|---|---|---|---|
| Avec brouette | 37 | 36 | 12 |
| Sans brouette | 65 | 43 | 7 |

Le $\chi^2$ de ce tableau, nous l'avons déjà calculé, vaut 5,24. L'effectif total vaut 200. La plus petite dimension du tableau est le nombre de lignes, qui vaut 2. On obtient donc le calcul suivant :

$$V = \sqrt{\frac{5,24}{200 \times (2-1)}} = 0,162$$

1. Penser à prononcer « Crameur » et non « Cramé ».

Les propriétés du $V$ à retenir sont les suivantes :

- la valeur du $V$ est toujours comprise entre 0 et 1 ;
- plus le $V$ est élevé, plus la dépendance entre les deux variables est forte. Plus le $V$ est faible, plus les variables se rapprochent de l'indépendance. Les cas extrêmes sont $V = 0$, dans le cas où les deux variables sont parfaitement indépendantes, et $V = 1$, dans le cas où les variables sont identiques ;
- le $V$ ne dépendant ni des effectifs ni des dimensions du tableau, il peut être comparé d'un tableau à l'autre.

Prenons comme d'habitude quelques exemples :

| | Homme | Femme |
|---|---|---|
| Choucroute | 20 | 20 |
| Brocolis | 20 | 20 |

$V = 0$

| | Homme | Femme |
|---|---|---|
| Choucroute | 10 | 30 |
| Brocolis | 30 | 10 |

$V = 0,5$

| | Homme | Femme |
|---|---|---|
| Choucroute | 0 | 40 |
| Brocolis | 40 | 0 |

$V = 1$

On voit bien avec ces trois tableaux que le $V$ varie bien en fonction du niveau de dépendance dans le tableau, de 0 (indépendance totale) à 1 (dépendance totale). C'est ce qui lui vaut le nom de *cœfficient de contingence* (la contingence étant l'inverse de l'indépendance) : plus la valeur du $V$ est élevée, plus la contingence dans le tableau est forte.

Par ailleurs, on peut montrer que la valeur du $V$ est insensible à l'effectif total du tableau :

| | Homme | Femme |
|---|---|---|
| Choucroute | 20 | 10 |
| Brocolis | 15 | 35 |
| Lasagnes | 38 | 21 |

$V = 0,34$

| | Homme | Femme |
|---|---|---|
| Choucroute | 200 | 100 |
| Brocolis | 150 | 350 |
| Lasagnes | 380 | 210 |

$V = 0,34$

| | Homme | Femme |
|---|---|---|
| Choucroute | 2 000 | 1 000 |
| Brocolis | 1 500 | 3 500 |
| Lasagnes | 3 800 | 2 100 |

$V = 0,34$

## 7.2 La correction de continuité de Yates

La correction de continuité de Yates vient du fait que les lois statistiques utilisées dans le test du $\chi^2$ sont par nature continues (elles peuvent prendre n'importe quelle valeur, y compris avec plein de zéros derrière la virgule) tandis que les effectifs des cases de notre tableau ne peuvent être que des nombres entiers. Ceci peut entraîner une surévaluation de la valeur du $\chi^2$ dans certains cas.

La correction de Yates consiste à enlever 0,5 à la valeur absolue des écarts entre les effectifs observés et les effectifs théoriques avant de les mettre au carré dans le calcul des $\chi^2$ partiels, ce qui donne la formule suivante :

$$\chi^2\text{partiel} = \frac{(|\text{Effectif observé} - \text{Effectif théorique}| - 0.5)^2}{\text{Effectif théorique}}$$

Les conditions d'application de cette correction ne font pas forcément l'objet d'un consensus. Parfois on la limite aux tableaux ayant 2 lignes et 2 colonnes, parfois non. En général elle est recommandée lorsque les effectifs sont insuffisants, mais là aussi les critères pour le « insuffisant » sont variés.

Bref, le mieux est de laisser faire le logiciel qui, s'il est bien élevé, devrait l'appliquer dans des conditions à peu près définies. Dans tous les cas cette correction n'a d'effet sensible que lorsque les effectifs sont faibles [2].

2. Modalisa, dans sa version 4, applique systématiquement cette correction aux cases dont les effectifs théoriques sont inférieurs à 5. R, lui, l'applique si le tableau est de dimension 2×2.

## 7.3 Le test exact de Fisher pour les tableaux 2 × 2

Le test exact de Fisher est une alternative au test du $\chi^2$, mais qui vise à tester la même hypothèse et s'interprète exactement de la même manière. La principale différence est qu'il s'agit d'un test *exact*, et non d'une approximation tirée d'une loi statistique.

La bonne nouvelle est donc que ce test peut s'appliquer quels que soient les effectifs théoriques du tableau. La mauvaise nouvelle est qu'il est assez gourmand en temps de calcul. C'est la raison pour laquelle on le limite en général aux tableaux de dimension $2 \times 2$. On peut cependant l'appliquer à des tableaux de plus grandes dimensions.

