STT-4400 / STT-6210

ANALYSE DE TABLEAUX DE FRÉQUENCES

Notes de cours rédigées par

Sophie Baillargeon
sophie.baillargeon@mat.ulaval.ca

à partir des notes et exercices réalisés par
Louis-Paul Rivest, Marc Simard, Nadia Ghazzali, Chantal Mérette,
Claude Belisle, Thierry Duchesne, Aurélie Labbe et Lajmi Lakhal Chaieb.

Automne 2013

Département de mathématiques et de statistique

Faculté des sciences et de génie

Table des matières

Préface

Ce manuel contient des notes réalisées pour le cours STT-4400 / STT-6210 « Analyse de tableaux de fréquences ». Elles se composent de théorie et d'exemples numériques présentant différents outils statistiques d'analyse de données catégoriques. Bien que plusieurs calculs des exemples aient été faits à l'aide d'un logiciel statistique, on ne retrouve pas de programmes ni de sorties informatiques dans le présent document. Il ne repose donc pas sur un logiciel statistique en particulier. Cependant, des programmes informatiques réalisant les calculs statistiques contenus dans ce document sont présentés dans le cours pendant les séances en classe.

Le cours STT-4400 / STT-6210 se nomme « Analyse de tableaux de fréquences », car à l'origine il traitait uniquement des tableaux de fréquences et des modèles loglinéaires pour analyser ceux-ci. Cependant, avec le temps, la régression logistique et la régression Poisson sont devenues plus populaires dans la pratique statistique. Ces types de régression répondent bien à certaines questions de recherche, car elles comportent une variable réponse, ce qui n'est pas le cas des tableaux de fréquences et des modèles loglinéaires pour ceux-ci. Ces derniers semblent de moins en moins employés. Afin de prioriser des méthodes plus susceptibles d'être utilisées par un statisticien sur le marché du travail, il a été décidé par le comité de programme du baccalauréat en statistique de ne plus enseigner les modèles loglinéaires pour tableaux de fréquences dans ce cours. Ces modèles étaient particulièrement utiles pour analyser des tableaux de fréquences à plusieurs variables. Le lecteur intéressé est référé à Bishop *et al.* (1975).

Plusieurs anciens enseignants du cours ont contribué à ces notes. Louis-Paul Rivest, à l'aide de Marc Simard à l'époque étudiant en statistique, a réalisé les premières notes dans les années 90. Ces notes étaient basées sur la

première édition du livre « An Introduction to Categorical Data Analysis » d'Agresti (2007). Elles ont été révisées à quelques reprises, par Louis-Paul Rivest lui-même, Nadia Ghazzali et Chantal Mérette. Elles ont maintes fois été utilisées pour donner le cours « Analyse de tableaux de fréquences ». En 2004, Claude Belisle a produit ses propres notes pour ce cours, alors qu'en 2005 et 2006, Thierry Duchesne a composé plusieurs exercices. En 2007, Aurélie Labbe a elle aussi produit des notes de cours, qui ont été reprises par Lajmi Lakhal Chaieb en 2008. Le présent manuel vise à unifier tout ce matériel pédagogique et à y faire quelques ajouts, notamment en bonifiant les rappels en introduction et en ajoutant de la nouvelle matière couvrant des modèles logistiques autres que binaires.

En terminant, je remercie le comité de perfectionnement du syndicat des chargé(e)s de cours de l'Université Laval (SCCCUL) pour les fonds qu'ils m'ont alloués pour la réalisation de ce matériel pédagogique.

Sophie Baillargeon
Université Laval
Novembre 2013

Introduction

Ces notes de cours présentent des méthodes statistiques usuelles pour l'analyse de données catégoriques. Des données catégoriques peuvent résulter, par exemple :

- d'un sondage d'opinion pour des consommateurs (ex. : niveau de satisfaction pour un service reçu) ;
- d'une étude dans le domaine de la santé cherchant à cerner les facteurs influençant l'occurrence d'une maladie (ex. : variable réponse prenant la valeur 'oui' pour les sujets malades, 'non' sinon) ;
- d'une étude sociologique sur les habitudes de vie des Québécois (ex. : moyen de transport pour se rendre au travail, nombre de consommations d'alcool par semaine) ;
- etc.

Ce type de données est très courant.

Plusieurs outils statistiques s'offrent à nous pour analyser des données catégoriques. Ces notes de cours visent à présenter les méthodes classiques et celles les plus utilisées en pratique. Dans ce cours, on n'utilisera pas de notions mathématiques poussées. Il s'agit d'un cours appliqué. Il a pour objectif de rendre les étudiants capables de cerner la bonne méthode statistique à utiliser, de l'appliquer correctement en comprenant les idées mathématiques de base derrière les méthodes et de bien interpréter les résultats obtenus.

Dans ces notes, les détails de certains calculs faits à la main pour de petits jeux de données sont présentés. Le but de cet exercice est uniquement de s'assurer de bien comprendre les formules. En pratique, l'ordinateur fait pour nous tous ces calculs.

Avant d'entrer dans le vif du sujet, rappelons certaines notions de statistiques souvent employées dans ce cours.

Définitions

Si vous suivez ce cours, vous connaissez déjà les termes : données, population, individu, échantillon, variable et observation. Afin de s'assurer que l'on part tous du même point, rappelons tout de même les définitions de ces termes :

Données : Des données sont des valeurs numériques (des nombres) ou alphanumériques (des chaînes de caractères) représentant les observations de certaines variables sur certains individus. Elles se présentent souvent sous la forme de jeux de données, c'est-à-dire de tableaux de valeurs, stockées dans un fichier informatique.

Population : La population est l'ensemble de référence sur lequel porte l'étude dans le cadre de laquelle les données ont été recueillies.

Individu ou unité statistique : Un individu est un élément de la population. L'ensemble des individus constitue la population. Chaque observation est associée à un individu.

Échantillon : L'échantillon est un sous-groupe de la population, composé des individus pour lesquels des observations ont été recueillies. Si des mesures ont été prises pour tous les individus de la population, on parle de recensement.

Variable : Le terme variable désigne la représentation d'une caractéristique des individus. Ainsi, une variable n'est pas la caractéristique elle-même, mais plutôt une mesure de cette caractéristique.

Observation : Une observation est l'ensemble des valeurs obtenues en mesurant des variables sur un individu de la population.

La figure 1 illustre les notions de population, d'individus et d'échantillon. Il est important de comprendre que l'on considère dans ce cours que l'échantillon est choisi au hasard. Il est donc aléatoire. Il pourrait être autre que celui réellement observé lors de la collecte des données. En principe, n'importe quel individu de la population aurait pu faire partie de l'échantillon. Ainsi, une valeur calculée à partir des observations d'un échantillon est une réalisation de ce que l'on appelle en statistique une « variable aléatoire ». C'est le caractère aléatoire de l'échantillon qui permet de calculer des erreurs-types associées

aux estimations, de faire des tests d'hypothèses, de construire des intervalles de confiance.

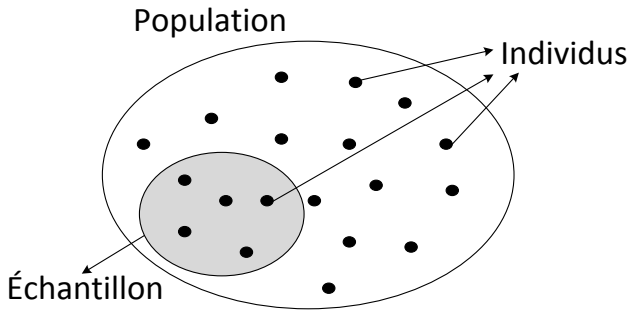


FIGURE 1 – Illustration d'une population, des individus qui la composent et d'un échantillon d'individus tiré de la population.

Lorsque l'on estime des paramètres de la population à l'étude à partir des observations d'un échantillon aléatoire de cette population, on fait de l'« inférence statistique ».

Pour approfondir la notion de variable

Types de variables

Les variables rencontrées en statistique sont de deux types, déterminés par l'ensemble des valeurs ou modalités possibles de la variable : catégorique (qualitative) ou numérique (quantitative).

Variables catégoriques : Les valeurs possibles d'une variable catégorique sont des catégories. Chacune des valeurs de la variable appartient à l'une des catégories. Par exemple, le sexe d'une personne est une variable catégorique, elle peut seulement prendre les valeurs 'homme' ou 'femme'. Les modalités d'une variable catégorique ne peuvent être mesurées numériquement, bien qu'elles puissent être représentées par des chiffres. Par exemple, un niveau de satisfaction peut être représenté par un chiffre de 1 à 5 alors que le chiffre 1 représente la catégorie « très insatisfait », « insatisfait » est représenté par le chiffre 2 et ainsi de suite.

On peut même définir deux types de variables catégoriques : nominales et ordinales. Les variables catégoriques **nominales**, tel le sexe d'une personne, ont des catégories qui ne suivent pas un ordre naturel. À l'opposé, les variables catégoriques dont les modalités peuvent être classées dans un certain ordre, tel un niveau de satisfaction, sont dites **ordinales**.

Variables numériques : Les variables numériques, telles que leur nom l'indique, peuvent être mesurées numériquement. Elles se subdivisent en deux types : discrètes et continues. Une variable **discrète** ne peut prendre qu'un nombre fini de valeurs, ou bien une infinité de valeurs si ces valeurs peuvent s'écrire sous la forme d'une suite $a_1, a_2, a_3, \ldots$. Ainsi une variable dont les valeurs sont 0, 1/5, 2/5, 3/5, 4/5, 1 est discrète ; il en est de même d'une variable pouvant prendre comme valeurs tous les entiers non négatifs 0, 1, 2, 3, ... Le résultat du lancer d'un dé est une variable discrète (valeurs $1, 2, 3, 4, 5, 6$), de même que le nombre de personnes frappées par la méningite dans une grande ville sur une période d'une année (valeurs $0, 1, 2, 3, \ldots$).

Une variable numérique est dite **continue** si elle peut prendre comme valeurs tous les points d'un intervalle de nombres réels (par exemple

$[0, 10]$). Des variables telles le temps, le poids ou la taille sont le plus souvent considérées continues, même si en pratique on n'observe qu'un nombre fini de valeurs de ces variables en raison de la précision limitée des instruments de mesure.

ATTENTION, *des données numériques ne sont pas forcément des observations d'une variable numérique.* C'est le cas par exemple du niveau de satisfaction représenté par un nombre de 1 à 5 décrit précédemment. Dans un jeu de données, les valeurs observées de certaines variables sur des individus sont saisies. Il n'est pas rare que des codes numériques soient utilisés pour représenter des modalités catégoriques de variables, car ils sont plus rapides à écrire ou taper que des chaînes de caractère.

Pour des variables catégoriques ordinales, ce code numérique s'avère parfois être un bon score pour représenter la variable. Ce score pourrait permettre un traitement numérique de la variable (voir ci-dessous). C'est probablement le cas pour le niveau de satisfaction.

Cependant, si la variable est catégorique nominale, ses modalités ne sont pas quantifiables. Mis à part un code numérique 0 et 1 pour une variable binaire, le code numérique ne peut être traité numériquement dans les analyses statistiques pour ce type de variable.

Ajoutons ici une précision concernant les variables mesurant un « niveau » ou un « degré » (par exemple un niveau de satisfaction, un niveau d'appréciation, un niveau de douleur, un niveau de fatigue, etc.). Selon la formulation de la question servant à mesurer la variable, elle sera de type catégorique ordinale ou numérique discrète. Si le choix de réponse est une série de libellés représentant différents niveaux (par exemple « très insatisfait », « insatisfait », « neutre », « satisfait », « très satisfait »), il s'agira d'une variable catégorique ordinale. Cependant, si la question est plutôt formulée de la façon suivante :

> « Sur une échelle de 1 à 5, 1 représentant le plus faible niveau et 5 le plus élevé, quel est votre niveau de ... »,

alors la variable mesurée sera de type numérique discrète.

Les deux variables ne sont pas identiques puisque l'utilisation de libellés laisse plus de place à l'interprétation. Chaque individu répondant à la question peut imaginer différemment la distance entre les modalités. La figure 2 vise à illustrer cet énoncé. Avec l'échelle numérique de 1 à 5, il est

clair que la distance entre toutes les modalités vaut 1. Avec les libellés, les individus imagineront probablement aussi une échelle avec des sauts égaux (comme l'individu 1 de la figure 2). Cependant, ceux qui mènent l'étude ne contrôlent pas l'interprétation que font les individus des libellés. Un individu peut considérer que la mention « satisfait » se rapproche plus du « très satisfait » que du niveau neutre. Ce serait le cas de l'individu 2 dans la figure 2. À l'inverse, un autre individu pourrait juger que la mention « satisfait » est plus similaire à un niveau neutre que de la mention « très satisfait » (individu 3 dans la figure 2).

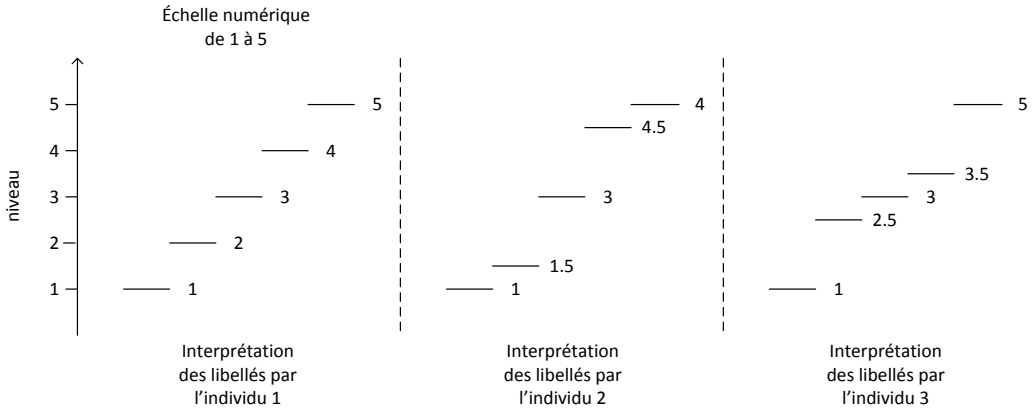


FIGURE 2 – Différentes interprétations possibles d'un niveau évalué par une question avec des choix de réponse sous forme de libellés.

On peut s'imaginer les différents niveaux comme des marches d'escalier. Les répondants peuvent avoir des conceptions différentes des hauteurs de ces marches. C'est pour cette raison que créer un score à partir d'une variable catégorique ordinale est une tâche difficile et subjective.

Représentation d'une caractéristique

Un jeu de données résultant d'une collecte de données contient des variables observées. Ces variables représentent des caractéristiques des individus d'une population, mais pas toujours sous le format reproduisant le mieux la réalité. Voici un exemple pour illustrer cet énoncé :

Exemple : variable pour représenter l'âge

L'âge d'une personne est naturellement représenté par une variable numérique. Cependant, dans certains sondages, par souci de confidentialité, cette variable est catégorisée. Ainsi, l'individu qui répond au sondage doit sélectionner, par exemple, une des classes suivantes : moins de 15 ans, 15 à 24 ans, 25 à 44 ans, 45 à 64 ans, 65 ans et plus. Alors la variable âge se retrouve dans le jeu de données sous un format catégorique ordinal et non numérique.

Il y a d'un côté la réalité et de l'autre côté la façon choisie pour mesurer cette réalité, en considérant toutes sortes de contraintes (instruments de mesure, confidentialité, budget, temps, etc.). Entre les deux, on pourrait imaginer une « variable théorique » qui collerait le plus possible à la réalité, comme sur la figure 3. Il faut cependant garder en tête que la variable observée ne correspond pas toujours à la variable théorique.

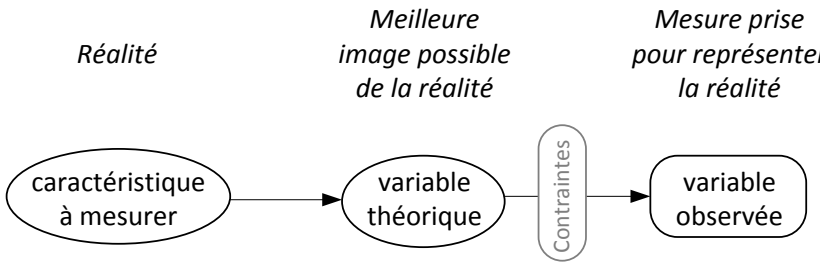


FIGURE 3 – Représentation d'une caractéristique par une variable.

Type de traitement statistique d'une variable

Dans une étude, lors du processus de collecte des données, des variables sont créées. Viendra ensuite le traitement statistique de ces variables. On peut classer les traitements statistiques en deux grands types, encore une fois catégorique et numérique.

ATTENTION, *le type du traitement d'une variable n'est pas nécessairement le même que le type d'origine de la variable*, comme en fait foi l'exemple suivant :

Exemple : variable pour représenter l'âge ... suite

Nous disposons d'une variable représentant l'âge, de type catégorique ordinal, prenant les modalités suivantes : moins de 15 ans, 15 à 24 ans, 25 à 44 ans, 45 à 64 ans, 65 ans et plus. Pour les analyses statistiques, nous souhaitons cependant traiter numériquement l'âge afin de l'intégrer facilement comme variable explicative dans un modèle de régression. Nous allons donc transformer la variable âge observée afin de la rendre de type numérique discrète. Pour ce faire, les observations catégoriques seront remplacées par des scores représentant les classes d'âge (par exemple 10, 20, 35, 55, 75, en respectant l'ordre des classes énoncé ci-dessus).

Ainsi, il arrive souvent que des variables observées soient transformées à postériori afin de les convertir d'un type à un autre avant leur traitement statistique. Pour les variables numériques discrètes, sans changer le format de saisie des observations, on peut les traiter de façon numérique ou catégorique.

Il y a donc une distinction à faire entre le type d'origine d'une variable et le type du traitement statistique de cette variable. La figure 4 présente les différentes combinaisons possibles de types de variable versus types de traitement statistique.

On peut choisir de traiter de façon catégorique les variables suivantes :
- toute variable catégorique ;
- une variable numérique discrète comprenant peu de modalités ;
- une variable numérique catégorisée en regroupant ses valeurs possibles en un petit nombre de classes.

ATTENTION, *traiter une variable numérique de façon catégorique est rarement un bon choix, car on perd de la puissance dans les tests statistiques.*

On peut choisir de traiter de façon numérique les variables suivantes :
- toute variable numérique ;
- une variable catégorique ordinale que l'on représente par un score numérique (ATTENTION, *rappelons que le choix d'un score est subjectif, il doit bien coller à la réalité pour conserver la validité des résultats*) ;
- une variable catégorique nominale binaire pour laquelle on attribue la valeur 1 à une catégorie et 0 à l'autre.

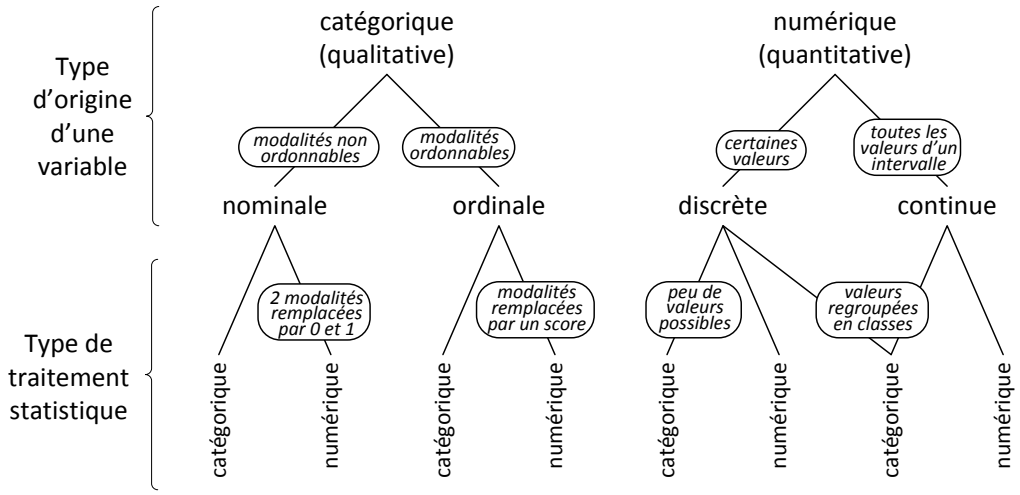


FIGURE 4 – Types de variable versus types de traitement statistique.

Variable dérivée pour le traitement statistique

Pour faire suite aux distinctions faites dans cette section, on se retrouve avec des variables observées d'un certain type, traitées statistiquement selon un type qui n'est pas nécessairement le même que le type de la variable observée. Afin de ne pas alourdir inutilement le texte, il est judicieux ici de définir une nouvelle catégorie de variables, celles utilisées pour faire les calculs statistiques. Nous allons donc dire qu'une variable est dérivée de la variable observée. Cette variable est définie de façon à avoir le bon format pour la méthode statistique choisie. La figure 5 positionne cette variable dans le schéma précédent de représentation d'une caractéristique.

Un statisticien a d'abord en main une variable observée (jeu de données d'origine). Ensuite, un choix est fait quant à la méthode statistique qui sera employée pour répondre à la question de recherche. Cette méthode peut traiter la variable observée de façon catégorique ou numérique. Le statisticien doit d'abord manipuler les données afin de créer, à partir de la variable observée, une « variable dérivée » qui est du type correspondant au traitement

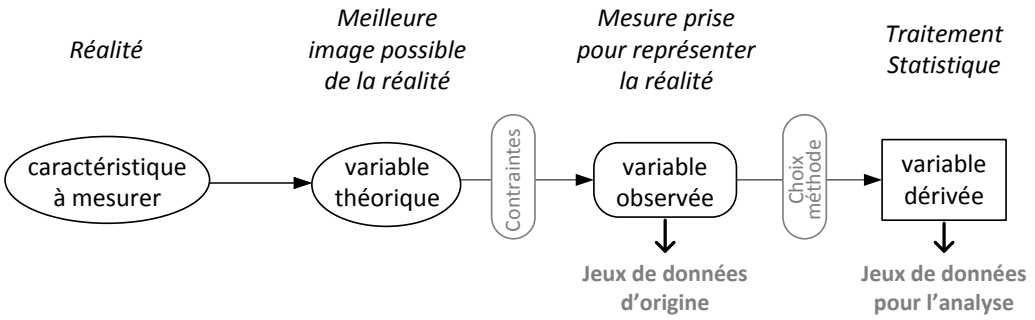


FIGURE 5 – Représentation d'une caractéristique par une variable : de la réalité jusqu'au traitement statistique.

statistique à effectuer. Ce travail doit être fait pour toutes les variables impliquées dans le traitement statistique. De plus, on essaie souvent plusieurs traitements statistiques sur des données. Si un autre traitement à faire demande un autre format pour une certaine caractéristique, on devra créer une autre variable dérivée. Ainsi, plusieurs variables peuvent être dérivées d'une même variable observée. Cependant, la variable observée est parfois déjà du bon type pour le traitement statistique. Dans ce cas, la variable dérivée est en fait une copie de la variable observée.

Lors d'une analyse statistique, la notion de « jeu de données » peut donc référer à deux entités différentes. Il est bon de distinguer le jeu de données d'origine, du jeu de données pour un traitement statistique. Le jeu de données d'origine contient les variables observées. Il est le produit du processus de collecte des données. Ensuite, ces données sont manipulées pour créer les variables dérivées, qui forment elles un jeu de données pour l'analyse (voir figure 5). Plusieurs jeux de données pour analyse peuvent être créés à partir d'un jeu de données d'origine.

Ainsi, lors du traitement statistique, on manipule des variables dérivées. Étant donné que dans ce cours on présente des méthodes statistiques de traitement de données, le terme variable fera toujours référence à une variable dérivée. L'expression « variable catégorique » reviendra souvent dans ce document. Elle fera référence à une variable dérivée pour l'analyse qui est de type catégorique.

Rappels concernant les tests d'hypothèses

Un test d'hypothèses est une méthode d'inférence statistique permettant d'évaluer, à partir d'observations, si une hypothèse statistique sur une population est ou non acceptée. Il nous permet d'arriver à une conclusion à partir de données.

Un test d'hypothèses confronte deux hypothèses : une nommée **hypothèse nulle** et notée H_0 et l'autre nommée **hypothèse alternative** et nommée H_1. Souvent, l'hypothèse nulle représente le statu quo, c'est-à-dire pas de changement par rapport à l'état actuel des connaissances, et l'hypothèse alternative représente un changement. Ainsi, l'hypothèse alternative est typiquement l'hypothèse de recherche de l'expérimentateur. La première étape d'un test d'hypothèses est de formuler ces deux hypothèses.

Le test mènera au rejet ou au non-rejet de H_0 ou, en d'autres mots, à l'acceptation ou à la non-acceptation de H_1. Cette décision sera prise à partir de données échantillonnales. Ces données interviennent dans le calcul de la valeur observée d'une statistique, la **statistique de test**, que nous noterons dans cette section W. Tout test d'hypothèses se base sur une statistique W dont on connaît la loi sous l'hypothèse nulle H_0. Si la valeur observée de W, notée w, tombe dans une certaine région dite critique, on rejettera H_0. Nous verrons plus loin comment définir cette région critique.