# Partie 8

# Aide-mémoire

*Cette partie récapitule les points importants à retenir de tout ce qui précède. On indique à chaque fois, entre crochets, le numéro de la page correspondant au passage où ce point a été traité.*

Le test du $\chi^2$ s'applique à un tableau croisant deux variables qualitatives [5]. Il vise à tester l'indépendance des lignes et des colonnes de ce tableau.

Dire que les lignes et les colonnes d'un tableau croisé sont indépendantes revient à dire l'une des phrases suivantes [6] :
- le fait d'appartenir à la modalité d'une des deux variables n'a aucune influence sur la modalité d'appartenance de l'autre variable ;
- les profils lignes du tableau croisé sont tous identiques ;
- les profils colonnes du tableau croisé sont tous identiques.

Le test du $\chi^2$ vise à déterminer la probabilité d'indépendance des lignes et des colonnes de notre tableau.

Pour cela, on commence par calculer les effectifs qu'on observerait si les lignes et les colonnes étaient parfaitement indépendants, en tenant notamment compte des contraintes sur les marges du tableau [9]. On obtient ainsi le tableau des effectifs théoriques sous l'hypothèse d'indépendance [10].

On calcule ensuite les écarts entre effectifs observés et effectifs théoriques et on les « standardise » pour qu'ils soient tous positifs et « comparables » : on obtient ainsi le $\chi^2$ partiel pour chaque case du tableau [17]. La somme de ces $\chi^2$ partiels donne la valeur du $\chi^2$ pour notre tableau. À partir de cette valeur et du nombre de degrés de libertés de notre tableau [19], la statistique nous permet de déduire un $p$ qui n'est autre que la probabilité d'obtenir le tableau croisé observé si nos variables étaient indépendantes [20].

Le tableau 8.1 page suivante donne quelques exemples de valeurs de $p$ que l'on peut obtenir et de l'interprétation qui peut en être faite [22].

L'interprétation du test du $\chi^2$ se fait en comparant les profils lignes ou les profils colonnes à leur profil moyen pour déterminer l'importance des écarts [23]. L'utilisation des résidus [26] permet de déterminer, à l'échelle de chaque case, quels sont les écarts qui sont statistiquement significatifs. Ils sont très utiles pour l'analyse notamment quand le nombre de cases est important, et peuvent même conduire à une représentation graphique du tableau croisé [29].

Certains points importants sont à prendre en compte quand on interprète le résultat du $\chi^2$ :

- le découpage des variables en modalités influe considérablement sur le résultat et peut faire apparaître ou masquer des écarts à l'indépendance [24] ;
- la valeur du $\chi^2$ et donc du $p$ est sensible à l'effectif total du tableau : un $p$ inférieur à 5 % peut signifier que les effectifs ne sont pas suffisamment important pour que le lien de dépendance soit statistiquement avéré [25] ;
- le résultat du test n'est pas un indicateur de la force du lien entre les deux variables [26] :

| $p$ | **Interprétation** |
|---|---|
| 1 | Les deux variables sont parfaitement indépendantes |
| 0,7 | Les deux variables sont indépendantes |
| 0,15 | En toute rigueur, on devrait considérer les deux variables comme indépendantes, mais il est possible qu'elles ne le soient pas et que les effectifs sont insuffisants pour le montrer |
| 0,05 | Les variables ne sont pas indépendantes au seuil classique de 5 % |
| 0,0001 | L'hypothèse d'indépendance doit être rejetée, il y a un lien entre les deux variables |
| 0 | L'hypothèse d'indépendance est tellement peu probable que le logiciel n'arrive même pas à afficher tous les zéros derrière la virgule. |

TAB. 8.1 – Exemples de valeur de $p$ et de son interprétation

comme la valeur du $\chi^2$ et du $p$ dépendent des effectifs et des dimensions du tableau, on ne peut comparer ces valeurs d'un tableau à l'autre. Pour ce genre de chose on utilise plutôt un cœfficient de contingence comme le $V$ de Cramer [33] ;
- le lien de dépendance entre les deux variables peut en fait être dû à une variable cachée à laquelle les deux variables étudiées sont liées [31].

Enfin, il faut tenir compte du fait que le test peut perdre en fiabilité dans le cas où des cases du tableau ont des effectifs théoriques faibles [30]. On peut cependant y remédier soit en calculant le $p$ par simulation, soit à l'aide du test exact de Fisher [35].

# Bibliographie


Alan Agresti : *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience, Hoboken, 2002.

Daniel Chessel : Comment comparer des fréquences très faibles ?, avril 2005. URL http://pbil.univ-lyon1.fr/R/querep/qrc.pdf.

Philippe Cibois : Le pem, pourcentage de l'écart maximum : un indice de liaison entre modalités d'un tableau de contingence. *Bulletin de méthodologie sociologique*, (40):43–63, septembre 1993. URL http://pagesperso-orange.fr/cibois/bms93.pdf.

R Development Core Team : *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL http://www.R-project.org. ISBN 3-900051-07-0.