Caractéristiques des tests d'hypothèses

On utilise parfois les adjectifs suivants pour décrire les tests d'hypothèses :

bilatéral ou unilatéral : Si l'hypothèse alternative H_1 comporte une direction particulière (voir exemples ci-dessous), le test est dit « unilatéral ». Si au contraire l'hypothèse alternative est le complément de l'hypothèse nulle, on qualifie le test de « bilatéral ».

exact ou asymptotique : Si la distribution sous H_0 de W, la statistique du test, est vraie peu importe la taille de l'échantillon, le test peut être qualifié d' « exact ». À l'inverse, si la distribution sous H_0 de W utilisée pour le test est vraie seulement lorsque la taille de l'échantillon est grande, on parlera d'un test « asymptotique ».

paramétrique ou non paramétrique : Un test est « paramétrique » si la distribution de la statistique de test repose sur un postulat quant à

la distribution des observations. Sinon, il est « non paramétrique ». Dans ce cas, on doit typiquement seulement postuler l'indépendance des observations.

La validité des résultats d'un test dépend grandement de la validité des postulats émis. Un test non paramétrique nécessite moins de postulats qu'un test paramétrique, ce qui représente un avantage. Cependant, les tests non paramétriques sont moins puissants (voir définition plus loin) que les tests paramétriques. Ils rejettent donc moins souvent H_0. Pour cette raison, à moins d'évidences fortes contre des postulats, les tests paramétriques sont souvent préférés. Cependant, pour de petits échantillons, lorsque les lois asymptotiques des statistiques de test ne sont pas fiables, ce sont plutôt les tests non paramétriques qui sont préférés.

Note : Le qualificatif non paramétrique pour un modèle ne signifie pas qu'aucune distribution n'est postulée pour les résidus du modèle. Le sens de « non paramétrique » n'est pas le même pour un modèle et pour un test. Un modèle est dit non paramétrique si sa structure n'est pas fixe. Plutôt que d'être composé de paramètres fixes, le modèle s'adapte aux données (par exemple comme le fait un histogramme).

Types de tests d'hypothèses

Voici quatre types de tests d'hypothèses usuels. Ces types sont définis en fonction de l'objectif des tests.

Les **tests de conformité** consistent à tester si un paramètre ou un vecteur de paramètres $\boldsymbol{\theta}$ est égal à un vecteur de valeurs préétablies $\boldsymbol{\theta}_0$. Par exemple, si un seul paramètre θ est testé, les hypothèses d'un test de conformité sont formulées comme suit :

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \text{ au choix } \begin{cases} \theta \neq \theta_0 & \text{ou} & & \text{test biltéral} \\ \theta > \theta_0 & \text{ou} & \theta < \theta_0 & \text{test unilatéral.} \end{cases}$$

Si $\boldsymbol{\theta}$ est de dimension supérieure à 1, il est compliqué de formuler une hypothèse alternative incluant une direction. Dans ce cas, les tests bilatéraux sont plus simples. Un test de conformité usuel, qui est vu dans ces notes de cours, est un test pour vérifier si une proportion prend la valeur 0.5.

Les **tests d'adéquation ou d'ajustement** (en anglais goodness of fit tests) permettent de tester si des observations s'ajustent bien à un certain modèle ou à une certaine distribution. Par exemple, le test F global en régression et en ANOVA teste l'ajustement du modèle, le test de Shapiro-Wilk teste l'adéquation d'observations à une loi normale. Les hypothèses d'un test d'adéquation sont typiquement formulées comme suit :

$$H_0 : \text{le modèle s'ajuste bien aux données}$$
$$H_1 : \text{le modèle ne s'ajuste pas bien aux données}$$

ou encore :

$$H_0 : \text{les données suivent une loi } \mathcal{L}$$
$$H_1 : \text{les données ne suivent pas une loi } \mathcal{L}$$

Ainsi formulé, il s'agit d'un test bilatéral.

Les **tests d'homogénéité (ou de comparaison)** visent à vérifier si 2 échantillons ou plus proviennent d'une même population. Les tests de comparaison de moyennes en sont des exemples. Si un seul paramètre est comparé et qu'on a seulement deux échantillons, les hypothèses sont formulées ainsi :

$$H_0 : \theta_1 = \theta_2 \quad \text{versus} \quad H_1 : \text{ au choix } \left\langle \begin{array}{llll} \theta_1 \neq \theta_2 & \text{ou} & & \text{test biltéral} \\ \theta_1 > \theta_2 & \text{ou} & \theta_1 < \theta_2 & \text{test unilatéral.} \end{array} \right.$$

Les **tests d'association**, en particulier les **tests d'indépendance**, servent à tester la présence d'un lien entre deux variables. Rappelons que le terme association est plus large que le terme dépendance. Voici comment les hypothèses d'un tel test peuvent être formulées :

$$H_0 : X \text{ et } Y \text{ ne sont pas associées}$$
$$H_1 : \text{ au choix } \left\langle \begin{array}{ll} X \text{ et } Y \text{sont associées} & \text{test biltéral} \\ \left.\begin{array}{l} X \text{ est associé positivement à } Y \\ X \text{ est associé négativement à } Y \end{array}\right\} & \text{test unilatéral.} \end{array} \right.$$

Un test d'association courant consiste à vérifier si un coefficient de corrélation ou encore un ou des paramètres d'un modèle sont nuls. Si on note ρ la corrélation entre X et Y, les hypothèses précédentes peuvent être formulées de façon équivalente comme suit :

$$H_0 : \rho = 0 \quad \text{versus} \quad H_1 : \text{ au choix } \left\langle \begin{array}{llll} \rho \neq 0 & \text{ou} & & \text{test biltéral} \\ \rho > 0 & \text{ou} & \rho < 0 & \text{test unilatéral.} \end{array} \right.$$

Définitions relatives aux tests d'hypothèses

Voici les définitions de certains termes relatifs aux tests d'hypothèses.

Ces termes seront illustrés à l'aide d'un test usuel : le test sur la moyenne d'une distribution normale de variance connue. Les hypothèses de ce test sont les suivantes :

$$H_0 : \mu = \mu_o \qquad \text{versus} \qquad H_1 : \mu > \mu_0$$

où μ_0 est une valeur préétablie qui provient d'études antérieures. Il s'agit d'un exemple de *test de conformité*. On suppose que l'échantillon X_1 à X_n est composé de variables aléatoires indépendantes et identiquement distribuées, d'espérance μ inconnue et de variance σ^2 connue. Si n est grand, ces postulats sont suffisants pour mener le test. Cependant, si n est petit, il faut aussi postuler que la distribution des données est normale.

La statistique de ce test est $W = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$. Sous H_0, cette statistique suit une distribution $\mathcal{N}(0, 1)$.

Pour se ramener aux caractéristiques de test vues précédemment, on peut affirmer que ce test est *unilatéral* puisque son hypothèse alternative comporte une direction. Il est aussi *paramétrique* puisque l'on postule que la variance de la distribution des observations est connue. Si on ne postule pas la normalité des observations et que l'on utilise le théorème limite central pour affirmer que la statistique de test suit une loi normale standard, le test est alors *approximatif*. Cependant, si la normalité des observations est postulée, le test est *exact*.

Région critique ou région de rejet : l'ensemble des valeurs possibles de la statistique de test W pour lesquelles H_0 doit être rejetée. On définit cette région critique en se basant sur le seuil du test, la forme de l'hypothèse alternative H_1 et bien sûr la distribution de W sous H_0. Le seuil du test détermine la taille de la région critique, alors que sa localisation sera déterminée conjointement par la forme de H_1 et la distribution de W sous H_0. Notons par le symbole R_c la région critique, on a donc :

$$\text{Règle de rejet de } H_0 : w \in R_c.$$

On verra plus loin comment définir cette région critique.

Pour le test unilatéral à droite sur la moyenne d'une distribution normale de variance connue, la région critique est définie par l'ensemble des valeurs supérieures à une certaine valeur critique m_c^Y. On verra plus tard comment déterminer la valeur exacte de m_c^Y. On a donc :

$$\text{Règle de rejet de } H_0 : w \geq m_c^Y.$$

La figure 6 présente cette région critique sur un graphique. Le graphique de gauche est sur l'échelle de la statistique de test et celui de droite est sur l'échelle de la moyenne $\bar{X}$. La valeur critique sur l'échelle de $\bar{X}$, notée μ_c, est obtenue en isolant $\bar{X}$ dans $m_c^Y = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$. On obtient $\mu_c = \mu_0 + m_c^Y \sigma/\sqrt{n}$. Le graphique sur l'échelle de $\bar{X}$ est introduit, car il simplifie la présentation des concepts suivants.

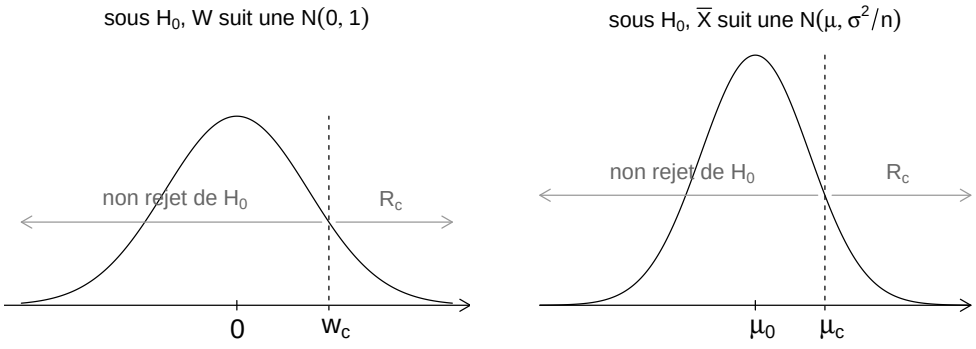


FIGURE 6 – Région critique pour le test unilatéral à droite sur la moyenne d'une distribution normale de variance connue (à gauche sur l'échelle de la statistique de test W, à droite sur l'échelle de la moyenne $\bar{X}$).

Erreurs de type I et II : Lorsque l'on effectue un test d'hypothèses, il y a deux types d'erreur que l'on est susceptible de commettre. Le tableau suivant présente ces erreurs.

		Réalité	
		H_0 **vraie**	H_0 **fausse**
Décision	**rejeter** H_0	erreur de type I	décision correcte
	ne pas rejeter H_0	décision correcte	erreur de type II

Dans l'exemple du test unilatéral à droite sur la moyenne d'une distribution normale de variance connue, H_0 fausse, ou en d'autres mots H_1 vraie, signifie que μ prend une certaine valeur μ_1 supérieure à μ_0. On peut visualiser les régions d'erreurs de type I et II sur la figure 7.

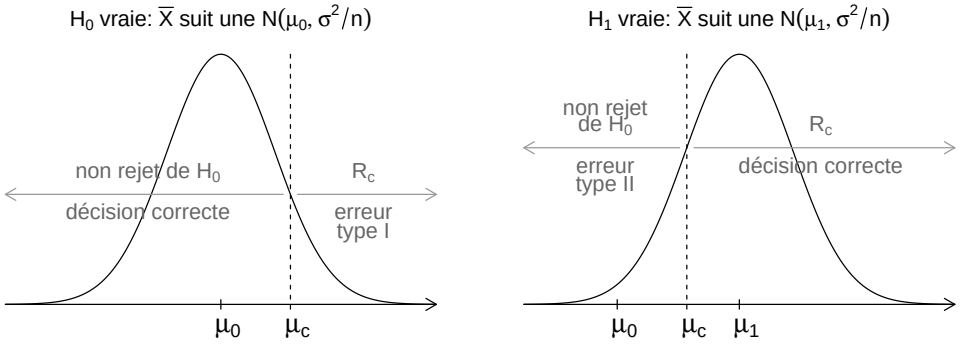


FIGURE 7 – Régions d'erreurs de type I et II pour le test unilatéral à droite sur la moyenne d'une distribution normale de variance connue (graphiques sur l'échelle de la moyenne $\bar{X}$).

Seuil ou niveau de signification : probabilité de commettre une erreur de type I :

$$P(\text{rejeter } H_0 \mid H_0 \text{ est vraie}).$$

Cette valeur est usuellement notée α.

Puissance : probabilité de ne pas commettre une erreur de type II :

$$P(\text{rejeter } H_0 \mid H_0 \text{ est fausse}).$$

On note usuellement β la probabilité d'erreur de type II et
$\beta = P(\text{ne pas rejeter } H_0 \mid H_0 \text{ est fausse}) = 1 - P(\text{rejeter } H_0 \mid H_0 \text{ est fausse})$.
Ainsi, on note souvent la puissance $1 - \beta$.

Dans l'exemple du test unilatéral à droite sur la moyenne d'une distribution normale de variance connue, nous allons réunir les deux courbes de loi de la figure 7 en un seul graphique afin d'illustrer le seuil et la puissance du test. Le seuil et la puissance étant des probabilités, il s'agit d'aires sous les courbes de densité. Le conditionnement dans la probabilité détermine la courbe de densité utilisée.

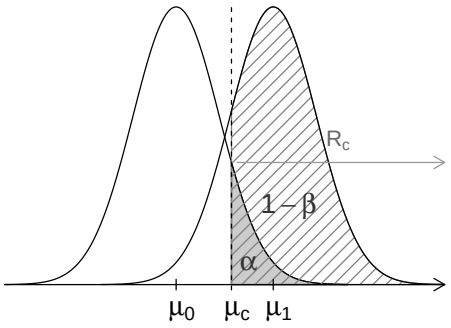


FIGURE 8 – Seuil α (aire de la région grise) et puissance $1-\beta$ (aire de la région hachurée) pour le test unilatéral à droite sur la moyenne d'une distribution normale de variance connue (graphique sur l'échelle de la moyenne $\bar{X}$).

Seuil observé (en anglais p-value) : probabilité, sous H_0, d'obtenir un résultat égal ou plus extrême que celui observé. L'hypothèse alternative H_1 dicte si le qualificatif extrême est attribué aux petites valeurs de la statistique de test W (test unilatéral à gauche), aux grandes valeurs (test unilatéral à droite) ou encore aux grandes et aux petites valeurs (test bilatéral). La loi de la statistique du test W sous H_0 dicte pour sa part à partir de quelle(s) valeur(s) critique(s) une valeur est dite extrême. Le terme « plus extrême » signifie en fait « moins probable » selon la distribution de W sous H_0.

La décision de rejeter ou non H_0 dans un test d'hypothèses peut être prise à partir d'une région critique, mais aussi à partir du seuil observé selon la règle suivante :

Règle de rejet de H_0 : seuil observé $\leq \alpha$.

Le seuil observé d'un test unilatéral à droite sur la moyenne d'une distribution normale de variance connue se calcule par la formule :

$$P(W \geq w),$$

car un résultat plus extrême serait une valeur observée encore plus grande. La figure 9 illustre comment le seuil observé se compare au seuil du test. Le seuil observé est l'aire de la région hachurée horizontalement. On voit clairement que si le seuil observé est supérieur au seuil du test α, c'est que w n'est pas dans la région critique. On ne rejette donc pas H_0. Cependant, si le seuil observé est égal ou inférieur à α, w est dans la région critique et on rejette H_0.

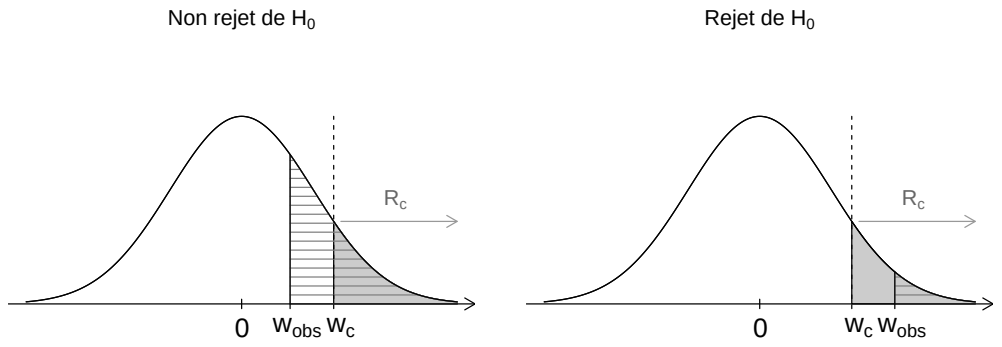


FIGURE 9 – Deux exemples de seuils observés pour le test unilatéral à droite sur la moyenne d'une distribution normale de variance connue (graphiques sur l'échelle de la statistique de test W).

Avec le seuil du test, on contrôle la probabilité d'erreur de type I. On considère souvent qu'il s'agit de l'erreur la plus grave à commettre. Cependant, la probabilité d'erreur de type II, c'est-à-dire de ne pas arriver à rejeter

H_0 même si H_0 est fausse, n'est pas contrôlée. Avec un test peu puissant, la probabilité de commettre une erreur de type II est forte. On dit parfois d'un test peu puissant qu'il est conservateur. Au contraire, un test très puissant a tendance à mener souvent au rejet de H_0. Il détecte comme significative la moindre déviance par rapport à l'hypothèse nulle H_0.

La puissance d'un test est fonction de la distribution de W sous H_1. Celle-ci dépend de l'ampleur de l'effet supposé ainsi que du nombre d'observations n utilisées pour réaliser le test. Plus l'effet supposé est grand, plus la puissance calculée sera grande. De même, plus n est grand, plus la dispersion de la distribution de W sous H_1 est petite, ce qui augmente la puissance du test.

Pour illustrer ces relations entre la puissance, l'ampleur de l'effet ainsi que la taille de l'échantillon, ramenons-nous une fois de plus à l'exemple d'un test unilatéral à droite sur la moyenne d'une distribution normale de variance connue. Sur la figure 8, on peut imaginer un effet plus grand, c'est-à-dire une valeur de μ_1 plus grande. Cela aurait pour impact de déplacer la cloche de droite encore plus loin vers la droite. En conséquence la puissance, soit l'aire sous cette courbe à droite de μ_c, serait encore plus grande. Une augmentation de n aurait quant à elle pour impact de rendre la cloche plus haute, mais moins large, sans la déplacer latéralement. Encore une fois, l'aire sous la courbe dans la région critique serait agrandie.

Comment définir la région critique ?

La région critique est composée de valeurs peu probables de W selon sa distribution sous H_0. Lorsque le test est bilatéral, cette région est parfois composée de deux sous-ensembles disjoints, aux deux extrémités opposées des valeurs possibles de W, mais ce n'est pas toujours le cas. Par exemple, il est très usuel pour un test d'hypothèses d'avoir une statistique de test W suivant sous H_0 une loi normale standard. Posons-nous ici dans ce cas. Nous utiliserons la lettre Z pour représenter cette statistique de test :

$$Z \xrightarrow[\text{sous } H_0]{} \mathcal{N}(0,1)$$

Si le test est bilatéral, on peut aussi utiliser la statistique $U = Z^2$ pour mener le test. On a alors que (voir section A.2.3) :

$$U \xrightarrow[\text{sous } H_0]{} \chi^2_1$$

Sous H_0, la distribution de Z nous indique que cette statistique devrait prendre une valeur proche de zéro. Si sa valeur s'éloigne trop de zéro, on peut douter de la véracité de H_0. Pour un test bilatéral, les valeurs de Z qui nous pousseraient à rejeter H_0 au profit de H_1 seraient les valeurs fortement positives ou négatives. Ainsi, la région critique comprendrait deux sous-régions, que l'on choisit habituellement de tailles égales. La figure 10 montre ces régions. Cependant, avec la statistique U, les valeurs fortement négatives de Z deviennent des valeurs fortement positives. Il n'y a plus de valeurs négatives possibles à cause de l'élévation au carré. La région critique du test, même s'il est bilatéral, est donc composée d'une seule région pour la statistique U, comme on le voit sur la figure 10 .

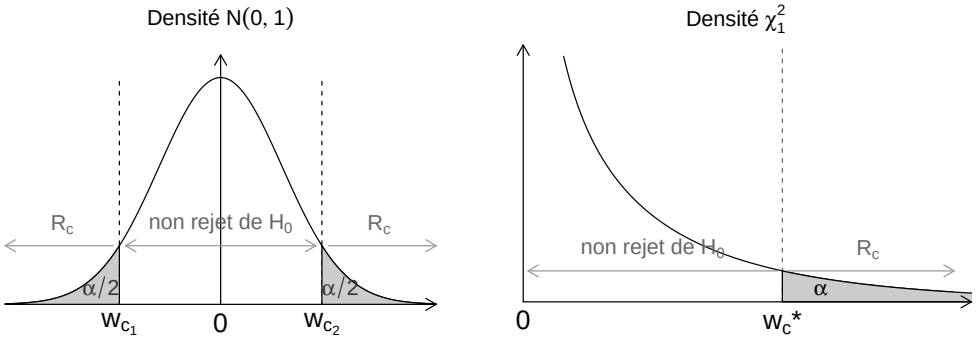


FIGURE 10 – Régions de rejet pour un test bilatéral de seuil α selon la densité de la statistique de test.

Les régions critiques sont délimitées par des **valeurs critiques**. Pour déterminer ces valeurs, on se ramène simplement aux distributions des statistiques de test sous H_0. Par exemple, pour trouver les valeurs de z_{c_1} et z_{c_1}, on doit résoudre les égalités suivantes :

$$P(Z \leq z_{c_1}) = \alpha/2 \qquad \text{et} \qquad P(Z \geq z_{c_2}) = \alpha/2.$$

On trouve que $z_{c_1} = \Phi^{-1}(\alpha/2)$ et $z_{c_2} = \Phi^{-1}(1-\alpha/2)$, donc z_{c_1} et z_{c_1} sont des quantiles de la loi normale standard. Étant donné que cette loi est symétrique en 0, on a que $z_{c_1} = -z_{c_2}$. Dans ce manuel, on notera :

$$z_\alpha = \Phi^{-1}(1-\alpha).$$

On a donc ici $z_{c_1} = -z_{\alpha/2}$ et $z_{c_2} = z_{\alpha/2}$.

Pour la statistique U qui suit sous H_0 une distribution χ^2_1, la valeur critique du test bilatéral est u_c. On trouve sa valeur en résolvant :

$$P(U \geq u_c) = \alpha.$$

On définit, tout comme pour les quantiles de la loi normale, le quantile de la loi khi-deux $\chi^2_{d,\alpha}$ par la valeur vérifiant l'équation :

$$P(\chi^2_d > \chi^2_{d,\alpha}) = \alpha$$

(comme dans la table C.2). Ainsi, on a dans notre exemple $u_c = \chi^2_{1,\alpha}$.

Étant donné qu'en élevant Z au carré on perd le signe de la statistique, on ne peut pas utiliser U pour effecteur un test unilatéral. Cependant, on peut le faire avec Z. Les régions critiques et leurs valeurs critiques correspondantes d'un test unilatéral à gauche et à droite sont présentées dans la figure 11

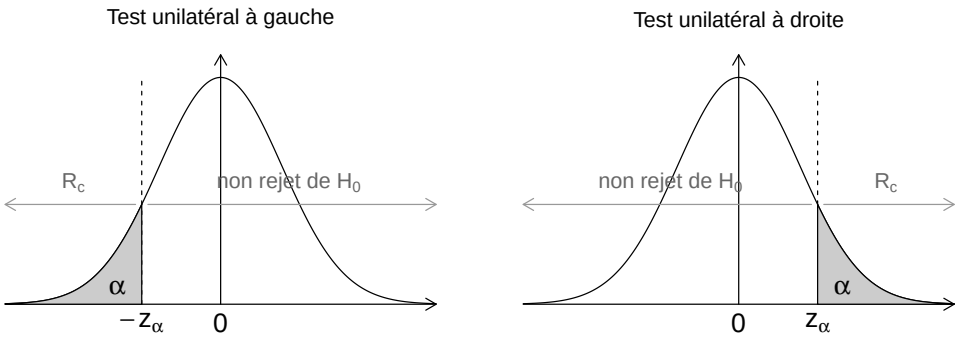


FIGURE 11 – Régions de rejet des tests unilatéraux de seuil α utilisant une statistique de test Z de loi $\mathcal{N}(0,1)$ selon la direction de l'hypothèse alternative.

On peut utiliser la région critique ou le seuil observé comme outil de décision dans un test d'hypothèses. Pour tous les cas traités précédemment (test bilatéral, unilatéral à gauche et à droite, statistique de test Z ou U), on retrouve dans le tableau 1 la définition du seuil observé.

H_1	comparaison à valeur critique	seuil observé (rejet si $\leq \alpha$)	
		notation 1	notation 2
bilatéral	$u \geq \chi^*_{1,\alpha}$ $\lvert z\rvert \geq z_{\alpha/2}$	$P(U \geq u)$ $2P(Z \geq \lvert z\rvert)$	$P(\chi^2_1 \geq u)$ $2P(\mathcal{N}(0,1) \geq \lvert z\rvert)$
unilatéral à gauche	$z \leq z_\alpha$	$P(Z \leq z)$	$P(\mathcal{N}(0,1) \leq z)$
unilatéral à droite	$z \geq z_\alpha$	$P(Z \geq z)$	$P(\mathcal{N}(0,1) \geq z)$

TABLE 1: Règles de rejet de H_0 selon l'hypothèse alternative du test, en utilisant la valeur observée d'une statistique $Z \xrightarrow[\text{sous } H_0]{} \mathcal{N}(0,1)$ ou $U \xrightarrow[\text{sous } H_0]{} \chi^2_1$.

Étapes de réalisation d'un test d'hypothèses

Voici une description des grandes étapes pour mener un test d'hypothèses.

1. Formuler l'hypothèse nulle H_0 et l'hypothèse alternative H_1.

2. Choisir un test approprié et définir une statistique W pour effectuer le test.

3. Déterminer la distribution de la statistique de test W sous H_0. Pour ce faire, il sera probablement nécessaire d'émettre des postulats concernant les données recueillies, par exemple de supposer l'indépendance des observations et aussi parfois de postuler qu'elles suivent une certaine distribution.

4. Choisir le seuil (ou niveau de signification) du test, noté α. Le seuil le plus usuel est 5%.

5. Établir la règle de décision de rejet pour ce test. Il y a deux outils possibles pour effectuer cette étape : une région critique ou un seuil observé.

   Région critique : Si on choisit de travailler avec une région critique, il faut d'abord définir cette région, notée R_c. On rejettera H_0 si w, la valeur observée de W, tombe dans cette région :

   Règle de rejet de H_0 : $w \in R_c$.

Seuil observé : Si on choisit plutôt d'utiliser un seuil observé, on doit définir la formule pour calculer cette valeur. On rejettera H_0 si le seuil observé est inférieur au seuil du test :

Règle de rejet de H_0 : seuil observé $\leq \alpha$.

6. Calculer à partir des données la valeur observée de la statistique de test, notée w ainsi que le seuil observé si on a choisi de travailler avec celui-ci.

7. Prendre une décision concernant l'hypothèse posée à partir de valeurs observées en suivant la règle établie précédemment, puis interpréter ce résultat, c'est-à-dire se ramener à la problématique pour expliquer ce que le résultat signifie.

Les deux façons de prendre la décision quant au résultat d'un test, soit avec une région critique ou un seuil observé, sont équivalentes : elles mènent à la même conclusion. Le seuil observé n'est pas calculable précisément avec une table de loi. Lorsqu'on ne peut pas utiliser un logiciel statistique (par exemple pendant un examen papier !), il est plus simple d'utiliser une région critique. Sinon, le seuil observé est plus souvent utilisé, car la règle de décision de rejet ou non de H_0 est très simple avec cet outil. Peu importe la distribution de W, le seuil observé doit toujours être comparé à la même valeur : le seuil du test. Une autre raison de la popularité du seuil observé est qu'il informe à lui seul de la force de la significativité. S'il est très petit (ex. ≤ 0.001), on rejette H_0 sans hésitation. Pour toutes ces raisons, le seuil observé est très courant dans les publications scientifiques.

Matière couverte dans ces notes de cours

Pour représenter les observations d'une variable catégorique, le principal outil est un tableau de fréquences, aussi appelé tableau de contingence. Ce tableau peut présenter une seule variable ou en croiser plusieurs. Ce manuel présente plusieurs tests et mesures associés à des tableaux de fréquences à une (chapitre 1), deux (chapitre 2) ou trois (chapitre 3) variables.

Ce manuel introduit aussi aux modèles linéaires généralisés, désignés par l'acronyme GLM (chapitre 4). Les GLM regroupent une grande variété de modèles, dont la régression linéaire classique et l'analyse de la variance. On s'intéresse cependant ici aux modèles linéaires généralisés permettant de traiter une variable réponse numérique discrète ou catégorique : la régression logistique binaire, conditionnelle, ordinale et multinomiale et la régression Poisson. Ainsi, ce manuel présente deux approches de traitement de variables catégoriques : l'approche par tests d'hypothèses (tableaux de fréquences) et l'approche par construction de modèles (GLM).

À quoi servent ces outils statistiques ?

La majorité des outils statistiques présentés dans ce manuel permettent de répondre à une question de recherche du type :

Quel est le lien entre les caractéristiques A et B des individus de la population à l'étude ?

Par exemple :

- Quel est le lien entre la couleur des cheveux et la couleur des yeux des Canadiens ?
- Est-ce que le niveau de scolarité est associé à la classe de revenu d'un individu ? Si oui, cette association est-elle positive ou négative ?
- Est-ce que le sexe d'une personne a une influence sur son risque de développer un cancer du poumon ? Si oui, de quelle façon ?

En termes statistiques, la question de recherche se traduit comme ceci :

Quel est le lien entre les variables X et Y ?

en supposant que X représente la caractéristique A et Y la caractéristique B.

Pour répondre à ces questions, il faut d'abord s'interroger sur l'existence ou non d'un lien (en d'autres mots d'une association). Les outils statistiques

présentés ici pour accomplir cette tâche sont des tests (tests d'association, tests sur les paramètres d'un modèle). Ensuite, si on arrive à la conclusion qu'un lien existe, on veut décrire ce lien. Pour ce faire, on utilise dans ce cours diverses mesures d'association ainsi que des paramètres de modèles.

Caractéristique de ces outils

Afin de choisir le bon outil statistique pour répondre à une question de recherche dans une situation donnée, il faut considérer les points suivants :

Nombre de variables impliquées

Lorsqu'on veut étudier le lien entre deux variables, on peut ne considérer que ces deux variables. C'est ce qu'on fait dans un tableau de fréquences à deux variables ou dans un modèle linéaire généralisé simple. Cependant, on possède parfois les observations d'autres variables, potentiellement liées aux variables étudiées et qui pourraient servir à corriger, pour la présence de cette variable, les statistiques calculées. On peut faire ça avec un tableau de fréquences à trois variables (ou plus) et avec un modèle linéaire généralisé multiple.

Type des variables impliquées

Les tableaux de fréquences traitent toutes les variables de façon catégorique. Certaines mesures associées à ces tableaux sont cependant propres aux variables catégoriques ordinales. Avec un GLM, la distribution de la variable réponse est choisie en fonction du type de cette variable. Les variables explicatives peuvent quant à elles être aussi bien numériques que catégoriques.

Direction de la relation

Certains de ces outils statistiques ne supposent aucune direction dans la relation entre les variables, donc ils ne supposent aucune causalité. C'est le cas des tableaux de fréquences. Toutes les variables ont le même intérêt dans un tableau de fréquences. On n'a pas à identifier une variable réponse et une variable explicative. On doit cependant le faire en régression logistique ou Poisson.

Rappelons ce qu'est une variable réponse versus une variable explicative :

variable réponse (ou dépendante) : Une variable dont les variations dépendent des variations d'autres variables (les variables explicatives). Cette variable est donc influencée par les autres.

variable explicative (ou indépendantee) : Une variable est dite explicative si elle influence une autre variable (la variable réponse).

Comparaison avec d'autres outils statistiques

Les outils présentés dans ce cours ne sont certes pas les seuls à pouvoir traiter des variables catégoriques. La première méthode statistique qui me vient à l'esprit lorsque je pense au traitement de variables catégoriques et l'analyse de la variance (ANOVA). Pour ce type de modèle, les variables explicatives, appelées dans ce cas facteurs, sont des variables catégoriques. Cependant, dans ce cours, on traite plutôt de méthodes d'analyse de données catégoriques pour lesquelles la variable réponse est catégorique ou numérique discrète (certains GLM) ou pour lesquelles il n'y a pas de direction supposée dans la relation entre les variables (tableaux de fréquences).

Aussi, le présent document traite de méthodes d'inférence statistique : estimation ponctuelle et par intervalle de confiance, tests et modèles. Cependant, quand le nombre de modalités des variables à l'étude est grand, il est difficile d'interpréter les résultats obtenus avec des méthodes inférentielles et les méthodes exploratoires multidimensionnelles de l'analyse de données sont toutes indiquées. L'analyse des correspondances binaire ou multiple (ACM) ainsi que l'analyse factorielle des correspondances (AFC) permettent de mettre en évidence les relations entre deux ou plusieurs variables catégoriques. Ces méthodes sont dites descriptives, elles ne testent pas la significativité des relations et ne supposent aucune direction dans les relations. Elles permettent de produire des graphiques exploratoires informatifs. Ces méthodes ne sont pas couvertes dans ce cours (voir Lebart *et al.*, 1997, pour plus d'information).

Le tableau 2 permet de comparer les méthodes statistiques couvertes dans ces notes de cours à quelques autres méthodes classiques que vous avez probablement déjà vues dans d'autres cours. Les méthodes sont comparées selon les caractéristiques décrites ci-dessus, soit le nombre de variables impliquées et le type de ces variables selon leur rôle (réponse, explicative ou ni un ni l'autre).

Classe de méthodes	**Méthode spécifique**		**Types des variables**		**Nombre de variables**
			réponse	**explicatives**	
Tableaux de fréquences	***divers tests et mesures d'association***		pas de direction supposée dans l'association		***1, 2, 3*** et plus → modèles loglinéaires
Analyse de données	ACM { méthodes } AFC { descriptives }		toutes catégoriques		2 → binaire 3+ → multiple
Modèles linéaires généralisés (GLM)	régression classique		numérique	numériques	2 → modèle simple 3 et plus → modèle multiple
	ANOVA			catégoriques	
	ANCOVA			numériques et catégoriques	
	régression Poisson		num. discrète		
	régression logistique	***binaire***	binaire		
		conditionnelle			
		ordinale	catégo. ordinale		
		multinomiale	catégo. nominale		

TABLE 2: Comparaison des méthodes présentées dans ce cours (en caractères italiques gras) avec quelques autres méthodes classiques couvertes dans d'autres cours.

Finalement, il existe aussi des graphiques pour représenter les observations de variables catégoriques, par exemple les diagrammes en secteurs, en bâtons, en mosaïque, etc. Quelques-uns de ces graphiques seront utilisés dans ce manuel pour illustrer les exemples, mais ils ne seront que très brièvement présentés d'un point de vue théorique (voir Friendly, 2000, pour plus d'information).

Indépendance entre les individus de l'échantillon

Notons aussi que les techniques vues dans ce cours supposent toujours l'indépendance entre les individus de l'échantillon. Si les individus de l'échantillon ont été sélectionnés par échantillonnage aléatoire simple, ce postulat est respecté. Cependant, il n'est pas rare que les données proviennent d'enquêtes utilisant un plan de sondage complexe. Dans ce cas, les individus n'ont pas des poids égaux dans l'échantillon. Le poids de sondage d'un individu de l'échantillon peut être interprété comme le nombre d'individus de la population entière qu'il représente. Ces poids servent à ajuster les formules afin de faire une inférence statistique correcte. Cependant, on n'apprendra pas comment calculer et utiliser des poids de sondage dans ce cours. Le cours gradué STT-7340 « Sondages : modèles et techniques » aborde ce problème. Une bonne référence à ce sujet est Lohr (2009).

Chapitre 1

Tableaux de fréquences à une variable : distributions utiles

Avant de présenter des méthodes qui permettent d'étudier le lien entre deux variables, voyons comment étudier une seule variable catégorique. Dans cette section, nous présenterons des outils permettant de répondre à la question de recherche suivante :

Quel est le portrait de la caractéristique A dans la population à l'étude ?
La reformulation statistique de cette question est la suivante :

De quoi ont l'air les observations de la variable Y ?
en supposant que la variable Y représente la caractéristique A.

Dans le présent chapitre, les différents éléments d'un tableau de fréquences univariées sont d'abord présentés. Quelques graphiques utilisables pour représenter un tel tableau sont mentionnés. Les sections suivantes sont consacrées à trois distributions parfois utilisées pour représenter ces données : les distributions Poisson, binomiale et multinomiale. On y voit comment, à partir de données, estimer ponctuellement et par intervalle de confiance, ainsi que tester les paramètres de ces lois.

1.1 Définitions et outils descriptifs

Soit $m_1^Y, \ldots, m_J^Y$ les modalités de la variable catégorique Y. La fréquence de la modalité m_j^Y, notée n_j, est le nombre d'observations dans l'échantillon prenant la valeur m_j^Y. On a alors $n = \sum_1^J n_j$ où n est le nombre total d'observations dans l'échantillon. Le tableau 1.1 est le tableau de fréquences univariées pour cette variable.

Modalité de Y	Fréquence observée
m_1^Y	n_1
$\vdots$	$\vdots$
m_j^Y	n_j
$\vdots$	$\vdots$
m_J^Y	n_J
Total	n

TABLE 1.1: Tableau de fréquences pour une variable catégorique Y.

On retrouve parfois, dans un tableau de fréquences univariées, des fréquences relatives. On définit la fréquence relative de la modalité m_j^Y comme étant la proportion

$$f_j = \frac{n_j}{n},$$

où l'on note que $\sum_1^J f_j = \sum_1^J n_j/n = 1$.

Si certaines modalités sont associées à de trop petites fréquences, il peut être avantageux de les regrouper. Les nouvelles modalités formées doivent rester interprétables bien sûr. Cette approche permet d'alléger la présentation des résultats. Par exemple, supposons qu'on étudie la variable : nombre d'enfants par famille dans la ville de Québec. On pourrait observer dans l'échantillon tous les entiers entre 0 et 10, mais les familles de 5 enfants et plus seraient probablement peu nombreuses. On pourrait alors calculer les fréquences pour les 6 catégories suivantes plutôt que les 11 d'origine : 0, 1, 2, 3, 4 ainsi que 5 et plus.

1.1.1 Différents formats de jeux de données

Lorsque l'on travaille avec des variables catégoriques, on ne possède pas toujours de jeu de données avec une ligne pour chaque individu. On a parfois seulement les fréquences comme dans le tableau 1.1. En fait, ces fréquences résument sans aucune perte d'information les données.

Il y a donc deux formats de données pour les observations d'une seule variable catégorique, illustrés dans le tableau 1.2. Nous les nommons « format individus » et « format fréquences ». Dans le format individus, le jeu de données comprend n lignes, alors qu'il en comprend J, soit le nombre de modalités possibles de la variable Y, dans le format fréquences. On utilisera l'indice u pour identifier les observations sous le format individus ($u = 1, \ldots, n$) et l'indice j pour celles sous le format fréquences ($j = 1, \ldots, J$).

Format individus :

Individu	Modalité de Y	
Id_1	m_1^Y	$\left.\begin{array}{c}\\ \\ \\ \end{array}\right\}$ n_1 lignes
$\vdots$	$\vdots$	
Id_{n_1}	m_1^Y	
$\vdots$	$\vdots$	$\left.\begin{array}{c}\\ \\ \\ \end{array}\right\}$ $n_2 + \cdots + n_{J-1}$ lignes
$\vdots$	$\vdots$	
Id_{n-n_J+1}	m_J^Y	$\left.\begin{array}{c}\\ \\ \\ \end{array}\right\}$ n_J lignes
$\vdots$	$\vdots$	
Id_n	m_J^Y	

Format fréquences :

Modalité de Y	Fréquence observée
m_1^Y	n_1
$\vdots$	$\vdots$
m_k^Y	n_J

TABLE 1.2: Formats possibles de données pour les observations d'une variable catégorique Y.

Les modalités m_1^Y à m_J^Y de Y sont parfois numériques. C'est le cas pour une variable numérique discrète traitée de façon catégorique. C'est aussi le cas pour une variable catégorique ordinale représentée par un score. Dans ce cas, on peut vouloir calculer des statistiques descriptives pour variables numériques, comme la moyenne et la variance échantillonnale. La formule pour calculer ces statistiques dépend du format des données, tel que présenté dans le tableau 1.3.

Statistique	Format individus	Format fréquences
moyenne ($\bar{y}$)	$\frac{\sum_{u=1}^{n} y_u}{n}$	$\frac{\sum_{j=1}^{J} m_j^Y n_j}{n}$
variance	$\frac{\sum_{u=1}^{n}(y_u-\bar{y})^2}{n-1} = \frac{\sum_{u=1}^{n} y_u^2 - n\bar{y}^2}{n-1}$	$\frac{\sum_{j=1}^{J}(m_j^Y)^2 n_j - n\bar{y}^2}{n-1}$

TABLE 1.3: Formules de statistiques descriptives numériques selon le format des données.

1.1.2 Graphiques

Les deux types de graphiques les plus utilisés pour représenter une variable catégorique sont le *diagramme en secteurs* (ou circulaire, ou en pointes de tarte) et le *diagramme en bâtons*.

Diagramme en secteurs

Dans un diagramme en secteurs, chaque modalité est représentée par un secteur circulaire dont l'angle est proportionnel à la fréquence de cette modalité. La figure 1.1 présente un exemple de diagramme en secteurs. Il est recommandé d'utiliser ce type de graphique avec parcimonie, car il a été prouvé que l'oeil est bon pour juger des mesures linéaires, mais mauvais pour comparer des aires. De plus, pour une variable comprenant un trop grand nombre de modalités, ce genre de graphique est illisible. Il demeure cependant utile dans certaines circonstances, notamment pour représenter de façon attrayante un petit nombre (2 ou 3) fréquences.

Diagramme en bâtons

Pour le diagramme en bâtons, chaque modalité est représentée par un rectangle (ou bâton) dont la hauteur ou la longueur est proportionnelle à la fréquence de cette modalité et dont la largeur est la même pour toutes les modalités. Les rectangles peuvent être placés verticalement ou horizontalement. De plus, les modalités de la variable peuvent être identifiées dans une légende indépendante ou directement sur un axe du graphique. La figure 1.4 est un exemple de diagramme en bâtons.

1.2 Expérience avec la loi Poisson

Exemple : armée prussienne

Un exemple historique pour représenter la distribution Poisson est celui concernant l'armée prussienne publié par Bortkiewicz en 1889 dans son livre sur la loi Poisson intitulé « La loi des petits nombres ». Ce jeu de données répertorie le nombre de soldats tués par ruade (coup de sabot donné par un cheval) au cours d'une année dans des corps de l'armée prussienne, qui devint l'armée impériale allemande. Le jeu de données se compose de 10 corps d'armée, étudiés pendant 20 ans (de 1875 à 1894). Étant donné qu'ici un individu de la population est en fait une année pour un corps d'armée, on a $n = 200$ individus. Les données sont les suivantes :

Nbre de soldats tués par année	0	1	2	3	4	5 ou plus	total
Fréquence observée	109	65	22	3	1	0	200

1.2.1 Rappel sur la loi Poisson

La loi de Poisson permet de modéliser le nombre de réalisations d'un événement dans un intervalle de temps et/ou d'espace. Par exemple, la loi Poisson serait une bonne candidate pour modéliser les dénombrements suivants :

- le nombre de clients se présentant à un guichet automatique d'une banque en une heure ;
- le nombre d'accidents par années à une intersection de la ville de Québec ;
- le nombre de centenaires dans une communauté.

On appelle parfois la loi Poisson « loi des évènements rares ».

La fonction de masse, aussi appelée fonction de probabilité, d'une variable aléatoire Poisson, notée $Y \sim Poisson(\lambda)$, est

$$P(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!} \qquad y = 0, 1, 2, \ldots$$

Le paramètre de la distribution est à la fois l'espérance et la variance de Y : $E[Y] = Var[Y] = \lambda$.

Les caractéristiques principales de la loi de Poisson sont les suivantes :

- pas de limite supérieure théorique aux valeurs prises par Y ;
- variance égale à la moyenne ;
- distribution plus asymétrique que la loi binomiale ;
- si Y représente le nombre d'évènements dans un intervalle de temps et/ou d'espace, alors λ est proportionnel à la longueur de l'intervalle. De plus, on suppose que les nombres d'évènements dans 2 intervalles disjoints sont indépendants.

Aussi, il est pertinent de noter que la distribution Poisson peut être approximée par la loi normale lorsque λ est grand. Dans ce cas, $Poisson(\lambda) \approx \mathcal{N}(\lambda, \lambda)$. Ainsi, pour modéliser le dénombrement d'événements pas très rares (nombre moyen d'occurrences λ assez élevé), il est inutile d'utiliser la loi Poisson. Dans un tel cas, la loi normale fait très bien l'affaire.

1.2.2 Estimation ponctuelle du paramètre λ

On peut estimer λ par la méthode du maximum de vraisemblance. On trouve :

$$\hat{\lambda} = \bar{Y} = \frac{\sum_{u=1}^{n} Y_u}{n}.$$

On peut calculer la moyenne et la variance de l'estimateur $\hat{\lambda}$. On trouve

$$\begin{aligned} E[\hat{\lambda}] &= E[\bar{Y}] \\ &= E\left[\sum_{u=1}^{n} \frac{Y_u}{n}\right] = \sum_{u=1}^{n} \frac{E[Y_u]}{n} = \frac{n\lambda}{n} = \lambda \text{ (Estimateur sans biais)}, \\ Var[\hat{\lambda}] &= Var[\bar{Y}] \\ &= Var\left[\sum_{u=1}^{n} \frac{Y_u}{n}\right] = \sum_{u=1}^{n} \frac{Var[Y_u]}{n^2} = \frac{n\lambda}{n^2} = \frac{\lambda}{n}. \end{aligned}$$

On peut finalement trouver la distribution asymptotique de l'estimateur $\hat{\lambda}$. Étant donné que $\hat{\lambda}$ est une moyenne de n variables indépendantes et identiquement distribuées (iid) $Poisson(\lambda)$, avec espérance λ et variance λ, le

théorème limite central dit que :

$$\frac{\hat{\lambda} - \lambda}{\sqrt{\lambda/n}} \xrightarrow{\text{asymp.}} \mathcal{N}(0,1).$$

Exemple : armée prussienne ... suite

La distribution de Poisson est en général une bonne candidate pour modéliser une variable aléatoire lorsque sa variance est voisine de sa moyenne. Calculons donc la moyenne et la variance échantillonnale des données de l'exemple de l'armée prussienne, en utilisant les formules pour des données présentées sous le « format fréquences » (voir section 1.1.1). Étant donné que les modalités possibles de la variable sont les nombres entiers non nuls $0, 1, 2, \ldots$, on remplace dans ces formules les modalités m_j^Y par j et on fait varier j de 0 à sa valeur maximum observée. On n'a pas besoin de faire varier j de 0 à l'infini puisque dans les formules les modalités sont multipliées par leurs fréquences, qui sont nulles pour toutes les valeurs supérieures à la valeur maximum observée.

$$\bar{y} = \frac{\sum_{j=0}^{4} j n_j}{n} = \frac{0 + 1 \times 65 + 2 \times 22 + 3 \times 3 + 4 \times 1}{200} = 0.61$$

$$\begin{aligned} s^2(y) &= \frac{\sum_{j=0}^{4} j^2 n_j - n\bar{y}^2}{(n-1)} \\ &= \frac{(0^2 + 1^2 \times 65 + 2^2 \times 22 + 3^2 \times 3 + 4^2 \times 1) - 200 \times 0.610^2}{199} \\ &= 0.611 \end{aligned}$$

Les deux statistiques sont pratiquement égales ici ! La loi Poisson semble donc une bonne candidate pour modéliser ces données.

1.2.3 Tests d'hypothèses sur le paramètre λ

Supposons que nous voulions confronter les hypothèses suivantes :

$$H_0 : \lambda = \lambda_0 \quad \text{versus} \quad H_1 : \text{ au choix } \left\langle \begin{array}{llll} \lambda \neq \lambda_0 & \text{ou} & & \text{test biltéral} \\ \lambda > \lambda_0 & \text{ou} & \lambda < \lambda_0 & \text{test unilatéral.} \end{array} \right.$$

Nous allons voir comment le faire avec un test de Wald, un test score et un test de rapport de vraisemblance (voir annexe A.4).

Ce test est une solution de remplacement au test classique sur une moyenne basé sur la loi normale. En effet, le paramètre testé ici, λ, est une espérance au même titre que le paramètre μ d'une loi $\mathcal{N}(\mu, \sigma^2)$. Il est plus approprié que le test basé sur la loi normale lorsque la variable étudiée prend des valeurs entières non nulles (tel un dénombrement).

Test de Wald sur le paramètre λ

Ce test se base sur la loi asymptotique de $\hat{\lambda}$ énoncée précédemment. Comme dans tout test de Wald, la variance de l'estimateur est estimée. Ainsi, dans le calcul de la variance de l'estimateur, λ est remplacé par son estimateur du maximum de vraisemblance. La statistique de ce test est la suivante :

$$Z_w = \frac{\hat{\lambda} - \lambda_0}{\sqrt{\hat{\lambda}/n}} \xrightarrow[H_0]{\text{asymp.}} \mathcal{N}(0,1).$$

Cette statistique de test permet d'effectuer un test bilatéral ou unilatéral.

Test score sur le paramètre λ

Le test score, contrairement au test de Wald, va utiliser l'hypothèse nulle $H_0 : \lambda = \lambda_0$ pour déterminer la variance de l'estimateur. La statistique de ce test est la suivante :

$$Z_s = \frac{\hat{\lambda} - \lambda_0}{\sqrt{\lambda_0/n}} \xrightarrow[H_0]{\text{asymp.}} \mathcal{N}(0,1).$$

Cette statistique de test permet elle aussi d'effectuer un test bilatéral ou unilatéral.

Test du rapport de vraisemblance sur le paramètre λ

Notons d'abord que l'annexe A.3 contient des rappels concernant les tests de rapport de vraisemblance. Les formules ci-dessous utilisent la même notation que dans l'annexe. La vraisemblance de paramètre λ est

$$L(\lambda|\boldsymbol{y}) = \prod_{u=1}^{n} \frac{e^{-\lambda}\lambda^{y_u}}{y_u!} = \frac{e^{-n\lambda}\lambda^{\sum_{u=1}^{n} y_u}}{\prod_{u=1}^{n} y_u!}.$$

Le logarithme naturel de cette valeur est :

$$\ln L(\lambda|\boldsymbol{y}) = -n\lambda + \ln(\lambda)\sum_{u=1}^{n} y_u - \ln\left(\prod_{u=1}^{n} y_u!\right).$$

La statistique du test du rapport des vraisemblances est donc :

$$\begin{aligned}
LR &= -2(\ln L(\lambda_0|\boldsymbol{Y}) - \ln L(\hat{\lambda}|\boldsymbol{Y})) \\
&= -2\left(-n\lambda_0 + \ln(\lambda_0)\sum_{u=1}^{n} Y_u - \ln\left(\prod_{u=1}^{n} Y_u!\right) + n\hat{\lambda} - \ln(\hat{\lambda})\sum_{u=1}^{n} Y_u + \ln\left(\prod_{u=1}^{n} Y_u!\right)\right) \\
&= -2\left(\ln\left(\frac{\lambda_0}{\hat{\lambda}}\right)\sum_{u=1}^{n} Y_u + n(\hat{\lambda} - \lambda_0)\right) \\
&= -2\left(n\hat{\lambda}\ln\left(\frac{\lambda_0}{\hat{\lambda}}\right) + n(\hat{\lambda} - \lambda_0)\right) \qquad \text{car} \quad \hat{\lambda} = \sum_{u=1}^{n}\frac{Y_u}{n}.
\end{aligned}$$

Sous H_0, cette statistique suit approximativement une loi du khi-deux à 1 degré de liberté. Le degré de liberté est 1, car on teste un seul paramètre et il n'y a aucun paramètre libre sous H_0 (on suppose que λ prend une valeur prédéfinie).

En résumé, la statistique de ce test de rapport de vraisemblance est :

$$LR = -2\left(n\hat{\lambda}\ln\left(\frac{\lambda_0}{\hat{\lambda}}\right) + n(\hat{\lambda} - \lambda_0)\right) \xrightarrow[H_0]{\text{asymp.}} \chi^2_1.$$

Cette statistique de test permet uniquement d'effectuer un test bilatéral.

1.2.4 Intervalle de confiance pour λ

Nous présentons ici uniquement l'intervalle de confiance le plus simple, celui de Wald :

$$\left[\hat{\lambda} - z_{\alpha/2}\sqrt{\frac{\hat{\lambda}}{n}}\ ,\ \hat{\lambda} + z_{\alpha/2}\sqrt{\frac{\hat{\lambda}}{n}}\right].$$

L'intervalle de confiance score peut être trouvé en exercice en s'inspirant de la démarche pour trouver l'intervalle de confiance score d'une proportion (voir section 1.3.4).

1.3 Expérience avec la loi binomiale

Exemple : opinion sur l'avortement

Les Américains sont-ils plutôt favorables ou défavorables à l'avortement ? Pour répondre à cette question, on observe un échantillon de 1223 Américains interrogés en 2010 dans le cadre de l'Enquête Sociale Générale aux États-Unis. Il s'agit d'une grande enquête qui existe depuis 1972 et qui possède plusieurs volets, dont certains internationaux (GSS, 2012). On a posé la question suivante aux participants : Pensez-vous qu'il devrait être possible pour une femme enceinte mariée qui ne veut plus d'enfants de se faire avorter légalement ? Au total, 587 personnes ont répondu oui à cette question, et 636 personnes ont répondu non. Ces réponses sont représentées dans le diagramme en secteurs de la figure 1.1.

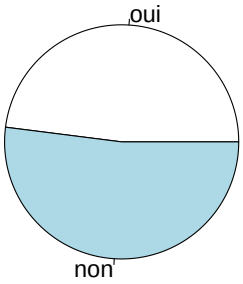


FIGURE 1.1 – Diagramme en secteurs des réponses dans l'exemple de l'opinion sur l'avortement.

1.3.1 Rappel sur la loi binomiale

Le contexte de la loi binomiale est le suivant. On fait une expérience qui peut prendre 2 résultats possibles : succès ou échec. On répète cette expérience de façon indépendante n fois, et on note S le nombre de succès obtenus. Notez que chaque expérience a la même probabilité de succès π. On dit alors que S suit une loi binomiale de paramètres n (nombre total d'essais) et π

(probabilité d'un succès). On note $S \sim Bin(n, \pi)$.

En résumé, les conditions pour qu'il y ait une expérience binomiale sont les suivantes :
- n essais ;
- deux résultats possibles pour chaque essai : succès et échec ;
- toujours la même probabilité de succès $= \pi$;
- indépendance entre chacun des n essais ;

Exemple : opinion sur l'avortement ... suite

Ici, les 4 conditions d'une expérience avec la loi binomiale sont respectées :
- $n = 1223$ essais $\rightarrow$ individus sondés ;
- 2 résultats possibles pour chaque essai : succès = oui, échec = non ;
- la probabilité de succès π est ici la proportion dans la population des Américains favorables à l'avortement ;
- on peut supposer qu'il y a indépendance entre chacun des essais puisque les individus participants à l'étude ont été sélectionnés avec un plan de sondage se rapprochant beaucoup de l'échantillonnage aléatoire simple.

La fonction de masse de la loi binomiale s'écrit de la façon suivante :

$$P(S = s) = \begin{cases} \binom{n}{s} \pi^s (1-\pi)^{n-s}, & \text{pour } s = 0, 1, \ldots, n; \\ 0 & \text{sinon.} \end{cases}$$

L'espérance et la variance de S sont :

$$\begin{aligned} E[S] &= n\pi \\ Var[S] &= n\pi(1-\pi). \end{aligned}$$

Notez que l'on peut approximer la loi binomiale par la loi normale lorsque n est grand en utilisant le résultat du Théorème Limite Central (voir annexe A.2.2). Pour ce faire, on peut adopter le point de vue suivant : observer une seule variable $Bin(n, \pi)$ revient à observer n variables $\mathbf{1}_1, \ldots, \mathbf{1}_n$ iid $Bin(1, \pi)$, soit la loi $Bernoulli(\pi)$. On a alors $S = \sum_{u=1}^{n} \mathbf{1}_u$, une somme de n variables iid $Bin(1, \pi)$. Donc :

$$S \xrightarrow{\text{asymp.}} \mathcal{N}(n\pi, n\pi(1-\pi)).$$

Notez aussi que si $n \to +\infty$, $\pi \to 0$ et $n\pi = \lambda$ constant, alors la loi $Bin(n, \pi)$ tend vers la loi Poisson. En d'autres mots,

$$\lim_{n\to\infty,\, \pi\to 0} P(S = s) = \frac{\lambda^s e^{-\lambda}}{s!}.$$

Correction pour la continuité

Rappelons qu'une correction pour la continuité peut améliorer l'approximation d'une binomiale par une loi normale lorsque n n'est pas grand. Par exemple, si on cherche à calculer $P(s_1 \leq S \leq s_2)$ en utilisant l'approximation normale avec une correction pour la continuité, on calculera en fait $P(s_1 - 1/2 \leq S \leq s_2 + 1/2)$. Les bornes des régions dont on calcule l'aire (car la probabilité est l'aire sous la courbe de densité) sont donc ajustées de $1/2$. Cet ajustement additionne ou soustrait $1/2$ selon que la borne soit inférieure ou supérieure, et selon que la borne soit incluse ou non dans la région. Plusieurs des tests et intervalles de confiance présentés ci-dessous peuvent aussi intégrer une correction pour la continuité.

1.3.2 Estimation ponctuelle d'une proportion π

On peut estimer le paramètre π par la méthode du maximum de vraisemblance (Casella et Berger, 2002, exemple 7.2.7). On trouve

$$\hat{\pi} = \frac{S}{n} = \frac{\text{Nombre de succès}}{\text{Nombre d'essais}}.$$

Dans certains ouvrages, on note cet estimateur p, mais ici nous le noterons $\hat{\pi}$. On peut calculer l'espérance et la variance de cet estimateur :

$$\begin{aligned} E[\hat{\pi}] &= \pi, \\ Var[\hat{\pi}] &= \frac{\pi(1-\pi)}{n}. \end{aligned}$$

Ainsi, pour estimer adéquatement π, comme pour tout paramètre d'une population, il faut observer un nombre assez grand de données. Sinon, l'estimateur est très imprécis (la variance de l'estimateur est grande).

On peut finalement trouver la distribution asymptotique de l'estimateur $\hat{\pi}$. Par le théorème limite central, encore en adoptant le point de vue que S est une somme de n variables iid $Bin(1, \pi)$, donc ayant pour espérance π et pour variance $\pi(1-\pi)$, on a

$$\frac{\hat{\pi} - \pi}{\sqrt{\pi(1-\pi)/n}} \xrightarrow{\text{asymp.}} \mathcal{N}(0,1).$$

Exemple : opinion sur l'avortement ... estimation de la proportion d'Américains favorables à l'avortement

La valeur observée de S est ici $s = 587$. La proportion d'Américains favorables à l'avortement π est donc estimée par $\hat{\pi} = 587/1223 = 0.48$.

1.3.3 Tests d'hypothèses sur une proportion π

Il est souvent d'intérêt de mener un test de conformité sur π afin de décider s'il est raisonnable de penser que cette probabilité prend une certaine valeur π_0. On appelle couramment ce test : test sur une proportion. Ce test peut être bilatéral ou unilatéral. Les hypothèses du test sont :

$$H_0 : \pi = \pi_0 \quad \text{versus} \quad H_1 : \text{ au choix } \left\langle \begin{array}{llll} \pi \neq \pi_0 & \text{ou} & & \text{test biltéral} \\ \pi > \pi_0 & \text{ou} & \pi < \pi_0 & \text{test unilatéral.} \end{array} \right.$$

Ce problème est classique, il est étudié depuis fort longtemps. Il existe plusieurs statistiques de tests pour confronter ces hypothèses. Nous verrons ici les statistiques pour le test de Wald, le test score ainsi que le test du rapport de vraisemblance (voir annexe A.4). Ces tests sont tous asymptotiques. Nous traiterons ensuite d'un test exact.

Test de Wald sur une proportion

Le test de Wald se base sur la statistique suivante :

$$Z_w = \frac{\hat{\pi} - \pi_0}{\sqrt{\hat{\pi}(1-\hat{\pi})/n}} \xrightarrow[H_0]{\text{asymp.}} \mathcal{N}(0,1).$$

Encore une fois, on voit que la variance de l'estimateur est estimée dans Z_w.

Test score sur une proportion

Le test score est le test le plus commun sur une proportion. La statistique du test est :

$$Z_s = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} \xrightarrow[H_0]{\text{asymp.}} \mathcal{N}(0,1).$$

Si la taille de l'échantillon n est grande et que l'hypothèse nulle est vraie, le test de Wald et le test score donneront des résultats pratiquement identiques. Cependant, si l'hypothèse nulle est en réalité fausse, la valeur observée de l'estimateur du maximum de vraisemblance $\hat{\pi}$ s'éloignera de π_0 et la variance de l'estimateur sera incorrectement estimée sous H_0 dans le test de Wald. C'est les tests score sont souvent préférés aux tests de Wald.

Test du rapport de vraisemblance sur une proportion

Nous avons s, une observation de la variable aléatoire S, qui suit une loi $Bin(n, \pi)$. La vraisemblance du paramètre π est :

$$L(\pi|s) = \binom{n}{s} \pi^s (1-\pi)^{n-s}.$$

La statistique du test de rapport de vraisemblance est :

$$LR = -2 \ln \left(\frac{L(\pi_0|S)}{L(\hat{\pi}|S)} \right)$$

où $\hat{\pi}$ est l'estimateur du maximum de vraisemblance de π. Ainsi :

$$\begin{aligned} LR &= -2 \ln \left(\frac{\binom{n}{S} \pi_0^S (1-\pi_0)^{n-S}}{\binom{n}{S} (\hat{\pi})^S (1-\hat{\pi})^{n-S}} \right) \\ &= -2 \left(S \ln \left(\frac{\pi_0}{\hat{\pi}} \right) + (n-S) \ln \left(\frac{1-\pi_0}{1-\hat{\pi}} \right) \right). \end{aligned}$$

Sous H_0, lorsque $n \to \infty$, cette statistique suit une loi du khi-deux à 1 degré de liberté : $LR \xrightarrow[H_0]{\text{asymp.}} \chi^2_1$. Cette statistique de test permet uniquement d'effectuer un test bilatéral, car elle prend la même valeur pour $\hat{\pi} = p$ et $\hat{\pi} = 1-p$.

Notons que nous aurions aussi pu déduire la forme de cette statistique de test en adoptant le point de vue que nous avons n observations (et non une seule) de n variables aléatoires iid de loi $Bernoulli(\pi)$ (et non de loi $Bin(n, \pi)$). On aurait alors écrit la vraisemblance comme suit :

$$L(\pi|\mathbf{1}_1, \ldots, \mathbf{1}_n) = \prod_{u=1}^{n} \pi^{\mathbf{1}_u}(1-\pi)^{1-\mathbf{1}_u} = \pi^{\sum_1^n \mathbf{1}_u}(1-\pi)^{\sum_1^n (1-\mathbf{1}_u)} = \pi^s(1-\pi)^{n-s},$$

car $s = \sum_1^n \mathbf{1}_u$, soit le nombre de succès observés. L'absence du facteur $\binom{n}{s}$ comparativement à la vraisemblance trouvée précédemment n'est pas problématique puisque ce facteur disparait par simplification en construisant la statistique LR du test de rapport de vraisemblance.

Exemple : opinion sur l'avortement ... test sur la proportion d'Américains favorables à l'avortement

Pour déterminer si les Américains manifestent majoritairement une préférence pour ou contre l'avortement, nous pouvons tester l'hypothèse nulle $\{H_0 : \pi = 0.5\}$ contre l'hypothèse alternative $\{H_1 : \pi \neq 0.5\}$. Pour répondre à la question de recherche ainsi formulée, on doit effectuer un test bilatéral.

Si on formulait plutôt la question de recherche ainsi : « Est-ce que la majorité des Américains sont défavorables à l'avortement ? », on devrait plutôt faire un test unilatéral. Les hypothèses confrontées seraient $\{H_0 : \pi = 0.5\}$ versus $\{H_1 : \pi < 0.5\}$. L'hypothèse alternative a été choisie vers la gauche plutôt que vers la droite, car des sondages précédents ont révélé que les Américains étaient majoritairement défavorables à l'avortement (ex. GSS 1991 : $\hat{\pi} = 424/950 = 0.446$). On se demande si c'est encore vrai.

La valeur observée de la statistique de test pour le test score est :

$$z_s = \frac{0.48 - 0.5}{\sqrt{0.5(1-0.5)/1223}} = -1.401144.$$

Le seuil observé du test bilatéral est le suivant (voir annexe) :

$$2P(\mathcal{N}(0,1) > |-1.401144|) = 0.1611709.$$

Ce seuil observé étant supérieur au seuil théorique de 5%, on ne peut pas rejeter l'hypothèse selon laquelle les Américains sont divisés sur la question

de l'avortement. Ainsi, ce test ne permet pas de conclure que, en 2010, les Américains manifestent majoritairement une préférence pour ou contre l'avortement.

Le seuil observé du test unilatéral énoncé ci-dessus est le suivant :

$$P(\mathcal{N}(0,1) < -1.401144) = 0.08058547.$$

Ce seuil n'est pas trop loin de 5%, mais il en est légèrement supérieur. On ne peut donc pas conclure que les Américains sont, en 2010, majoritairement contre l'avortement. On voit ici que le test unilatéral est plus puissant que le test bilatéral.

Ici, la taille de l'échantillon n est grande, de plus, il semble que H_0 soit vraie. Alors le test de Wald devrait donner des résultats très similaires au test score. La valeur observée de la statistique du test de Wald est :

$$z_w = \frac{0.48 - 0.5}{\sqrt{0.48(1-0.48)/1223}} = -1.459625.$$

Le seuil observé du test bilatéral est le suivant :

$$2P(\mathcal{N}(0,1) > |-1.459625|) = 0.1443932.$$

Les seuils observés diffèrent un peu entre les tests score et de Wald, mais l'inférence statistique est la même.

Effectuons maintenant le test du maximum de vraisemblance pour illustrer son fonctionnement. La valeur observée de la statistique de test est

$$\begin{aligned} lr &= -2\left(587\ln\left(\frac{0.5}{0.48}\right) + (1223-587)\ln\left(\frac{1-0.5}{1-0.48}\right)\right) \\ &= 1.963730811. \end{aligned}$$

Le seuil observé du test (valeur-p) est

$$P(\chi^2_1 > 1.963730811) = 0.1611149$$

Le test du rapport de vraisemblance aboutit à la même conclusion que le test score. Les seuils observés pour les 2 tests bilatéraux sont pratiquement égaux.

Remarquez que si on avait observé la même proportion, mais sur un plus grand nombre d'individus, le test aurait été significatif. Par exemple, si $n = 3000$ individus avaient répondu au sondage, la statistique du test score aurait valu $z_s = (0.48 - 0.5)/\sqrt{0.5(1-0.5)/3000} = -2.194473$, ce qui est supérieur en valeur absolue à 1.96, la valeur critique du test bilatéral au seuil de 5%. En fait, la figure 1.2 présente la proportion critique en fonction de la taille d'échantillon pour le test score bilatéral sur une proportion lorsque l'hypothèse nulle est $H_0 : \pi = 0.5$. Voici comment interpréter ce graphique : pour un n donné, si on observe dans notre échantillon une proportion inférieure à la courbe du bas ou supérieure à la courbe du haut, le test va conclure au rejet de H_0. On voit bien que plus n est grand, plus le test déclare significatifs de petits écarts à 0.5. La proportion observée dans notre échantillon, soit 0.48, serait déclarée significative pour une taille d'échantillon supérieure ou égale 2393.

Attention, *lorsqu'une taille d'échantillon est très grande, les paramètres sont estimés avec tellement de précision que même de petites différences sont détectées comme étant significatives*. Il est donc important de ne pas utiliser aveuglément les seuils observés. Il faut toujours évaluer avec sa logique si une différence est grande. Ici, on estime la proportion à 0.48. Ce chiffre est très proche de 0.5. Il semble exagéré de conclure qu'une différence d'à peine 2% est significative, même si c'est ce que le test nous dirait si la taille d'échantillon était de 3000.

Test exact sur une proportion

Tous les tests vus précédemment se basent sur des lois asymptotiques. Cependant, on peut aussi faire un test exact. En effet, on connaît la distribution exacte de S, le nombre de succès parmi les n essais. Il s'agit d'une distribution binomiale. On peut donc utiliser S comme statistique de test. Sous l'hypothèse nulle $H_0 : \pi = \pi_0$, la statistique de test suit une distribution $Bin(n, \pi_0)$. Pour un test bilatéral ($H_1 : \pi \neq \pi_0$), la région critique du test est définie de façon à créer deux zones, une avec les plus petites valeurs possibles de S et l'autre avec les plus grandes, dont l'aire sous la fonction de masse est inférieure ou égale à la moitié du seuil α.

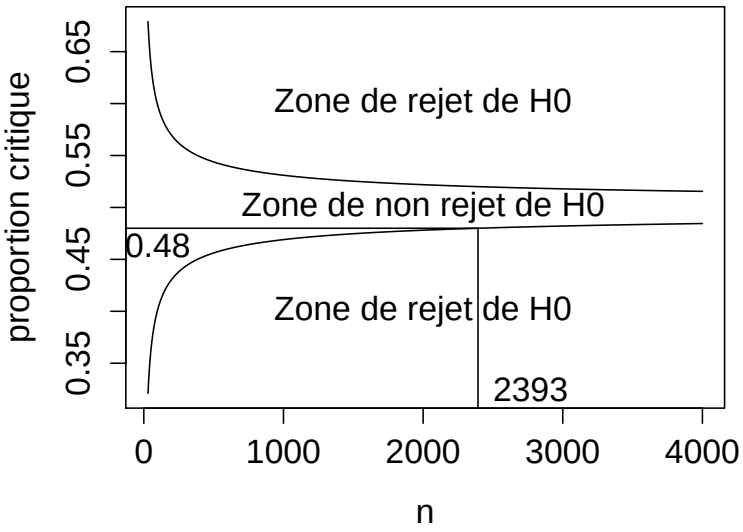


FIGURE 1.2 – Proportion critique en fonction de la taille d'échantillon pour le test score bilatéral sur une proportion lorsque l'hypothèse nulle est $H_0 : \pi = 0.5$ avec positionnement de $\hat{\pi}$ pour l'exemple de l'opinion sur l'avortement.

Par exemple, si $n = 10$ et $\pi_0 = 0.5$, la région critique du test au seuil de 5% serait l'ensemble suivant : $\{0, 1, 9, 10\}$. Les valeurs 2 et 8 ne sont pas incluses dans la région critique, car les inclure porterait le seuil à une valeur supérieure à 5%. En effet,

$$\begin{aligned} P(S \geq 9) = P(S \leq 1) &= \sum_{s=0}^{1} \binom{10}{s} 0.5^s (1 - 0.5)^{10-s} \\ &= 0.000977 + 0.009766 = 0.0107, \end{aligned}$$

mais $P(S = 8) = P(S = 2) = \binom{10}{2} 0.5^2 (1 - 0.5)^{10-2} = 0.0439$. Donc la probabilité de tomber dans la région critique $\{0, 1, 9, 10\}$, en supposant que H_0 est vraie, est de $2 \times 1.07\% = 2.14\%$, alors que cette probabilité vaut $2.14 + 2 \times 4.39\% \approx 10.94\%$ pour la région critique $\{0, 1, 2, 8, 9, 10\}$.

Dans l'exemple, le seuil réel du test, soit la probabilité de tomber dans la région critique, ne peut pas être exactement 5% comme lorsque la distri-

bution de la statistique de test sous H_0 est continue. Pour un test exact se basant sur une loi discrète, le seuil réel du test est, la grande majorité du temps, inférieur au seuil ou niveau de signification visé (par exemple 5%). Ce genre de test est donc conservateur.

Cette caractéristique d'être un peu trop conservateur, donc de ne pas rejeter assez souvent H_0, peut être corrigée en travaillant avec un seuil observé modifié appelé en anglais « mid p-value ». Comme on l'a vu dans l'introduction, un seuil observé se définit par la probabilité, sous H_0, d'obtenir un résultat égal ou plus extrême que celui observé. Cette définition signifie que pour calculer le seuil observé il faut sommer les probabilités de toutes les valeurs possibles de la statistique de test qui ont une probabilité inférieure ou égale à la probabilité du résultat observé ($P(S = s_{obs})$), et qui respectent la direction de l'hypothèse alternative du test si celle-ci est unilatérale.

Dans l'exemple précédent avec $n = 10$, calculons le seuil observé du test bilatéral pour deux valeurs de π_0 : 0.5 et 0.45. Nous aurons besoin de la fonction de masse de S sous H_0 pour ces deux valeurs de π_0 :

s	$P(S = s\|\pi_0 = 0.5)$	$P(S = s\|\pi_0 = 0.45)$
0	**0.0010**	**0.0025**
1	**0.0098**	**0.0207**
2	**0.0439**	0.0763
3	0.1172	0.1665
4	0.2051	0.2384
5	0.2461	0.2340
6	0.2051	0.1596
7	0.1172	0.0746
8	**0.0439**	**0.0229**
9	**0.0098**	**0.0042**
10	**0.0010**	**0.0003**

Supposons que l'on observe dans notre échantillon $s_{obs} = 8$.

Si $\pi_0 = 0.5$, la probabilité du résultat observé est $P(S \geq 8) = 0.0439$. Les résultats égaux ou plus extrêmes pour un test bilatéral sont donc 0, 1, 2, 8, 9 et 10, car ces valeurs sont associées à des probabilités inférieures ou égales

à 0.0439. Le seuil observé du test est donc

$$P(S \geq 8) + P(S \leq 2) = 2 \times (0.0439 + 0.0010 + 0.0098) = 0.1094.$$

Le calcul de ce seuil observé est facilité par la symétrie de la distribution sous H_0. Si π_0 ne vaut pas 0.5, la distribution de la statistique de test sous H_0 n'est plus symétrique. Par exemple, si $\pi_0 = 0.45$, la probabilité du résultat observé devient $P(S \geq 8) = 0.0229$. Le résultat 2 n'est plus un résultat dit extrême puisque la probabilité de ce résultat (0.0763) est supérieure à celle du résultat observé. Le seuil observé du test devient

$$P(S \geq 8)+P(S \leq 1) = (0.0229+0.0042+0.0003)+(0.0207+0.0025) = 0.0506.$$

Ainsi, on ne peut rejeter au seuil de 5% ni l'hypothèse $H_0 : \pi_0 = 0.5$, ni l'hypothèse $H_0 : \pi_0 = 0.45$.

Un mid p-value est, par définition, la moitié de la probabilité du résultat observé plus la probabilité d'un résultat plus extrême. Pour le calculer, on additionne les mêmes probabilités que dans le calcul du seuil observé ordinaire. Cependant, on divise préalablement par deux la probabilité que S soit égal à la valeur observée s_{obs}. Si d'autres valeurs possibles de S sont associées à la même probabilité que celle de la valeur observée, on divise aussi par deux leurs probabilités. Ainsi, le mid p-value sera nécessairement plus petit que le seuil observé. L'utiliser signifie que l'on rejette plus souvent H_0 que si on utilisait le seuil observé ordinaire. Le test devient ainsi moins conservateur.

Calculons les deux mid p-values correspondants aux seuils observés que l'on vient tout juste de calculer. Dans le cas où, $\pi_0 = 0.5$, le mid p-value pour la valeur observée $s_{obs} = 8$ est :

$$\frac{1}{2}P(S = 8) + \frac{1}{2}P(S = 2) + P(S > 8) + P(S < 2) =$$
$$2 \times (0.0439/2 + 0.0010 + 0.0098) = 0.0655.$$

Dans le cas où, $\pi_0 = 0.45$, pour la même valeur observée, le mid p-value est

$$\frac{1}{2}P(S = 8) + P(S > 8) + P(S \leq 1) =$$
$$0.0229/2 + (0.0042 + 0.0003) + (0.0207 + 0.0025) = 0.0392.$$

Ce seuil observé modifié nous mènerait au rejet de l'hypothèse nulle $H_0 : \pi_0 = 0.45$, mais pas au rejet de $H_0 : \pi_0 = 0.5$.

La figure 1.3 illustre le calcul du seuil observé ordinaire et du mid p-value.

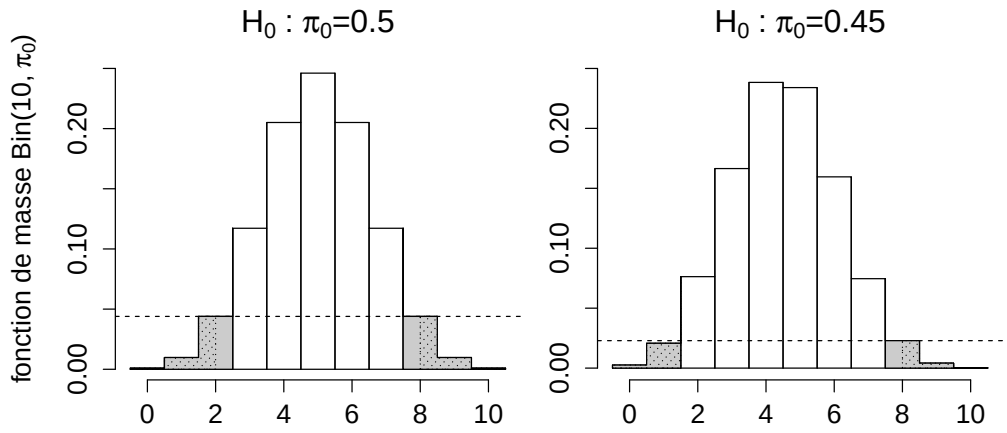


FIGURE 1.3 – Seuil observé pour un test exact sur une proportion avec $n = 10$, le seuil observé ordinaire est l'aire de la région en gris, le mid p-value est l'aire de la région hachurée.

1.3.4 Intervalle de confiance pour une proportion π

Au fil des ans, plusieurs auteurs ont proposé des intervalles de confiance pour une proportion. Nous n'allons pas tous les voir dans ce cours, car ils sont trop nombreux. Agresti et Coull (1998), ainsi que Fleiss *et al.* (2003), sont de bonnes références à ce sujet.

Chacun des tests mentionnés à la section précédente peut être inversé afin de créer un intervalle de confiance. L'intervalle de confiance exact est dû à Clopper et Pearson (1934). Il porte parfois le nom de ses auteurs. La formule pour cet intervalle ne sera pas donnée ici, mais on peut le calculer facilement avec un logiciel statistique. Pendant longtemps, cet intervalle a été considéré comme étant le meilleur. Cependant, Agresti et Coull (1998) ont

montré que cet intervalle était inutilement grand. Ils ont plutôt trouvé que le meilleur intervalle de confiance, en terme de pourcentage de couverture, était l'intervalle score, et ce, même pour de petits échantillons !

Ces notes présentent donc l'intervalle score, considéré comme étant le meilleur, et l'intervalle de confiance de Wald, car c'est un intervalle de confiance classique en statistique.

Intervalle de confiance de Wald pour une proportion

L'intervalle de confiance le plus simple est l'intervalle de Wald, dont la formule est la suivante :

$$\left[\hat{\pi} - z_{\alpha/2}\sqrt{\hat{\pi}(1-\hat{\pi})/n}\ ,\ \hat{\pi} + z_{\alpha/2}\sqrt{\hat{\pi}(1-\hat{\pi})/n}\right]$$

où $\hat{\pi}$ est l'estimateur du maximum de vraisemblance de π. Cette formule pour les bornes de l'intervalle a été simplement obtenue en résolvant pour π l'équation suivante :

$$\frac{|\hat{\pi} - \pi|}{\sqrt{\hat{\pi}(1-\hat{\pi})/n}} = z_{\alpha/2}$$

(voir annexe A.5 pour plus de détails).

Intervalle de confiance score pour une proportion

L'intervalle de confiance score pour une proportion est aussi nommé intervalle de confiance de Wilson, car il aurait été présenté pour la première fois par Wilson (1927). Obtenir la formule de l'intervalle de confiance score n'est pas aussi simple que pour l'intervalle de confiance de Wald. Les bornes de l'intervalle sont les deux solutions pour π de l'équation suivante :

$$\frac{|\hat{\pi} - \pi|}{\sqrt{\pi(1-\pi)/n}} = z_{\alpha/2}.$$

La différence avec l'équation à résoudre pour l'intervalle de confiance de Wald est la présence de π au dénominateur en plus de sa présence au numérateur de la partie gauche de l'équation, ce qui rend π plus difficile à isoler. On se retrouve maintenant à devoir trouver les racines d'une équation de second

degré. Les calculs sont laissés en exercice. On obtient l'intervalle $[L, U]$ avec :

$$\begin{aligned} L &= \frac{n}{n + z_{\alpha/2}^2} \left(\hat{\pi} + \frac{z_{\alpha/2}^2}{2n} - z_{\alpha/2} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n} + \frac{z_{\alpha/2}^2}{4n^2}} \right) \\ U &= \frac{n}{n + z_{\alpha/2}^2} \left(\hat{\pi} + \frac{z_{\alpha/2}^2}{2n} + z_{\alpha/2} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n} + \frac{z_{\alpha/2}^2}{4n^2}} \right). \end{aligned}$$

Agresti et Coull (1998) ont proposé une façon simple de calculer pratiquement les mêmes bornes que l'intervalle score avec un niveau de confiance de 95%, mais avec une formule beaucoup plus simple. Ils ont prouvé qu'en ajoutant deux succès et deux échecs aux données (S devient $S + 2$ et n devient $n + 4$), puis en calculant l'intervalle de confiance de Wald, on tombe à peu près sur l'intervalle de confiance score.

1.4 Expérience avec la loi multinomiale

Exemple : intentions de vote des Québécois

Le journal La Presse publiait, le 31 octobre 2007, un article de Denis Lessard présentant les résultats d'un sondage sur les intentions de vote des Québécois. À l'époque, les trois principaux partis se partageant les intentions de vote étaient le Parti Québécois (PQ), le Parti libéral du Québec (PLQ) et l'Action démocratique du Québec (ADQ). Considérons ici uniquement ces partis. Les intentions de vote se répartissaient ainsi :

Parti politique	PQ	PLQ	ADQ	total
Fréquence observée	264	264	238	766

Le graphique 1.4 met clairement en évidence que ces intentions de vote sont partagées de façon pratiquement égale entre les trois partis.

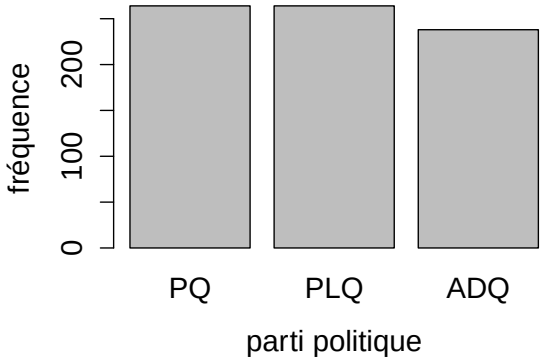


FIGURE 1.4 – Diagramme en bâtons des réponses dans l'exemple des intentions de vote des Québécois.

1.4.1 La loi multinomiale

Le contexte de la loi multinomiale est le suivant. On fait une expérience qui peut prendre J résultats possibles $\{m_1^Y, \ldots, m_J^Y\}$. On répète cette expérience de façon indépendante n fois, et on note S_j le nombre de résultats m_j^Y obtenus ($j = 1, \ldots, J$). En fait, on pourrait noter Y_1 à Y_n l'échantillon découlant de ces n expériences (souvent n individus sondés). Ces variables aléatoires prennent une valeur dans l'ensemble des J résultats possibles. De cet échantillon, on peut construire les fréquences S_j pour $j = 1, \ldots, J$:

Valeur de Y	Fréquence
m_1^Y	S_1
$\vdots$	$\vdots$
m_j^Y	S_j
$\vdots$	$\vdots$
m_J^Y	S_J
Total	n

On vient ainsi de créer un tableau de fréquences à une variable. Les fréquences sont ici représentées par des S_j majuscules plutôt que par des petits n_j comme au début du chapitre afin de faire ressortir le fait que l'on considère maintenant qu'il s'agit de variables aléatoires. Les fréquences sont aléatoires, car elles proviennent d'un échantillon aléatoire. Une réalisation des variables aléatoires $(S_1, \ldots, S_J)$ sera dans cette section notée $(s_1, \ldots, s_J)$, qui sont en fait l'équivalent des n_j du début du chapitre.

Pour chaque expérience, π_j représente la probabilité d'obtenir le résultat m_j^Y : $\pi_j = P(Y = m_j^Y)$. Ce contexte est similaire à celui pour une distribution binomiale, à la différence que le nombre d'issues possibles à l'expérience peut être supérieur à 2. On dit que le vecteur $\boldsymbol{S} = (S_1, \ldots, S_J)$ suit une loi multinomiale de paramètres n (nombre total d'essais) et $(\pi_1, \ldots, \pi_J)$ (probabilités des résultats de chacun des types). On a bien sûr les 2 contraintes suivantes :

$$\begin{aligned} S_1 + \ldots + S_J &= n \quad \text{et} \\ \pi_1 + \ldots + \pi_J &= 1 \end{aligned}$$

avec $0 \leq \pi_j \leq 1$ pour $j = 1, \ldots, J$. En raison de la première contrainte, les S_j ne sont pas des variables indépendantes. On note

$$\boldsymbol{S} \sim Multinomiale(n, \pi_1, \ldots, \pi_J).$$

De façon marginale, on a

$$S_j \sim Bin(n, \pi_j)$$

pour $j = 1, \ldots, J$. Notez que lorsque $J = 2$, il devient inutile de travailler avec le vecteur $\boldsymbol{S} = (S_1, S_2)$ étant donné que la valeur d'une variable peut être déduite de l'autre variable ($S_1 = n - S_2$). On définit plutôt une des deux modalités possibles de l'expérience comme étant un succès (disons la modalité 1) et on travaille uniquement avec $S_1 \sim Bin(n, \pi_1)$. La distribution binomiale est plus simple à manipuler que la distribution multinomiale puisqu'elle est univariée plutôt que multivariée.

La fonction de masse de la loi multinomiale s'écrit de la façon suivante :

$$P(S_1 = s_1, \ldots, S_J = s_J) = \frac{n!}{s_1! \ldots s_k!} \pi_1^{s_1} \ldots \pi_J^{s_J},$$

pour $\{(s_1, \ldots, s_J) \in \mathbb{N}^J : s_1 + \ldots + s_J = n\}$, soit l'ensemble des nombres naturels (entiers non négatifs) de dimension J, tel que la somme des s_j vaut n. L'espérance et la matrice de variance-covariance de $\mathbf{S}$ sont :

$$\begin{aligned} E[S_j] = &\quad n\pi_j &&\text{pour } j = 1, \ldots, k \\ Var[S_j] = &\quad n\pi_j(1 - \pi_j) &&\text{pour } j = 1, \ldots, k \\ Cov(S_j, S_{i'}) = &\quad -n\pi_j\pi_j &&\text{pour } j \neq j' \end{aligned}$$

1.4.2 Estimation ponctuelle du paramètre $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_J)$

L'estimateur du maximum de vraisemblance (Hogg *et al.*, 2005, exemple 6.4.5) de $\boldsymbol{\pi}$ est $\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \ldots, \hat{\pi}_J)$ avec

$$\hat{\pi}_j = \frac{S_j}{n}, \text{ pour } j = 1, \ldots, J.$$

On peut calculer l'espérance et la variance de $\hat{\boldsymbol{\pi}}$. On trouve :

$$\begin{aligned} E[\hat{\pi}_j] = &\quad \pi_j &&\text{pour } j = 1, \ldots, J \\ Var[\hat{\pi}_j] = &\quad \frac{\pi_j(1-\pi_j)}{n} &&\text{pour } j = 1, \ldots, J \\ Cov(\hat{\pi}_j, \hat{\pi}_{i'}) = &\quad \frac{-\pi_j\pi_{i'}}{n} &&\text{pour } j \neq j'. \end{aligned}$$

Sous forme matricielle, on a donc :

$$E[\hat{\boldsymbol{\pi}}] = (\pi_1, \ldots, \pi_J)$$

$$Var[\hat{\boldsymbol{\pi}}] = \frac{1}{n} \begin{pmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 & \cdots & -\pi_1\pi_J \\ -\pi_2\pi_1 & \pi_2(1-\pi_2) & \cdots & -\pi_2\pi_J \\ \vdots & \vdots & \vdots & \vdots \\ \pi_J\pi_1 & -\pi_J\pi_2 & \cdots & \pi_J(1-\pi_J) \end{pmatrix}.$$

On peut trouver la distribution asymptotique de l'estimateur $\hat{\pi}_j$ (pour $j = 1, \ldots, J$) en notant que $\hat{\pi}_j$ est une moyenne de n variables iid $Bin(1, \pi_j)$ avec espérance π_j et variance $\pi_j(1-\pi_j)$. Par le théorème limite central, on a donc

$$\frac{\hat{\pi}_j - \pi_j}{\sqrt{\pi_j(1-\pi_j)/n}} \xrightarrow{\text{asymp.}} \mathcal{N}(0,1)$$

On peut même déterminer la distribution asymptotique du vecteur $\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \ldots, \hat{\pi}_J)$. On trouve

$$\hat{\boldsymbol{\pi}} \xrightarrow{\text{asymp.}} \mathcal{N}(\boldsymbol{\pi}, \Sigma),$$

avec $\Sigma = Var[\hat{\boldsymbol{\pi}}]$. Ici, la loi normale est multivariée, de dimension J.

1.4.3 Test d'hypothèses sur la valeur de $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_J)$

On cherche à tester ici les hypothèses suivantes.

$$H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}_0 \quad \text{ou} \quad (\pi_1, \ldots, \pi_J) = (\pi_{0,1}, \ldots, \pi_{0,J})$$
$$H_1 : \boldsymbol{\pi} \neq \boldsymbol{\pi}_0 \quad \text{ou} \quad (\pi_1, \ldots, \pi_J) \neq (\pi_{0,1}, \ldots, \pi_{0,J})$$

Il s'agit d'un test bilatéral multivarié. Nous ne verrons pas dans ce cours comment déduire de façon générale des tests de Wald ou score multivariés. Cependant, il est facile de confronter ces hypothèses avec un test de rapport de vraisemblance. Nous allons aussi introduire ici un test qui reviendra à de multiples reprises dans ce cours : le test du khi-deux de Pearson.

Test de rapport de vraisemblance sur la valeur de $\boldsymbol{\pi}$

Nous avons $\boldsymbol{s} = (s_1, \ldots, s_J)$, une observation de la variable aléatoire $\boldsymbol{S}$, qui suit une loi $Multinomiale(n, \boldsymbol{\pi})$. La vraisemblance de paramètre $\boldsymbol{\pi}$ est

$$L(\boldsymbol{\pi}|\boldsymbol{s}) = \frac{n!}{s_1! \ldots s_J!} \pi_1^{s_1} \ldots \pi_J^{s_J}.$$

Le logarithme naturel de cette valeur est :

$$\ln L(\boldsymbol{\pi}|\boldsymbol{s}) = \ln\left(\frac{n!}{s_1!\dots s_J!}\right) + \sum_{j=1}^{J} s_j \ln(\pi_j).$$

La statistique du test du rapport des vraisemblances est donc :

$$\begin{aligned} G^2 &= -2(\ln L(\boldsymbol{\pi}_0|\boldsymbol{S}) - \ln L(\hat{\boldsymbol{\pi}}|\boldsymbol{S})) \\ &= -2\left(\ln\left(\frac{n!}{S_1!\dots S_J!}\right) + \sum_{j=1}^{J} S_j \ln(\pi_{0,j}) - \ln\left(\frac{n!}{S_1!\dots S_J!}\right) - \sum_{j=1}^{J} S_j \ln(\hat{\pi}_j)\right) \\ &= -2\sum_{j=1}^{J} S_j \ln\left(\frac{\pi_{0,j}}{\hat{\pi}_j}\right). \end{aligned}$$

Notez que dans ce contexte, cette statistique est souvent notée G^2, c'est pourquoi on adopte cette notation ici. Sous H_0, cette statistique suit approximativement une loi du khi-deux à $J-1$ degrés de liberté. Pour justifier ces degrés de liberté, il faut se rappeler qu'ils sont la différence entre la dimension de l'espace des valeurs possibles de $\boldsymbol{\pi}$, ici $J-1$ à cause de la contrainte $\pi_1 + \ldots + \pi_J = 1$, et le nombre de paramètres libres sous l'hypothèse nulle, ici 0 (voir annexe A.4.3).

En résumé, la statistique de ce test de rapport de vraisemblance est :

$$G^2 = -2\sum_{j=1}^{J} S_j \ln\left(\frac{\pi_{0,j}}{\hat{\pi}_j}\right) \xrightarrow[H_0]{\text{asymp.}} \chi^2_{J-1}.$$

Test du khi-deux de Pearson sur la valeur de $\boldsymbol{\pi}$

Une solution de rechange classique au test du rapport de vraisemblance dans le cas d'un test sur les paramètres d'une loi multinomiale est le test du khi-deux de Pearson. La statistique du khi-deux de Pearson a été proposée par Karl Pearson en 1900. C'est à ce même statisticien britannique que l'on doit, entre autres, la corrélation de Pearson. Notons cependant que l'intervalle de confiance exact pour une proportion appelé intervalle de Clopper-Pearson, que nous avons vu précédemment, est plutôt attribuable à son fils, Egon Sharpe Pearson.

Dans le contexte d'un test sur les paramètres d'une loi multinomiale, Pearson a proposé d'utiliser la statistique suivante :

$$X^2 = \sum_{j=1}^{J} \frac{(S_j - n\pi_{0,j})^2}{n\pi_{0,j}} \xrightarrow[H_0]{\text{asymp.}} \chi^2_{J-1}.$$

Il a prouvé que sous l'hypothèse nulle $H_0 : (\pi_1, \ldots, \pi_J) = (\pi_{0,1}, \ldots, \pi_{0,J})$, cette statistique suit asymptotiquement une loi du khi-deux à $J-1$ degrés de liberté, tout comme la statistique du test de rapport de vraisemblance.

La preuve formelle de ce résultat ne sera pas faite en classe, car elle fait intervenir des statistiques multivariées. Elle est présentée dans (Agresti, 2002, sections 1.5.4 et 14.3). On peut cependant justifier ce résultat de façon simple en supposant que les fréquences S_j pour $j = 1, \ldots, J$ sont des variables aléatoires Poisson, de paramètre $\lambda_j = E[S_j] = n\pi_{0,j}$. Cette supposition n'est pas farfelue considérant qu'une loi binomiale peut être approximée par une loi Poisson sous certaines conditions. Comme on l'a vu précédemment, pour de grandes valeurs de λ_j, on peut supposer que $Z_j = (S_j - \lambda_j)/\sqrt{\lambda_j} = (S_j - n\pi_{0,j})/\sqrt{n\pi_{0,j}}$ suit une loi normale centrée réduite. Supposons pour l'instant que les J variables aléatoires sont indépendantes. On aurait donc que la somme des Z_j élevés au carré, soit $X^2 = \sum_{j=1}^{J}(S_j - n\pi_{0,j})^2/n\pi_{0,j}$, suit une loi asymptotique khi-deux à J degrés de liberté. En réalité, les J variables aléatoires ne sont pas indépendantes puisqu'elles sont soumises à la contrainte $\sum_{j=1}^{J} S_j = n$. On perd un degré de liberté à cause de cette contrainte, qui représente la conversion de la loi Poisson à la loi multinomiale.

Remarquez que si $J = 2$, la statistique du khi-deux de Pearson se simplifie à la statistique du test score sur une proportion, élevée au carré. En fait, pour une valeur de J quelconque, le test du khi-deux de Pearson est, en réalité, le test score sur les paramètres d'une loi multinomiale.

Exemple : intentions de vote des Québécois ... suite

On veut tester si les votes sont vraiment répartis uniformément entre les partis politiques :

$H_0 : (\pi_1, \pi_2, \pi_3) = (1/3, 1/3, 1/3)$ versus $H_1 : (\pi_1, \pi_2, \pi_3) \neq (1/3, 1/3, 1/3)$.

La statistique du test de rapport de vraisemblance prend la valeur suivante :

$$G^2_{obs} = -2\left(2 \times 264 \ln\left(\frac{1/3}{264/766}\right) + 238 \ln\left(\frac{1/3}{238/766}\right)\right) = 1.786.$$

Le seuil observé de ce test est donc $P(\chi^2_2 \geq 1.786) = 0.40942$. Cette valeur est nettement plus grande que 0.05, on conclut donc au non-rejet de H_0.

La statistique du test du khi-deux de Pearson prend quant à elle la valeur suivante :

$$X^2_{obs} = 2 \times \frac{(264 - 766/3)^2}{766/3} + \frac{(238 - 766/3)^2}{766/3} = 1.765.$$

Le seuil observé de ce test est donc $P(\chi^2_2 \geq 1.765) = 0.41375$. Cette valeur est aussi nettement plus grande que 0.05.

Les deux tests nous mènent donc à la conclusion que les votes sont réellement répartis de façon uniforme entre les partis politiques.

1.4.4 Intervalle de confiance pour $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_J)$

Si $\boldsymbol{S} = (S_1, \ldots, S_J)$ suit une loi $Multinomiale(n, \pi_1, \ldots, \pi_J)$, alors $S_j \sim Bin(n, \pi_j)$. Un intervalle de confiance (IC) pour π_j peut donc être calculé comme à la section 1.3.4. Cependant, si on fait J intervalles de confiance de niveau $1 - \alpha^*$, le niveau de confiance global sera inférieur à $1 - \alpha^*$. En effet, l'inégalité de Bonferroni dit que pour tout événement A_1 à A_J,

$$P(\cap_{j=1}^J A_j) \geq 1 - \sum_{j=1}^J P(A_j^c)$$

où A^c est le complément de l'événement A, donc $P(A^c) = 1 - P(A)$. Ainsi, on a que :

$$P((\pi_1 \in \text{IC}_1) \text{ et} \ldots \text{et } (\pi_J \in \text{IC}_J)) \geq 1 - \sum_{j=1}^J P(\pi_j \notin \text{IC}_j)$$

La probabilité $P((\pi_1 \in \text{IC}_1) \text{ et} \ldots \text{et } (\pi_J \in \text{IC}_J))$ est ce que l'on appelle le niveau de confiance global des J intervalles de confiance.

Si on fait des intervalles de confiance individuels de niveau $1-\alpha^*$, on a que $P(\pi_j \notin \mathrm{IC}_j) = \alpha^*$ pour tout $j = 1, \ldots, k$. En conséquence, on a que :

$$P((\pi_1 \in \mathrm{IC}_1) \text{ et} \ldots \text{et } (\pi_J \in \mathrm{IC}_J)) \geq 1 - k\alpha^*.$$

Ainsi, pour s'assurer que le niveau de confiance global des J intervalles de confiance est d'au moins $(1-\alpha)$, il faut que :

$$1 - \alpha = 1 - k\alpha^* \qquad \Rightarrow \qquad \alpha^* = \frac{\alpha}{k}.$$

Il faut donc faire des intervalles individuels de niveau $(1-\alpha/k)$. Cet ajustement des niveaux de confiance des intervalles de confiance individuels est nommé « correction de Bonferroni ».

1.5 Pour aller plus loin

1.5.1 Généralisation du test sur les paramètres d'une loi multinomiale

La statistique du khi-deux de Pearson peut être utilisée dans plusieurs contextes, notamment pour faire un test sur les paramètres d'une loi multinomiale (cas traité précédemment), un test d'adéquation de données à une loi (voir section 1.5.2), un test d'homogénéité des populations ou d'indépendance dans un tableau de fréquences à deux variables et un test d'adéquation d'un modèle linéaire généralisé. La formule pour définir la statistique varie un peu d'un contexte à l'autre, mais l'idée de base est toujours la même. Cette statistique permet de tester une hypothèse nulle selon laquelle les probabilités d'occurrence de certains événements, observés dans un échantillon, sont égales à certaines valeurs théoriques. Les événements doivent être mutuellement exclusifs et leurs probabilités d'occurrence doivent sommer à 1. Notons O_j les fréquences observées dans l'échantillon des J événements à tester et E_j les fréquences espérées de ces événements, sous l'hypothèse nulle. La statistique du khi-deux de Pearson est :

$$X^2 = \sum_{j=1}^{J} \frac{(O_j - E_j)^2}{E_j} = \sum_{j=1}^{J} \frac{O_j^2}{E_j} - \sum_{j=1}^{J} E_j.$$

La valeur de cette statistique sera petite si les fréquences observées prennent des valeurs se rapprochant des fréquences espérées.

Il existe toujours un test du rapport de vraisemblance équivalent à un test du khi-deux de Pearson. La statistique de ce test peut s'écrire comme suit :

$$G^2 = -2\sum_{j=1}^{J} O_j \ln\left(\frac{E_j}{O_j}\right) = 2\sum_{j=1}^{J} O_j \ln\left(\frac{O_j}{E_j}\right).$$

Sous l'hypothèse nulle que les fréquences espérées sont vraies, les statistiques X^2 et G^2 suivent asymptotiquement une loi du khi-deux à d degrés de liberté. Ces degrés de liberté sont la différence entre le nombre de paramètres libres dans l'espace de toutes les valeurs possibles des fréquences O_1 à O_J et le nombre de paramètres libres sous l'hypothèse nulle.

Mise en garde contre de faibles fréquences espérées

Notez que l'approximation de la statistique du khi-deux de Pearson X^2, ainsi que de la statistique de rapport de vraisemblance équivalente G^2, par une loi de khi-deux est en générale correcte, mais sous la condition qu'au moins 80% des classes $j = 1, \ldots, J$ soient telles que $E_j \geq 5$. Ainsi, si plus de 20% des fréquences espérées sont inférieures à 5, l'approximation par la khi-deux n'est pas très juste.

Une solution possible dans un tel cas est de regrouper des classes de façon à augmenter les valeurs des fréquences espérées. On perd cependant des degrés de liberté pour la loi asymptotique de la statistique de test et, en conséquence, le test perd en puissance.

1.5.2 Test d'adéquation de données à une loi

On peut utiliser les statistiques X^2 et G^2 pour mener un test permettant de vérifier l'ajustement d'un ensemble de données observées à une loi spécifique, peu importe laquelle.

Il y a deux hypothèses nulles potentielles pour ce test. La première est la suivante :

$$\text{type 1} \quad \rightarrow \quad H_0 : \text{ La loi } \mathcal{L}(\boldsymbol{\theta}_0) \text{ s'ajuste bien aux données}$$

où $\mathcal{L}(\boldsymbol{\theta}_0)$ représente n'importe quelle distribution avec vecteur de paramètres $\boldsymbol{\theta}_0$ dont les valeurs sous H_0 sont prédéfinies. L'autre hypothèse nulle possible est la suivante :

$$\text{type 2} \quad \rightarrow \quad H_0 : \text{ La famille de loi } \mathcal{L} \text{ s'ajuste bien aux données.}$$

Dans cette hypothèse nulle, les valeurs des paramètres de la loi ne sont pas spécifiées. Ils peuvent prendre n'importe quelle valeur. Pour mener le test dans ce cas, on va estimer les paramètres de la loi par maximum de vraisemblance. Cette estimation fera perdre des degrés de liberté à la loi khi-deux asymptotique des statistiques X^2 et G^2.

L'hypothèse alternative est toujours ici le complément de l'hypothèse nulle. Il s'agit donc d'un test bilatéral.

On considère un échantillon Y_1 à Y_n de n variables aléatoires indépendantes et identiquement distribuées dont la distribution est inconnue. Les n observations sont disposées dans un tableau de fréquence ayant J classes. Si les variables aléatoires Y_u sont catégoriques ou numériques discrètes, ces classes sont typiquement les modalités possibles des variables. Si les Y_u sont plutôt continues, les classes sont des intervalles disjoints couvrant entièrement le support des valeurs possibles des Y_u. Ces intervalles sont choisis de façon arbitraire.

On note O_j la fréquence observée dans la classe j, soit le nombre d'observations tombant dans cette classe. On note aussi E_j la fréquence espérée sous l'hypothèse nulle pour la classe j. On a donc les fréquences observées et espérées suivantes :

Valeurs des Y_j	classe 1	$\cdots$	classe k	total
Fréquence observée	O_1	$\cdots$	O_J	n
Fréquence espérée	E_1	$\cdots$	E_J	n

Comment calcule-t-on les fréquences espérées ? Sous l'hypothèse nulle, on peut calculer $P_j = P(\text{Observer une donnée dans la classe } j|H_0)$. Le calcul de ces probabilités fait intervenir la fonction de densité de la loi $\mathcal{L}$:

$$P_j = P(Y \in \text{ classe } j|Y \sim \mathcal{L}(\boldsymbol{\theta}))$$

On doit attribuer une certaine valeur aux paramètres $\boldsymbol{\theta}$ de la loi afin de pouvoir effectuer ces calculs. Avec l'hypothèse nulle du type 1 énoncé précédemment, ces valeurs sont prédéfinies : $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Cependant, avec l'hypothèse nulle du type 2, on utilise les estimateurs de maximum de vraisemblance des paramètres calculés à partir des données : $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. Puisqu'on observe n données, on espère donc en moyenne obtenir nP_j observations dans la classe j. Ainsi, $E_j = nP_j$.

On effectue donc le test avec la statistique X^2 ou G^2 définie ci-dessus. La loi asymptotique de ces statistiques est khi-deux. Cependant, les degrés de liberté de cette khi-deux diffèrent selon l'hypothèse nulle (type 1 ou 2). Rappelons que les degrés de liberté sont la différence entre le nombre de paramètres libres dans l'espace de toutes les valeurs possibles des fréquences O_1 à O_J et le nombre de paramètres libres sous l'hypothèse nulle. Le premier élément de cette différence vaut $J-1$, car les fréquences sont toujours contraintes à sommer à n. Le deuxième élément vaut 0 avec H_0 de type 1, car les valeurs des paramètres sont fixées. Avec H_0 de type 2, il vaut p, soit le nombre de paramètres de la loi $\mathcal{L}$ que l'on estime à partir des données. Notez que l'on pourrait même prédéfinir la valeur de certains paramètres sous H_0 et en estimer d'autres.

En réalité, le test décrit ici, peu importe que l'on choisisse l'hypothèse nulle du premier ou du deuxième type, revient au test de l'hypothèse nulle :

$$H_0 : \text{ Les données suivent une loi } Multinomiale(n, \boldsymbol{\pi}_0),$$

avec $\boldsymbol{\pi}_0 = (\pi_{0,1}, \ldots, \pi_{0,J})$ et $\pi_{0,j} = P_j$, tel que défini précédemment. Ainsi, lorsque l'on effectue un test d'adéquation de données à une loi avec la statistique X^2 ou G^2, cela revient à faire un test sur les paramètres d'une loi multinomiale.

Notons qu'une solution de rechange au test d'adéquation de données à une loi avec la statistique X^2 ou G^2 est le test de Kolmogorov-Smirnov se basant sur la fonction de répartition empirique.

Exemple : armée prussienne ... test d'adéquation à la loi Poisson

On veut tester ici si les données suivent une loi Poisson quelconque. Les hypothèses confrontées sont donc les suivantes :

H_0 : La famille de loi Poisson s'ajuste bien aux données
H_1 : La famille de loi Poisson ne s'ajuste pas bien aux données

Pour mener le test, nous devons d'abord calculer des fréquences espérées sous l'hypothèse nulle. Étant donné que celle-ci ne spécifie aucune valeur prédéfinie pour le paramètre λ de la loi Poisson, nous allons d'abord estimer

ce paramètre par maximum de vraisemblance. On obtient $\hat{\lambda} = \bar{x} = 0.61$ (ce calcul a été fait précédemment). On peut ensuite calculer les fréquences observées :

Nbre de décès	0	1	2	3	4	5 ou plus	total
Fréquence observée	109	65	22	3	1	0	200
Fréquence espérée	108.7	66.3	20.2	4.1	0.6	0.1	200

Par exemple, la fréquence espérée de la modalité 3 est :

$$nP(Y = 3|Y \sim Poisson(0.61)) = 200\frac{e^{-0.61}0.61^3}{3!} = 4.1.$$

La fréquence espérée de la dernière modalité est simplement 200 moins les fréquences espérées de toutes les autres modalités.

Deux fréquences espérées sont très faibles ici. Afin de nous assurer de la validité de la loi asymptotique de la statistique de test, nous allons regrouper les 3 dernières classes :

valeur de S	0	1	2	3 ou plus	total
Fréquence observée	109	65	22	4	200
Fréquence espérée	108.7	66.3	20.2	4.8	200

Les valeurs observées des statistiques de test sont donc les suivantes :

$$\begin{aligned} X^2_{obs} &= \frac{(109-108.7)^2}{108.7} + \frac{(65-66.3)^2}{66.3} + \frac{(22-20.2)^2}{20.2} + \frac{(4-4.8)^2}{4.8} = 0.324 \\ G^2_{obs} &= 2\left(109\ln\left(\frac{109}{108.7}\right) + 65\ln\left(\frac{65}{66.3}\right) + 22\ln\left(\frac{22}{20.2}\right) + 4\ln\left(\frac{4}{4.8}\right)\right) \\ &= 0.328 \end{aligned}$$

Sous H_0, ces statistiques suivent asymptotiquement une loi du khi-deux à $4-1-1 = 2$ degrés de liberté. Les seuils observés des tests sont donc $P(\chi^2_2 \geq 0.324) \approx P(\chi^2_2 \geq 0.328) \approx 0.85$. Ce seuil observé est nettement supérieur à 5%. Comme on s'y attendait puisque l'espérance et la variance échantillonnale de ces données étaient pratiquement égales, on accepte l'hypothèse que ces données suivent une loi Poisson.

1.5.3 Commentaire à propos du caractère non paramétrique de certains des tests présentés dans ce chapitre

Posons-nous maintenant cette question : est-ce que les tests présentés dans cette section sont paramétriques ou non paramétriques ? La réponse est que les tests pour analyser des données nominales ne sont pas considérés paramétriques. Ainsi, un test sur les paramètres d'une loi binomiale ou multinomiale est un test non paramétrique. Il en est de même pour tout test basé sur la statistique généralisée X^2 ou G^2. Conover (1999) le mentionne dans son livre sur les statistiques non paramétriques. Cependant, un test sur le paramètre d'une loi Poisson est considéré paramétrique.

1.6 Résumé des formules concernant les tableaux de fréquences à une variable

Résumé des informations relatives aux trois distributions étudiées :

	Poisson	**Binomiale**	**Multinomiale**
Échantillon sur l'échelle catégorique	-	Y_1 à $Y_n \in \{\text{succès}, \text{échec}\}$	Y_1 à $Y_n \in \{m_1^Y, \ldots, m_J^Y\}$
Échantillon sur l'échelle numérique	Y_1 à $Y_n \in \{0, 1, 2, \ldots\}$	$\mathbf{1}_u = \begin{cases} 1 & \text{si } Y_u = \text{succès} \\ 0 & \text{si } Y_u = \text{échec} \end{cases}$ pour $u = 1, \ldots, n$	$\mathbf{1}_{uj} = \begin{cases} 1 & \text{si } Y_u = m_j^Y \\ 0 & \text{sinon} \end{cases}$ pour $u = 1, \ldots, n$ et $j = 1, \ldots, J$
Définition de la ou des variables aléatoires (v.a.)	n v.a. indépendantes Y_u : nombre de réalisations d'un événement dans un intervalle de temps et/ou d'espace pour chaque individu de l'échantillon	une seule v.a. $S = \sum_u \mathbf{1}_u$: nombre de succès parmi les n individus de l'échantillon	un vecteur de J v.a. $\boldsymbol{S} = (S_1, \ldots, S_J)$ avec $S_j = \sum_u \mathbf{1}_{uj}$: nombre d'individus de l'échantillon pour lesquels $Y = m_j^Y$, pour $j = 1, \ldots, J$
Tableau des fréquences observées	$\begin{array}{c\|cccc} Y & 0 & 1 & \cdots & \text{total} \\ \hline \text{fréq.} & n_0 & n_1 & \cdots & n \end{array}$	$\begin{array}{c\|cc\|c} Y & \text{succès} & \text{échec} & \text{total} \\ \hline \text{fréq.} & S & n-S & n \end{array}$	$\begin{array}{c\|ccc\|c} Y & m_1^Y & \cdots & m_J^Y & \text{total} \\ \hline \text{fréq.} & S_1 & \cdots & S_J & n \end{array}$
Notation	Y_u *iid* $Poisson(\lambda)$ pour $u = 1, \ldots, n$	$S \sim Bin(n, \pi)$	$\boldsymbol{S} = (S_1, \ldots, S_J) \sim Multinomiale(n, \pi_1, \ldots, \pi_J)$
valeurs possibles	$\{0, 1, 2, \ldots\}$	$\{0, \ldots, n\}$	$\{0, \ldots, n\}$ $\forall$ S_j sous la contrainte que $S_1 + \ldots + S_J = n$
Fonction de masse	$P(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!}$	$P(S = s) = \binom{n}{s}\pi^s(1-\pi)^{n-s}$	$P(\boldsymbol{S} = (s_1, \ldots, s_J)) = \frac{n!}{s_1! \ldots s_J!}\pi_1^{s_1} \ldots \pi_J^{s_J}$
Espérance	$E(Y) = \lambda$	$E(S) = n\pi$	$E(S_j) = n\pi_j$ pour $j = 1, \ldots, J$
Variance	$Var(Y) = \lambda$	$Var(S) = n\pi(1-\pi)$	$Var(S_j) = n\pi_j(1-\pi_j)$ pour $j = 1, \ldots, J$ et $Cov(S_j, S_{j'}) = -n\pi_j\pi_{j'}$ pour $j \neq j'$
Paramètre d'intérêt	$\lambda = E(Y)$	$\pi = P(Y = \text{succès})$	$\boldsymbol{\pi} = (\pi_1, \ldots \pi_J)$ où $\pi_j = P(Y = m_j^Y)$, sous la contrainte $\sum_j \pi_j = 1$
Estimateur max. vrais. du paramètre	$\hat{\lambda} = \frac{\sum_{u=1}^n Y_u}{n}$	$\hat{\pi} = \frac{S}{n}$	$\hat{\pi}_j = \frac{S_j}{n}$ pour $j = 1, \ldots, J$

	Poisson	**Binomiale**	**Multinomiale**
Hypothèses d'un test sur le paramètre	$H_0 : \lambda = \lambda_0$ $H_1 : \lambda \neq$ ou $>$ ou $< \lambda_0$	$H_0 : \pi = \pi_0$ $H_1 : \pi \neq$ ou $>$ ou $< \pi_0$	$H_0 : \boldsymbol{\pi} = \boldsymbol{\pi_0} = (\pi_{0,1}, \ldots, \pi_{0,J})$ $H_1 : \boldsymbol{\pi} \neq \boldsymbol{\pi_0}$
Statistique du test de Wald	$Z_w = \frac{\hat{\lambda}-\lambda_0}{\sqrt{\hat{\lambda}/n}} \xrightarrow[H_0]{\text{asymp.}} \mathcal{N}(0,1)$	$Z_w = \frac{\hat{\pi}-\pi_0}{\sqrt{\hat{\pi}(1-\hat{\pi})/n}} \xrightarrow[H_0]{\text{asymp.}} \mathcal{N}(0,1)$	-
Statistique du test score	$Z_s = \frac{\hat{\lambda}-\lambda_0}{\sqrt{\lambda_0/n}} \xrightarrow[H_0]{\text{asymp.}} \mathcal{N}(0,1)$	$Z_s = \frac{\hat{\pi}-\pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} \xrightarrow[H_0]{\text{asymp.}} \mathcal{N}(0,1)$	$X^2 = \sum_{j=1}^{J} \frac{(S_j - n\pi_{0,j})^2}{n\pi_{0,j}} \xrightarrow[H_0]{\text{asymp.}} \chi^2_{J-1}$
Statistique test rapport vraisemblance (test bilatéral seulement)	$LR = -2n\left(\hat{\lambda}\ln\left(\frac{\lambda_0}{\hat{\lambda}}\right) + (\hat{\lambda}-\lambda_0)\right) \xrightarrow[H_0]{\text{asymp.}} \chi^2_1$	$LR = -2\left(S\ln\left(\frac{\pi_0}{\hat{\pi}}\right) + (n-S)\ln\left(\frac{1-\pi_0}{1-\hat{\pi}}\right)\right) \xrightarrow[H_0]{\text{asymp.}} \chi^2_1$	$G^2 = -2\sum_{j=1}^{J} S_j \ln\left(\frac{\pi_{0,j}}{\hat{\pi}_j}\right) \xrightarrow[H_0]{\text{asymp.}} \chi^2_{J-1}$
Statistique du test exact	-	$S \underset{H_0}{\longrightarrow} Bin(n, \pi_0)$	-
IC Wald de niveau $1-\alpha$	$\hat{\lambda} \pm z_{\alpha/2}\sqrt{\frac{\hat{\lambda}}{n}}$	$\hat{\pi} \pm z_{\alpha/2}\sqrt{\hat{\pi}(1-\hat{\pi})/n}$	correction de Bonferroni : IC individuels de niveau de confiance $1-\alpha/k$ pour les J π_j comme pour le paramètre de la binomiale
IC score de niveau $1-\alpha$	$[L_\lambda, U_\lambda]$ tel que défini sous le tableau	$[L_\pi, U_\pi]$ tel que défini sous le tableau	

$$L_\lambda = \hat{\lambda} + \frac{z^2_{\alpha/2}}{2n} - z_{\alpha/2}\sqrt{\frac{\hat{\lambda}}{n} + \frac{z^2_{\alpha/2}}{4n^2}} \text{ et } U_\lambda = \hat{\lambda} + \frac{z^2_{\alpha/2}}{2n} + z_{\alpha/2}\sqrt{\frac{\hat{\lambda}}{n} + \frac{z^2_{\alpha/2}}{4n^2}},$$
$$L_\pi = \frac{n}{n+z^2_{\alpha/2}}\left(\hat{\pi} + \frac{z^2_{\alpha/2}}{2n} - z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n} + \frac{z^2_{\alpha/2}}{4n^2}}\right) \text{ et}$$
$$U_\pi = \frac{n}{n+z^2_{\alpha/2}}\left(\hat{\pi} + \frac{z^2_{\alpha/2}}{2n} + z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n} + \frac{z^2_{\alpha/2}}{4n^2}}\right).$$

Remarque : Pour trouver les valeurs observées des estimateurs et des statistiques de test, il suffit de remplacer les variables aléatoires S, Y_u ou S_j par leurs valeurs observées s, y_u ou s_j, respectivement.

Formules de statistiques descriptives numériques selon le format des données :

Statistique	**Format individus**	**Format fréquences**
moyenne ($\bar{y}$)	$\frac{\sum_{u=1}^{n} y_u}{n}$	$\frac{\sum_{j=1}^{J} m_j^Y n_j}{n}$
variance	$\frac{\sum_{u=1}^{n}(y_u-\bar{y})^2}{n-1} = \frac{\sum_{u=1}^{n} y_u^2 - n\bar{y}^2}{n-1}$	$\frac{\sum_{j=1}^{J}(m_j^Y)^2 n_j - n\bar{y}^2}{n-1}$

Formes générales pour les statistiques du khi-deux de Pearson (X^2) et du rapport de vraisemblance (G^2) :

$$\chi^2 = \sum_{j=1}^{J} \frac{(O_j - E_j)^2}{E_j} \qquad \text{et} \qquad G^2 = 2\sum_{j=1}^{J} O_j \ln\left(\frac{O_j}{E_j}\right)$$

où les O_j sont des fréquences observées et les E_j sont des fréquences espérées. Sous l'hypothèse nulle que les fréquences espérées sont vraies, les statistiques X^2 et G^2 suivent asymptotiquement une loi du khi-deux à d degrés de liberté. Ces degrés de liberté sont la différence entre le nombre de paramètres libres dans l'espace de toutes les valeurs possibles des fréquences O_1 à O_J et le nombre de paramètres libres sous l'hypothèse nulle.

Remarque : La validité de loi asymptotique des statistiques X^2 et G^2 peut être mise en doute lorsque plus de 20% des fréquences espérées sont inférieures à 5.

Test d'adéquation de données à une loi avec X^2 et G^2 :
Étape préalable : déterminer arbitrairement J classes qui couvrent tout le support des valeurs possibles de la variable à tester. Pour $j = 1, \ldots, J$, O_j est le nombre d'observations de l'échantillon tombant dans la classe J.
Selon le type de l'hypothèse nulle, les statistiques de test sont les suivantes :

type 1 $\rightarrow$ H_0 : La loi $\mathcal{L}(\theta_0)$ s'ajuste bien aux données
X^2 ou $G^2 \xrightarrow[H_0]{\text{asymp.}} \chi^2_{J-1}$
avec $E_j = nP(Y \in \text{ classe } j | Y \sim \mathcal{L}(\boldsymbol{\theta}_0))$

type 2 $\rightarrow$ H_0 : La famille de loi $\mathcal{L}$ s'ajuste bien aux données
X^2 ou $G^2 \xrightarrow[H_0]{\text{asymp.}} \chi^2_{J-1-p}$
avec $E_j = nP(Y \in \text{ classe } j | Y \sim \mathcal{L}(\hat{\boldsymbol{\theta}}))$

où p est le nombre de paramètres de la loi $\mathcal{L}$ que l'on estime à partir des données et $\hat{\boldsymbol{\theta}}$ est l'estimateur du maximum de vraisemblance de $\boldsymbol{\theta}$.

Chapitre 2

Tableaux de fréquences à deux variables : relation entre deux variables catégoriques

Ce chapitre traite d'outils pour répondre à la question suivante :

« Quel est le lien entre les caractéristiques A et B des individus de la population à l'étude ? »

lorsque ces caractéristiques sont représentées par des variables catégoriques. Pour l'instant, on utilise uniquement la variable X qui représente la caractéristique A et la variable Y qui représente la caractéristique B. Dans les prochains chapitres, on verra comment tenir compte de l'effet d'autres variables dans l'étude de la relation entre X et Y.

Pour étudier le lien entre deux variables, il est bon de procéder en suivant les étapes suivantes :

1. **Visualiser les données** : avec un tableau de fréquences à deux variables et/ou des graphiques.

2. **Tester l'association entre X et Y** : on veut déterminer s'il existe un lien entre les variables, on choisira le meilleur test en fonction du caractère nominal ou ordinal des variables, de la taille de l'échantillon et du nombre de modalités des variables.

3. **Si elle est présente, décrire l'association** : avec différentes statistiques et mesures d'association.

Parfois, deux variables sont en fait deux mesures de la même caractéristique (par exemple prises à différents moments dans le temps, prises par deux instruments de mesure, évaluées par deux personnes distinctes, etc.). On parle alors de données pairées et la question de recherche énoncée ci-dessus n'est plus pertinente. On traitera de ces données particulières dans la dernière section de ce chapitre.

Les autres sections du chapitre couvrent des définitions de base, les tests d'association entre deux variables catégoriques nominales, les outils statistiques pour décrire une telle association, puis le cas particulier des variables catégoriques ordinales.

2.1 Définitions et outils descriptifs

Le contexte traité dans ce chapitre est celui où l'on étudie deux variables catégoriques notées X et Y. Soit $m_1^X, \ldots, m_I^X$ les modalités de la première variable X et $m_1^Y, \ldots, m_J^Y$ les modalités de la deuxième variable Y. On possède un échantillon de n observations indépendantes de ces deux variables. À partir de cet échantillon, on calcule un certain nombre de fréquences, que l'on représente la plupart du temps dans un tableau comme suit :

$X \setminus Y$	m_1^Y	$\ldots$	m_j^Y	$\ldots$	m_J^Y	Total
m_1^X	n_{11}	$\ldots$	n_{1j}	$\ldots$	n_{1J}	$n_{1\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
m_i^X	n_{i1}	$\ldots$	n_{ij}	$\ldots$	n_{iJ}	$n_{i\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
m_I^X	n_{I1}	$\ldots$	n_{Ij}	$\ldots$	n_{IJ}	$n_{I\bullet}$
Total	$n_{\bullet 1}$	$\ldots$	$n_{\bullet j}$	$\ldots$	$n_{\bullet J}$	$n_{\bullet\bullet}$

Les fréquences de ce tableau sont définies dans les paragraphes suivants. Un tel tableau de fréquences (ou de contingence) peut porter plusieurs noms :

- tableau de fréquences croisées ;
- tableau de fréquences à deux variables ;
- tableau de fréquences bivariées ;
- tableau de fréquences $I \times J$.

Tous ces noms désignent la même chose. Dans le cas particulier où les deux variables n'ont que deux modalités, on parle souvent de tableau de fréquences 2×2. En anglais, ce tableau est parfois désigné par l'expression « fourfold table ».

Emplacement des variables

Même si la causalité n'est pas prise en compte dans un tableau de fréquences, si on peut identifier une variable réponse (disons Y) et une variable explicative (disons X), il est usuel de placer la variable réponse en colonnes et la variable explicative en lignes. L'emplacement des variables dans le tableau est important pour le calcul de certaines statistiques couvertes dans ce chapitre.

Exemple d'emplacement de variables : intentions de vote selon le sexe

Dans le cadre de l'Enquête Sociale Générale aux États-Unis en 1991, $n = 980$ personnes ont été interrogées, notamment, à propos de leurs intentions de vote. L'enquête a aussi permis d'identifier le sexe des répondants. On a donc deux variables catégoriques nominales :
X = le sexe d'une personne, prenant soit la valeur $m_1^X =$ « Femme », soit la valeur $m_2^X =$ « Homme » et
Y = le parti politique pour lequel une personne à l'intention de voter, soit $m_1^Y =$ « Démocrate », $m_2^Y =$ « Indépendant » ou $m_3^Y =$ « Républicain ».
Puisque le sexe est une caractéristique intrinsèque d'un individu, si une des deux variables influence l'autre ici, c'est certainement le sexe qui influence les intentions de vote, et non l'inverse. Ainsi, si on postulait un lien de causalité entre les variables X et Y, on dirait que la variable réponse est Y, les intentions de vote, et la variable explicative est X, le sexe. La variable X sera donc placée en lignes dans le tableau de fréquences, et Y en colonnes.

Fréquences croisées

Pour $i = 1, ..., I$ et $j = 1, ..., J$, la fréquence n_{ij} est le nombre total d'observations dans l'échantillon pour lesquelles la valeur de X est m_i^X et la valeur de Y est m_j^Y simultanément. Les n_{ij} sont ce que l'on appelle les *fréquences croisées*. Elles se retrouvent dans les cellules centrales d'un tableau de fréquences à deux variables. On a toujours la relation suivante : $n = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$ où n est le nombre total d'observations dans l'échantillon. Ainsi, $n_{\bullet\bullet}$ est la même chose que n.

Exemple de fréquences croisées : intentions de vote selon le sexe

On a observé dans l'échantillon de 980 répondants les fréquences croisées suivantes :

		Y		
		Démocrate	Indépendant	Républicain
X	Femme	$n_{11} = 279$	$n_{12} = 73$	$n_{13} = 225$
	Homme	$n_{21} = 165$	$n_{22} = 47$	$n_{23} = 191$

On constate que, par exemple, le nombre de femmes de l'échantillon ayant l'intention de voter démocrate (n_{11}) vaut 279.

Fréquences marginales

Dans la ligne et la colonne « Total » du tableau général du début de la section, on retrouve les *fréquences marginales* de chacune des variables. Ces fréquences marginales sont en fait les fréquences univariées vues au chapitre 1. Dans un cadre bivarié, $n_{i\bullet} = \sum_{j=1}^{J} n_{ij}$ et $n_{\bullet j} = \sum_{i=1}^{I} n_{ij}$.

Exemple de fréquences marginales : intentions de vote selon le sexe

Ajoutons des marges au tableau de fréquences croisées précédent :

		Y			
		Démocrate	Indépendant	Républicain	Total
X	Femme	279	73	225	577
	Homme	165	47	191	403
	Total	444	120	416	980

Les fréquences marginales du sexe sont ($n_{1\bullet} = 577, n_{2\bullet} = 403$) et les fréquences marginales des intentions de vote sont ($n_{\bullet 1} = 444, n_{\bullet 2} = 120, n_{\bullet 3} = 416$). Ainsi l'échantillon comprend 577 femmes et 403 hommes. Parmi ces individus, 444 ont l'intention de voter démocrate, 120 de voter indépendant et 416 de voter républicain.

Fréquences conditionnelles

Les fréquences des modalités d'une variable en fixant l'autre variable à une certaine modalité sont appelées *fréquences conditionnelles*. Par exemple, les fréquences de X sachant que $Y = 2$ sont n_{12} à n_{I2}. Une ligne ou une colonne à l'intérieur du tableau de fréquences croisées représente donc un groupe de fréquences conditionnelles.

Exemple de fréquences conditionnelles : intentions de vote selon le sexe

Les fréquences des modalités des intentions de vote Y en conditionnant par rapport au fait d'être une femme sont $(n_{11} = 279, n_{12} = 73, n_{13} = 225)$, soit la première ligne du tableau de fréquences croisées.

Fréquences relatives

On peut maintenant définir des *fréquences relatives*. Elles peuvent être croisées, marginales ou conditionnelle, tout comme les fréquences non relatives. On les dits relatives, car on ramène les fréquences à des proportions en les divisant par une fréquence totale.

Pour les fréquences croisées et marginales, la somme des fréquences vaut n. Ainsi, les fréquences relatives croisées sont définies par $f_{ij} = n_{ij}/n$ pour $i = 1, ..., I$ et $j = 1, ..., J$. La somme de ces $I \times J$ fréquences vaut 1. Les fréquences relatives marginales se calculent par $f_{i\bullet} = n_{i\bullet}/n$ pour $i = 1, ..., I$ et $f_{\bullet j} = n_{\bullet j}/n$ pour $j = 1, ..., J$. La somme des I fréquences relatives de la marge verticale vaut 1, tout comme la somme des J fréquences relatives de la marge horizontale.

Finalement, les fréquences relatives conditionnelles dépendent quant à elles de la variable de conditionnement. Ce n'est pas une division par n qui permet de calculer ces fréquences. On fixe une variable à une certaine valeur. Disons que l'on fixe la variable Y à sa $j^{\text{ième}}$ modalité. L'échantillon comporte $n_{\bullet j}$ individus rencontrant cette condition. En d'autres mots, la somme des fréquences conditionnelles à ce que Y vaille m_j^Y est $n_{\bullet j}$. Ainsi, les fréquences relatives conditionnelles de X par rapport à Y sont définies par $f_{i|j} = n_{ij}/n_{\bullet j}$ pour $i = 1, ..., I$ et pour une valeur de j fixe. De façon similaire, les fréquences relatives conditionnelles de Y par rapport à X sont définies par $f_{j|i} = n_{ij}/n_{i\bullet}$ pour $j = 1, ..., J$ et pour une valeur de i fixe.

Exemple de fréquences relatives : intentions de vote selon le sexe

Les fréquences relatives croisées (f_{ij} pour $i = 1, 2$ et $j = 1, 2, 3$) et marginales ($f_{i\bullet}$ pour $i = 1, 2$ et $f_{\bullet j}$ pour $j = 1, 2, 3$) sont les suivantes :

	Démocrate	Indépendant	Républicain	Total
Femme	$f_{11} = \frac{279}{980} = 0.2847$	$f_{12} = \frac{73}{980} = 0.0745$	$f_{13} = \frac{225}{980} = 0.2296$	$f_{1\bullet} = \frac{577}{980} = 0.5888$
Homme	$f_{21} = \frac{165}{980} = 0.1684$	$f_{22} = \frac{47}{980} = 0.0480$	$f_{23} = \frac{191}{980} = 0.1949$	$f_{2\bullet} = \frac{403}{980} = 0.4112$
Total	$f_{\bullet 1} = \frac{444}{980} = 0.4531$	$f_{\bullet 2} = \frac{120}{980} = 0.1224$	$f_{\bullet 3} = \frac{416}{980} = 0.4245$	

Comme il se doit, la somme de toutes les fréquences relatives croisées vaut 1 et la somme des fréquences relatives dans chacune des marges vaut aussi 1. On remarque aussi que les sommes en lignes et en colonnes des fréquences relatives croisées retombent sur les fréquences relatives marginales, car ces fréquences sont toutes issues d'une division par le même nombre, la taille d'échantillon n.

Les fréquences relatives conditionnelles à la valeur de X, le sexe, sont les suivantes :

	Démocrate	Indépendant	Républicain
Femme	$f_{1\|i=1} = \frac{279}{577} = 0.4835$	$f_{2\|i=1} = \frac{73}{577} = 0.1265$	$f_{3\|i=1} = \frac{225}{577} = 0.3899$
Homme	$f_{1\|i=2} = \frac{165}{403} = 0.4094$	$f_{2\|i=2} = \frac{47}{403} = 0.1166$	$f_{3\|i=2} = \frac{191}{403} = 0.4739$

La fréquence relative $f_{1|i=1}$ nous dit, par exemple, que 48.35% des femmes de l'échantillon ont l'intention de voter démocrate. Ici, les fréquences relatives de chaque ligne somment à 1, car ces fréquences sont calculées en divisant les fréquences croisées par les fréquences dans la marge verticale. Cette marge est celle de la variable X.

Les fréquences relatives conditionnelles à la valeur de Y, les intentions de vote, sont les suivantes :

	Démocrate	Indépendant	Républicain
Femme	$f_{1\|j=1} = \frac{279}{444} = 0.6284$	$f_{2\|j=2} = \frac{73}{120} = 0.6083$	$f_{3\|j=3} = \frac{225}{416} = 0.5409$
Homme	$f_{1\|j=1} = \frac{165}{444} = 0.3716$	$f_{2\|j=2} = \frac{47}{120} = 0.3917$	$f_{3\|j=3} = \frac{191}{416} = 0.4591$

Ici, les fréquences relatives de chaque colonne somment à 1, car ces fréquences sont calculées en divisant les fréquences croisées par les fréquences dans la marge horizontale. Cette marge est celle de la variable Y.

Probabilités d'intérêt

Nous allons définir ici une notation pour représenter les probabilités en lien avec un tableau de fréquences à deux variables croisant les variables X et Y. Selon la nature des variables X et Y, on s'intéressera à certaines probabilités plutôt que d'autres. Cependant, à cause de la façon dont les données ont été recueillies (type d'échantillonnage vu plus loin), il ne sera pas toujours possible d'estimer les probabilités d'intérêt à partir des données (voir section 2.1.2). Les probabilités en lien avec un tableau de fréquences à deux variables sont les suivantes, pour $i = 1, \ldots, I$ et $j = 1, \ldots, J$:

probabilités conjointes :	$\pi_{ij} = P(X = m_i^X, Y = m_j^Y)$
probabilités marginales :	$\pi_{i\bullet} = P(X = m_i^X)$ $\pi_{\bullet j} = P(Y = m_j^Y)$
probabilités conditionnelles :	$\pi_{i\|j} = P(X = m_i^X \| Y = m_j^Y)$ $\pi_{j\|i} = P(Y = m_j^Y \| X = m_i^X)$

Exemple de probabilités d'intérêt : intentions de vote selon le sexe

Exemples de probabilités qui pourraient nous intéresser :

Type	Définition de la probabilité
conjointe	$\pi_{11} = P$(être une femme et voter démocrate)
marginale	$\pi_{1\bullet} = P$(être une femme) $\pi_{\bullet 1} = P$(voter démocrate)
conditionnelle	$\pi_{1\|j=1} = P$(être une femme | vote démocrate) $\pi_{1\|i=1} = P$(voter démocrate | sexe féminin)

2.1.1 Types d'échantillonnage

Les données à représenter dans un tableau de fréquences bivariées peuvent être collectées selon différents modes d'échantillonnage. Deux caractéristiques seront utilisées ici pour désigner la façon dont l'échantillonnage est effectué :

- Est-ce que la taille d'échantillon est fixe ou non ?
- Est-ce qu'un seul ou plusieurs échantillons sont tirés ?

Échantillonnage multinomial versus Poisson

Lorsque l'on tire un échantillon afin de mesurer deux variables catégoriques X et Y, si la taille de l'échantillon est fixée avant la collecte des données, on dit que l'échantillonnage est multinomial. C'est souvent le cas dans des enquêtes par sondage. Si, au contraire, on échantillonne les sujets selon la survenue d'un certain événement dans un intervalle de temps ou d'espace, sans fixer préalablement la taille de l'échantillon, on parle d'échantillonnage Poisson.

Exemple d'échantillonnage multinomial : Dans l'exemple des intentions de vote selon le sexe traité jusqu'ici dans ce chapitre, les données proviennent d'un échantillonnage multinomial, car la taille totale de l'échantillon $n = 980$ est considérée fixe dans l'Enquête Sociale Générale.

Exemple d'échantillonnage Poisson : On observe les accidents dans la côte de l'autoroute Robert-Bourassa en 2004. À partir des rapports de police des accidents, on relève les informations quant au sexe du conducteur (variable X) et aux conditions routières lors de l'accident (variable Y). On pourrait, par exemple, obtenir les données suivantes :

	Conditions routières		
	difficiles	normales	Total
Chauffeur homme	11	5	16
Chauffeur femme	9	4	13
Total	20	9	29

Au total, 29 accidents sont inclus dans l'échantillon. Ce nombre n'était pas prédéterminé. Il est plutôt représentatif du nombre d'accidents à survenir dans l'intervalle de temps et d'espace choisi, et pour lesquels un rapport de police a été produit.

Interprétation statistique : Les échantillonnages multinomial et Poisson portent ces noms en raison de la distribution postulée pour les fréquences du tableau croisant les variables X et Y. En échantillonnage multinomial, on suppose que le vecteur de toutes les fréquences du tableau suit une loi

multinomiale :

$$(n_{ij}, i = 1, \ldots, I; j = 1, \ldots, J) \sim$$
$$Multinomiale(n, \pi_{ij}, i = 1, \ldots, I; j = 1, \ldots, J).$$

En échantillonnage Poisson, on suppose plutôt que les $I \times J$ fréquences du tableau sont des variables aléatoires Poisson indépendantes :

$$n_{ij} \sim Poisson(\lambda_{ij}) \qquad \text{indépendantes pour} \quad i = 1, \ldots, I \quad \text{et} \quad j = 1, \ldots, J.$$

Notons qu'en échantillonnage multinomial, on ne peut pas postuler l'indépendance de ces fréquences, car la taille d'échantillon n fixe induit la contrainte que la somme des fréquences soit égale à n. En conséquence, n'importe laquelle de ces fréquences croisées peut être déduite des $IJ - 1$ autres.

Notons aussi qu'au chapitre 1, on a mentionné que la loi Poisson servait à modéliser des dénombrements. Elle est ici utilisée pour modéliser des fréquences. Un dénombrement est un nombre de réalisations d'un événement par individu. Une fréquence est plutôt le nombre d'individus pour lesquels une certaine condition est observée.

Lien entre les deux types d'échantillonnage : En échantillonnage Poisson, si on suppose à postériori que la taille d'échantillon n était fixée (bien que l'on sache qu'elle ne l'était pas en réalité), on retombe sur le cas de l'échantillonnage multinomial. En effet, on peut prouver que, sous la contrainte $\sum_{ij} n_{ij} = n$, on a :

$(n_{ij}, i = 1, \ldots, I; j = 1, \ldots, J)$
$\sim Multinomiale(n, \frac{\lambda_{ij}}{\sum_{ij} \lambda_{ij}}, i = 1, \ldots, I; j = 1, \ldots, J),$

(plus de détails dans Agresti, 2002, section 1.2.5).

En pratique, l'échantillonnage Poisson est moins courant que l'échantillonnage multinomial. Cependant, en théorie, l'échantillonnage Poisson facilite le développement algébrique de certaines statistiques. Plusieurs formules de variance complexes pour différentes mesures peuvent s'obtenir aussi bien à partir de la loi multinomiale que de la loi Poisson. Les dérivations mathématiques avec la loi Poisson sont beaucoup plus simples à cause de l'hypothèse d'indépendance entre les fréquences. Cependant, dans ce cours, ce type d'échantillonnage sera peu utilisé.

Échantillonnage simple versus multiple

Si on tire un seul échantillon de la population cible de l'étude, on dit que l'échantillonnage est simple. Cependant, on veut parfois s'assurer d'avoir suffisamment d'individus ayant certaines caractéristiques. Dans ce cas, on détermine d'abord des groupes de sujets à échantillonner, que l'on appellera ici sous-populations (on parle souvent de strates), puis on échantillonne aléatoirement des sujets dans chacune des sous-populations. On effectue donc plusieurs échantillonnages indépendants, un par sous-population.

Exemple d'échantillonnage simple : Les exemples d'échantillonnage multinomial et Poisson présentés précédemment sont deux exemples d'échantillonnage simple. En effet, dans les deux cas, un seul échantillon est tiré à partir de la population cible.

Exemple d'échantillonnage multiple : Statistique Canada mène plusieurs enquêtes couvrant tout le Canada. Par exemple, un volet de l'Enquête Sociale Générale canadienne concerne la victimisation et cherche, en autres, à estimer la proportion de gens victimes d'actes criminels. Statistique Canada souhaite comparer ces proportions entre les provinces. Pour ce faire, on pourrait croiser dans un tableau de fréquences les variables catégoriques suivantes :
Y = une indicatrice d'avoir été victime ou non d'au moins un acte criminel dans la dernière année et
X = la province de résidence.
Il est important pour Statistique Canada d'obtenir des estimations suffisamment précises partout. Cependant, un échantillonnage aléatoire simple parmi tous les Canadiens risquerait de ne pas contenir suffisamment de gens provenant des plus petites provinces, notamment l'Île-du-Prince-Édouard, afin de bien estimer les proportions dans ces provinces. Statistique Canada utilise donc souvent l'échantillonnage stratifié, en stratifiant sur les provinces. C'est un autre nom pour l'échantillonnage multiple, dans lequel les sous-populations sont les provinces.

Dans le dernier exemple, la formation des sous-populations (la stratification) a été faite à partir de la variable X. Si on peut identifier une variable

explicative X et une variable réponse Y, la stratification est typiquement faite en fonction de la variable explicative X. Ainsi, si on a respecté la règle de mettre X en lignes et Y en colonnes, chaque ligne du tableau de fréquences représente un échantillon d'une sous-population.

Pour aller plus loin dans notre présentation de l'échantillonnage multiple, nous allons maintenant utiliser en exemple des études typiques dans le domaine médical, notamment l'essai clinique et l'étude cas-témoin. L'annexe B décrit un peu plus en détail en quoi consistent ces études. Dans un essai clinique, on veut comparer des individus exposés à un certain facteur de risque de développer une maladie, à d'autres non exposés. La variable X représente l'exposition au facteur de risque et la variable Y représente le développement ou non de la maladie. Dans ce type d'étude, on fait un échantillonnage multiple, car les tailles des groupes d'exposition au facteur de risque sont prédéterminées.

Exemple d'étude expérimentale similaire à un essai clinique : Une étude à propos de la relation entre la prise quotidienne d'aspirine et l'infarctus du myocarde a été menée à l'école de médecine de l'Université Harvard (Steering Committee of the Physicians' Health Study Research Group, 1989). Au total, 22071 médecins américains ont participé à l'étude, d'une durée de 5 ans. Les participants ont été attribués de façon aléatoire, en nombres presque égaux, à l'un des deux groupes d'exposition au facteur de risque : ceux qui prennent de l'aspirine quotidiennement et ceux qui n'en prennent pas. Pendant l'étude, tous les médecins devaient prendre une pilule par jour sans savoir s'il s'agissait d'une aspirine ou d'un placebo. On a observé si les participants ont été victimes ou non d'un infarctus pendant l'étude. Voici les résultats obtenus :

X	Y : Infarctus		
Groupe	Oui	Non	Total
Placebo	239	10795	11034
Aspirine	139	10898	11037
Total	378	21693	22071

Ici, la variable explicative X est la prise quotidienne d'aspirine ou non. La variable réponse Y est la survenue ou non d'un infarctus. L'échantillonnage

est multiple, avec des sous-populations formées conditionnellement à la variable explicative X. Aussi, l'échantillonnage est considéré multinomial, car les tailles des deux échantillons ont été fixées. Ce sont les chercheurs qui ont décidé de mettre approximativement le même nombre de participants dans chaque groupe.

Il existe une exception notable à la règle du conditionnement par rapport à la variable explicative : l'étude cas-témoin. Dans ce type d'étude, la formation des sous-populations se fait par rapport à la variable réponse Y qui est dans ce cas la survenue d'une maladie. Les nombres de cas et de témoins (les deux sous-populations) sont préétablis. Ainsi, dans une étude cas-témoin, si on a respecté la règle de mettre X en lignes et Y en colonnes, c'est chaque colonne du tableau de fréquences qui représente un échantillon d'une sous-population.

Exemple d'étude cas-témoin : Une étude a été menée pour étudier la relation entre le fait de fumer et l'infarctus du myocarde. L'étude s'est limitée à la population des femmes italiennes de moins de 70 ans. On voulait s'assurer d'avoir un assez grand nombre de victimes d'un infarctus dans l'échantillon. On a donc divisé la population en deux groupes : les femmes ayant déjà été victimes d'un infarctus et celles n'ayant jamais été victimes d'un infarctus. Pour rejoindre la sous-population des femmes ayant déjà été victime d'un infarctus, on a recruté des participantes parmi des patientes d'hôpitaux italiens, entre 1983 et 1988, admises à l'hôpital en raison d'un infarctus. Pour chaque participante recrutée dans cette sous-population, qu'on appelle les cas, on a recruté deux témoins, soit deux femmes n'ayant jamais été victimes d'un infarctus. Ces femmes ont été recrutées parmi les patientes des mêmes hôpitaux, dans les mêmes années, souffrant d'un autre trouble grave. On a ainsi obtenu un échantillon de 262 cas et 519 témoins. Le nombre de témoins n'est pas exactement égal au double du nombre de cas, probablement à cause de l'abandon de certaines participantes initiales. On a demandé aux participantes si elles étaient ou avaient déjà été fumeuses. Les données observées sont les suivantes (Gramenzi *et al.*, 1989) :

À déjà fumé	Infarctus du myocarde	Groupe témoin	Total
Oui	172	173	345
Non	90	346	436
Total	262	519	781

Ainsi les deux variables croisées sont dichotomiques. La variable X, considérée comme une variable explicative potentielle, est une indicatrice du fait d'avoir déjà fumé et la variable Y, considérée comme la variable réponse, est une indicatrice du fait d'avoir été victime d'un infarctus. L'échantillonnage est donc multiple, avec des sous-populations formées conditionnellement à la variable réponse Y. Aussi, l'échantillonnage semble Poisson, car on a recruté le plus de participantes possible dans certains hôpitaux, pendant certaines années. Par contre, on a fixé $n_{\bullet 1} = 2n_{\bullet 2}$. Ainsi, il y a une contrainte dans la marge de Y et, en conséquence, le modèle mathématique le plus juste est celui de l'échantillonnage multinomial multiple.

Interprétation statistique de l'échantillonnage multiple : On suppose que les sous-populations sont indépendantes, donc que les vecteurs des fréquences croisées dans les sous-populations sont indépendants. Si les tailles des échantillons ne sont pas fixes, on dira aussi que les fréquences théoriques à l'intérieur de chaque sous-population sont des variables aléatoires indépendantes, toutes de loi Poisson. Ainsi, un échantillonnage Poisson multiple aboutit au même modèle mathématique qu'un échantillonnage Poisson simple.

Cependant, si les tailles des échantillons sont fixes, on postulera que les vecteurs des fréquences croisées dans chaque sous-population suivent des distributions multinomiales indépendantes.

Dans le cas où la stratification a été faite par rapport à la variable X, on suppose qu'il y a I sous-populations indépendantes telles que :

$$(n_{i1}, \ldots, n_{iJ}) \sim Multinomiale(n_i, \pi_{1|i}, \ldots, \pi_{J|i}) \quad \text{pour } i = 1, \ldots, I$$

où les n_i sont en fait la même chose que les $n_{i\bullet}$ vus auparavant, mais considérés fixes. Le vecteur de probabilités $(\pi_{1|i}, \ldots, \pi_{J|i})$ est en fait la fonction de masse de Y conditionnellement à ce que X prenne la valeur m_i^X.

Si la stratification a plutôt été faite par rapport à la variable Y, comme dans une étude cas-témoin, on suppose qu'il y a J sous-populations indépendantes telles que :

$$(n_{1j}, \ldots, n_{Ij}) \sim Multinomiale(n_j, \pi_{1|j}, \ldots, \pi_{I|j}) \quad \text{pour } j = 1, \ldots, J$$

où les n_j sont la même chose que les $n_{\bullet j}$ vus auparavant, mais considérés fixes. Les paramètres $\pi_{1|j}$ à $\pi_{I|j}$ sont les probabilités pour X de prendre cha-

cune de ses modalités possibles, étant donné que Y prenne la valeur m_j^Y. Ces probabilités sont donc conditionnelles à être dans la $j^{\text{ième}}$ sous-population.

2.1.2 Estimation des probabilités d'intérêt

On souhaite connaître la valeur des probabilités d'intérêt. Ces probabilités sont en fait des proportions dans la population à l'étude. Que faire si on ne possède pas des observations pour tous les individus de la population, mais seulement pour un échantillon aléatoire d'individus provenant de cette population ? On va utiliser l'échantillon pour estimer les probabilités.

Pour dériver des estimateurs du maximum de vraisemblance des probabilités d'intérêt, nous allons uniquement considérer le mode d'échantillonnage multinomial. C'est donc dire que nous considérons les tailles d'échantillons fixes. Si des données à traiter provenaient en réalité d'un échantillonnage Poisson, on conditionnerait sur la ou les valeurs des tailles d'échantillons. Ainsi, on retomberait sur de l'échantillonnage multinomial et les estimateurs présentés ici seraient aussi pertinents.

Dans cette sous-section, on se ramène souvent à de la matière vue au chapitre 1. Ainsi, pour toutes les probabilités d'intérêt, on va uniquement voir ici comment faire une estimation ponctuelle. Si on souhaite faire un test de conformité sur la valeur d'une proportion ou d'un vecteur de proportions, ou encore si on veut faire des intervalles de confiance pour ces proportions, on peut déduire comment faire à partir du chapitre 1. Il est donc inutile de présenter de nouveau cette matière ici.

Le but de cette section n'est pas uniquement d'apprendre à faire de l'inférence sur les probabilités d'intérêt. Elle vise aussi à mettre en lumière les circonstances dans lesquelles certaines de ces probabilités ne sont pas estimables à partir des données.

Probabilités conjointes

La probabilité conjointe π_{ij}, pour $i = 1, \ldots, I$ et $j = 1, \ldots, J$, est simplement la probabilité que, simultanément, la variable aléatoire X prenne la valeur m_i^X et la variable aléatoire Y prenne la valeur m_j^Y. Les π_{ij} sont facilement estimables si les données proviennent d'un échantillonnage multinomial simple. Dans ce cas, le vecteur de toutes les fréquences du tableau suit une

loi multinomiale :

$$[n_{ij}] \sim Multinomiale(n, [\pi_{ij}]).$$

On voit que les probabilités conjointes sont des paramètres de cette distribution. Ainsi, de ce que l'on a appris sur la distribution multinomiale au chapitre 1, on peut tirer des estimateurs de maximum de vraisemblance des π_{ij} comme suit :

$$\hat{\pi}_{ij} = \frac{n_{ij}}{n} \qquad \text{pour } i = 1, \ldots, I \text{ et } j = 1, \ldots, J.$$

Il s'agit donc des fréquences relatives croisées f_{ij}. Cependant, on ne les voit pas ici comme des valeurs observées, mais plutôt comme des quantités aléatoires, puisque calculées à partir d'un échantillon aléatoire.

Exemple d'estimation de probabilités conjointes :
intentions de vote selon le sexe

Revenons à l'exemple des intentions de vote selon le sexe. On estime les probabilités conjointes par les fréquences relatives croisées. Les valeurs observées des estimateurs des probabilités conjointes obtenues dans cette enquête sont donc les suivantes :

	Démocrate	Indépendant	Républicain
Femme	$\hat{\pi}_{11} = f_{11} = 0.2847$	$\hat{\pi}_{12} = f_{12} = 0.0745$	$\hat{\pi}_{13} = f_{13} = 0.2296$
Homme	$\hat{\pi}_{21} = f_{21} = 0.1684$	$\hat{\pi}_{22} = f_{22} = 0.0480$	$\hat{\pi}_{23} = f_{23} = 0.1949$

Ainsi, on estimait à 28.47%, en 1991, la probabilité d'être à la fois une femme et d'avoir l'intention de voter démocrate dans la population cible qui est ici toute la population des États-Unis. Remarquez que lorsque l'on avait introduit les fréquences observées, on ne faisait pas d'inférence sur une population cible. On interprétait la fréquence relative croisée f_{11} simplement en disant que 28.47% de l'échantillon était constitué de femmes ayant l'intention de voter démocrate.

Attention, *si l'échantillonnage est multinomial multiple, les données ne nous permettent pas d'estimer ces probabilités.* En effet, étant donné que les tailles des sous-populations ont été fixées, elles ne sont plus représentatives

des nombres d'individus dans la population cible globale appartenant à ces sous-populations. Nous verrons à la fin de cette section un exemple illustrant ce bais introduit par l'échantillonnage.

Probabilités marginales

Les probabilités marginales $\pi_{i\bullet}$ et $\pi_{\bullet j}$, pour $i = 1, \ldots, I$ et $j = 1, \ldots, J$, sont simplement les fonctions de densité univariées des variables X et Y respectivement. On peut les estimer lorsque les données proviennent d'un échantillonnage multinomial simple, car dans ce cas la distribution multinomiale du vecteur complet des fréquences croisées implique que les deux vecteurs de fréquences marginales suivent aussi des distributions multinomiales :

$$\begin{aligned}(n_{1\bullet}, \ldots, n_{I\bullet}) &\sim Multinomiale(n, \pi_{1\bullet}, \ldots, \pi_{I\bullet}) \\ (n_{\bullet 1}, \ldots, n_{\bullet J}) &\sim Multinomiale(n, \pi_{\bullet j}, \ldots, \pi_{\bullet J})\end{aligned}$$

Les probabilités conjointes sont des paramètres de ces distributions et, en conséquence, les estimateurs de maximum de vraisemblance des probabilités marginales sont les suivants :

$$\begin{aligned}\hat{\pi}_{i\bullet} &= \frac{n_{i\bullet}}{n} \qquad \text{pour } i = 1, \ldots, I, \\ \hat{\pi}_{\bullet j} &= \frac{n_{\bullet j}}{n} \qquad \text{pour } j = 1, \ldots, J.\end{aligned}$$

Il s'agit donc des fréquences relatives marginales $f_{i\bullet}$ et $f_{\bullet j}$, vues comme des quantités aléatoires.

Exemple d'estimation de probabilités marginales :
intentions de vote selon le sexe

On estime ici la probabilité pour un individu vivant aux États-Unis d'être une femme par $\hat{\pi}_{1\bullet} = 0.5888$ et la probabilité d'être un homme par $\hat{\pi}_{2\bullet} = 0.4112$. Ces estimations ne semblent pas être tout à fait exactes ! Le recensement américain de 2010 dit plutôt que la population des États-Unis est composée à 49.2% d'hommes et à 50.8% de femmes. Pourtant, l'Enquête Sociale Générale contacte autant d'hommes que de femmes pour répondre à son questionnaire. On constate ici que proportionnellement plus de femmes ont accepté de répondre au sondage que d'hommes. C'est typique dans un

sondage du genre. On pourrait parler d'un biais de sélection qui rend erronée l'estimation de la proportion d'hommes et de femmes dans la population américaine à partir des données de cette enquête. Pourtant, l'échantillonnage n'était pas multiple. On n'a pas stratifié la population selon le sexe.

ATTENTION, *si l'échantillonnage est multinomial multiple, seulement un des deux vecteurs de probabilités marginales est estimable à partir des données.* Si la stratification s'est faite à partir de la variable X, les fréquences marginales $(n_{1\bullet}, \ldots, n_{I\bullet})$ ne sont pas aléatoires. Elles sont fixes et on les notent $(n_1, \ldots, n_I)$. Il serait incorrect de s'en servir pour estimer les $\pi_{i\bullet}$ puisque ces valeurs, ayant été fixée préalablement, ne sont pas nécessairement représentatives de la population cible.

Cependant, les $\pi_{\bullet j}$ sont bien estimables. Elles s'estiment de la même façon que si l'échantillonnage était simple, car on a que :

$$(n_{i1}, \ldots, n_{iJ}) \sim Multinomiale(n_i, \pi_{1|i}, \ldots, \pi_{J|i}) \quad \text{pour } i = 1, \ldots, I.$$

En sommant les I vecteurs de fréquences $(n_{i1}, \ldots, n_{iJ})$, on obtient le vecteur de fréquences marginales $(n_{\bullet 1}, \ldots, n_{\bullet J})$ qui suit alors la distribution $Multinomiale(n, \pi_{\bullet 1}, \ldots, \pi_{\bullet J})$. Le premier paramètre de la distribution est devenu $n = \sum_i n_i$. Les autres paramètres sont maintenant les probabilités formant la fonction de masse de Y peu importe la valeur de X.

Probabilités conditionnelles

Les probabilités conditionnelles nous intéressent particulièrement si on peut identifier X comme étant une variable explicative et Y comme étant une variable réponse. Dans ce cas, leurs valeurs de ces probabilités révèlent si les variables X et Y sont associées ou non. La comparaison des I vecteurs de probabilités $(\pi_{1|i}, \ldots, \pi_{J|i})$ où, on le rappelle, $\pi_{j|i} = P(Y = m_j^Y | X = m_i^X)$, dévoilera si les variables sont liées. Si ces I vecteurs sont égaux, les variables ne sont pas associées. Au contraire, si ces I vecteurs diffèrent, c'est le signe qu'il y a une association entre X et Y : la valeur de X influence la fonction de densité de Y.

Comment pouvons-nous estimer les probabilités conditionnelles $\pi_{j|i}$ à partir des données ? Si celles-ci proviennent d'un échantillonnage multinomial simple, on conditionne sur les valeurs des fréquences marginales. On considère donc la marge en X fixe. On peut ainsi déduire la distribution des

vecteurs de fréquences conditionnelles :

$$(n_{i1}, \ldots, n_{iJ}) \sim Multinomiale(n_i, \pi_{1|i}, \ldots, \pi_{J|i}) \quad \text{pour } i = 1, \ldots, I$$

où $n_i = n_{i\bullet}$. On obtient donc les estimateurs du maximum de vraisemblance suivants :

$$\hat{\pi}_{j|i} = \frac{n_{ij}}{n_i}.$$

Il s'agit de fréquences relatives conditionnelles $f_{j|i}$, considérées aléatoires.

Ce résultat demeure vrai si l'échantillonnage est multinomial multiple et que la stratification a été faite par rapport à la variable X. En fait, avec un échantillonnage multinomial multiple, on peut estimer les probabilités conditionnelles à la variable de stratification, mais pas celles conditionnelles à l'autre variable.

Exemple d'estimation de probabilités conditionnelles :
intentions de vote selon le sexe

Estimons maintenant les probabilités pour les Américaines, et celles pour les Américains, de voter démocrate, indépendant ou républicain. Les valeurs observées à partir de l'échantillon des estimateurs de ces probabilités sont les suivantes :

	Démocrate	Indépendant	Républicain
Femme	$\hat{\pi}_{1\|i=1} = f_{1\|i=1} = 0.4835$	$\hat{\pi}_{2\|i=1} = f_{2\|i=1} = 0.1265$	$\hat{\pi}_{3\|i=1} = f_{3\|i=1} = 0.3899$
Homme	$\hat{\pi}_{1\|i=2} = f_{1\|i=2} = 0.4094$	$\hat{\pi}_{2\|i=2} = f_{2\|i=2} = 0.1166$	$\hat{\pi}_{3\|i=2} = f_{3\|i=2} = 0.4739$

Selon ces estimations, la probabilité pour une femme de voter démocrate (0.4835) semble plus grande que celle pour un homme (0.4094). À l'inverse, les hommes semblent avoir plus l'intention de voter républicain que les femmes.

Ces estimations nous permettent de comparer les intentions de vote selon le sexe. Pour comparer le sexe selon les intentions de vote, nous aurions utilisé les probabilités conditionnelles à l'intention de vote. Cependant, étant donné que nous savons que l'échantillon comporte proportionnellement plus de femmes que la population cible, les proportions conditionnelles d'être une femme dans chacune des trois sous-populations d'intentions de vote auraient surestimé ces proportions dans la population générale.

ATTENTION, *si la stratification a été faite par rapport à la variable Y, on ne peut pas estimer les probabilités conditionnelles $\pi_{j|i}$ à partir des données.* On peut cependant estimer les probabilités conditionnelles à la valeur de Y, soit les $\pi_{i|j}$. Par contre, si Y est la variable réponse et X la variable explicative, ces probabilités ne nous intéressent pas ! On a ce genre de problème dans une étude cas-témoin. On s'intéresse aux probabilités conditionnelles $\pi_{j|i}$ afin de savoir si X influence Y. Cependant, on ne peut pas estimer ces probabilités à partir des données.

Exemple de probabilités d'intérêt non estimables à partir des données : étude cas-témoin

Revenons sur l'étude cas-témoin présentée précédemment. Les données observées étaient les suivantes :

À déjà fumé	Infarctus du myocarde	Groupe témoin	Total
Oui	172	173	345
Non	90	346	436
Total	262	519	781

Ici, l'échantillonnage est multiple, conditionnellement à la variable Y, soit la survenue d'un infarctus. On va considérer le nombre de cas ($n_{\bullet 1} = 262$) et le nombre de témoins ($n_{\bullet 2} = 519$) comme étant fixes. On va donc pouvoir se baser sur des distributions multinomiales pour estimer les probabilités d'intérêt.

Le tableau suivant présente des exemples de probabilités d'intérêt accompagnées de leurs estimations potentielles.

Définition d'une probabilité		Estimation potentielle		Bonne
générale	exemple	générale	exemple	estim. ?
π_{11}	P(fumeuse et infarctus)	n_{11}/n	$172/781 = 0.22$	non
$\pi_{1\bullet}$	P(fumeuse)	$n_{1\bullet}/n$	$345/781 = 0.44$	oui
$\pi_{\bullet 1}$	P(infarctus)	$n_{\bullet 1}/n$	$262/781 = 0.34$	non
$\pi_{1\|j=1}$	P(fumeuse | infarctus)	$n_{11}/n_{\bullet 1}$	$172/262 = 0.66$	oui
$\pi_{1\|i=1}$	P(infarctus | fumeuse)	$n_{11}/n_{1\bullet}$	$172/345 = 0.50$	non

Essayons maintenant d'utiliser notre logique pour évaluer si ces estimations sont plausibles. Pour ce faire, il faut d'abord savoir que l'infarctus est un

problème de santé plutôt rare. D'autres études, notamment celle à propos de la relation entre la prise quotidienne d'aspirine et le risque d'infarctus présentée précédemment, ont permis d'estimer que la probabilité de subir un infarctus est de l'ordre de 1% ou 2% dans la population générale. On estimerait ici cette probabilité à 34% pour la population cible des femmes italiennes de moins de 70 ans dans les années 80 ! Ce pourcentage est beaucoup trop élevé pour être plausible. Il prend cette valeur uniquement parce que les chercheurs menant cette étude avaient préalablement fixé à environ un tiers la proportion de cas (gens ayant subi un infarctus) dans leur échantillon. De même, il est tout à fait illogique d'estimer à 22% la proportion des individus de cette population cible qui ont déjà subi un infarctus et qui ont déjà fumé, et d'estimer à 50% la proportion des fumeuses ou ex-fumeuses qui ont déjà subi un infarctus. On voit clairement ici que ces estimations ne sont pas bonnes. Cependant, estimer la proportion de gens ayant déjà fumé dans la population cible de l'étude à 44% est plausible, tout comme estimer à 66% la proportion de gens ayant déjà fumé parmi les individus de la population cible qui ont déjà subi un infarctus.

Pour répondre à la question de recherche « Est-ce que la cigarette augmente le risque d'infarctus ? », on pourrait comparer les probabilités conditionnelles de subir un infarctus pour les femmes ayant déjà fumé ($\pi_{1|i=1}$) à celle des femmes n'ayant jamais fumé ($\pi_{1|i=2}$). Cependant, on a ici un problème : ces probabilités ne sont pas estimables ! On verra dans ce chapitre des stratégies pour répondre à la question de recherche malgré ce problème.

Résumé : Voici un tableau résumant quels sont les estimateurs potentiels des probabilités mentionnées dans cette sous-section. Pour les échantillonnages multinomiaux simple et Poisson, ces estimateurs sont bons, à condition qu'on n'ait pas rencontré de biais de sélection ou autres biais comme dans l'exemple des intentions de vote selon le sexe. Cependant, pour un échantillonnage multinomial multiple, il est certain qu'un biais vient empêcher l'estimation correcte de certaines probabilités à partir de l'échantillon. La dernière colonne du tableau dit dans quelles circonstances les estimateurs sont appropriés pour ce type d'échantillonnage.

Type de probabilité	Probabilité	Estimateur potentiel	Bon estimateur si éch. multiple
conjointe	π_{ij}	$\hat{\pi}_{ij} = n_{ij}/n$	jamais
marginale	$\pi_{i\bullet}$	$\hat{\pi}_{i\bullet} = n_{i\bullet}/n$	si var. stratif. $= Y$
	$\pi_{\bullet j}$	$\hat{\pi}_{\bullet j} = n_{\bullet j}/n$	si var. stratif. $= X$
conditionnelle	$\pi_{i\|j}$	$\hat{\pi}_{i\|j} = n_{ij}/n_{\bullet j}$	si var. stratif. $= Y$
	$\pi_{j\|i}$	$\hat{\pi}_{j\|i} = n_{ij}/n_{i\bullet}$	si var. stratif. $= X$

Rappelons que les probabilités, désignées par la lettre grecque π, sont en fait des proportions dans la population. Les estimateurs, notés $\hat{\pi}$, sont pour leurs parts des proportions dans l'échantillon. On considère ces quantités aléatoires puisque l'échantillon est aléatoire.

2.1.3 Qu'est-ce que l'association entre deux variables catégoriques ?

De façon générale, on dira qu'il y a un lien ou une association entre deux variables si modifier la valeur d'une variable affecte la valeur de l'autre variable. Le terme « association » est plutôt général. Pour des variables catégoriques, si la fonction de densité d'une variable varie conditionnellement à la valeur de l'autre variable, on peut dire que la valeur d'une variable influence la valeur de l'autre variable. Elles seraient donc associées. La fonction de masse d'une variable, disons Y, conditionnelle à la valeur de l'autre variable, disons X, est le vecteur $(\pi_{1|i}, \ldots, \pi_{J|i})$. On a I vecteurs de ce type, pour les I modalités de X. Chacun de ces vecteurs s'estime par le vecteur des fréquences relatives conditionnelles à X $(f_{1|i}, \ldots, f_{J|i})$. Ainsi, un signe que deux variables sont reliées est que les vecteurs de fréquences relatives conditionnelles $(f_{1|i}, \ldots, f_{J|i})$ diffèrent selon la valeur de X.

On peut aussi faire le même raisonnement en conditionnant par rapport à Y, mais si X est vraiment une variable pouvant influencer Y, le conditionnement par rapport à X a plus de sens.

Certains concepts mathématiques sont reliés à la notion d'association ou de non-association entre deux variables, par exemple l'indépendance et l'homogénéité de sous-populations. Leurs définitions seront vues plus loin. Pour

l'instant, on cherche seulement à dire de façon exploratoire si les données d'un tableau de fréquences présentent une association ou non entre les variables.

Exemple d'étude exploratoire de l'association entre deux variables catégoriques : intentions de vote selon le sexe

Nous avons auparavant estimé les deux vecteurs de probabilités des intentions de votes conditionnellement au sexe. Nous avons obtenu :

	Démocrate	Indépendant	Républicain
Femme	$\hat{\pi}_{1\|i=1} = f_{1\|i=1} = 0.4835$	$\hat{\pi}_{2\|i=1} = f_{2\|i=1} = 0.1265$	$\hat{\pi}_{3\|i=1} = f_{3\|i=1} = 0.3899$
Homme	$\hat{\pi}_{1\|i=2} = f_{1\|i=2} = 0.4094$	$\hat{\pi}_{2\|i=2} = f_{2\|i=2} = 0.1166$	$\hat{\pi}_{3\|i=2} = f_{3\|i=2} = 0.4739$

À cause de la probabilité pour une femme de voter démocrate que l'on estime plus grande que celle pour un homme (0.4835 versus 0.4094), et à l'inverse de la probabilité pour une femme de voter républicain que l'on estime plus petite que celle pour un homme (0.3899 versus 0.4739), les deux vecteurs de probabilités conditionnelles au sexe présentent des différences. De façon exploratoire, on voit donc une certaine association entre le sexe des Américains et leurs intentions de vote. Il sera intéressant de tester la significativité de cette relation.

2.1.4 Graphiques

Voyons maintenant comment visualiser un tableau de fréquences à deux variables et utiliser des graphiques pour comparer les fréquences relatives conditionnelles afin de juger de façon exploratoire de l'association entre deux variables catégoriques. Un tableau de fréquences bivariées peut être représenté par un diagramme en bâtons groupés ou empilés, ou encore par un diagramme en mosaïque. Ces graphiques présentent tous des fréquences conditionnelles. Dans un diagramme en mosaïque, ces fréquences sont toujours relatives. Dans un diagramme en bâtons, elles peuvent être relatives ou non. Aussi, les bâtons peuvent être positionnés à la verticale ou à l'horizontale. Dans les exemples qui suivent, les bâtons sont toujours présentés à la verticale. Ces exemples utilisent les données suivantes :

Exemple : Population de la ville de Québec selon l'âge et le sexe (Statistique Canada, recensement 2006).

Âge \ Sexe	Homme	Femme	Total
0 à 14 ans	33 970	32 955	66 925
15 à 64 ans	169 065	175 810	344 875
65 ans et plus	30 990	48 345	79 335
Total	234 025	257 115	491 135

Diagramme en bâtons groupés

Dans un diagramme en bâtons groupés, des bâtons dont les longueurs sont proportionnelles aux fréquences conditionnelles d'une variable étant donnée l'autre variable sont mis côte à côte. On y trouve un groupe de bâtons pour chaque modalité de la variable de conditionnement. La figure 2.1 présente deux diagrammes en bâtons groupés possibles de créer à partir des données ci-dessus. Le graphique de gauche est utile pour comparer les groupes d'âge pour chaque sexe et le graphique de droite permet de comparer les sexes à l'intérieur des groupes d'âge.

FIGURE 2.1 – Exemples de diagrammes en bâtons groupés.

Si l'on souhaite juger de l'association potentielle entre les variables, on

changera l'échelle des fréquences conditionnelles pour l'échelle des fréquences relatives conditionnelles (voir graphique 2.2). Dans le graphique conditionnant par rapport à l'âge (graphique de droite), il devient évident que plus la population est âgée, plus la proportion de femmes est grande.

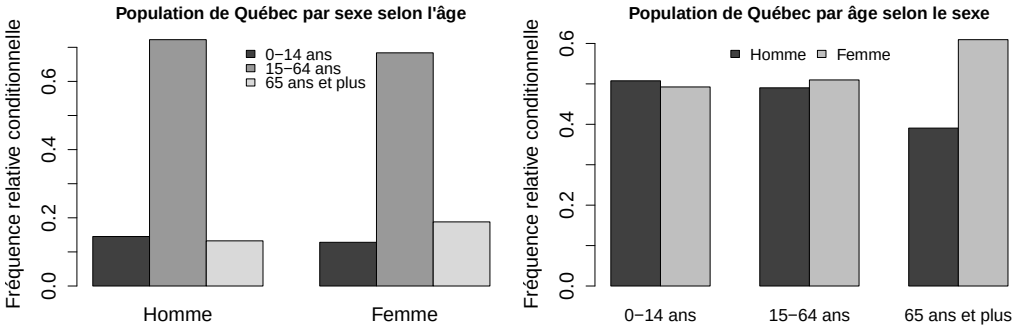


FIGURE 2.2 – Exemples de diagrammes en bâtons groupés sur l'échelle des fréquences relatives conditionnelles.

Si les deux variables n'étaient pas associées, les groupes de bâtons du graphique sur l'échelle relative seraient similaires entre eux. Donc les premiers bâtons de tous les groupes seraient environ de même longueur, même chose pour les deuxièmes bâtons, et ainsi de suite jusqu'aux bâtons en dernière position.

Diagramme en bâtons empilés

Dans un diagramme en bâtons empilés, des bâtons dont les longueurs sont relatives aux fréquences conditionnelles sont empilés les uns par-dessus les autres plutôt que d'être mis côte à côte. La figure 2.3 présente les diagrammes en bâtons empilés équivalents aux diagrammes de la figure 2.1.

Le graphique de gauche partitionne, pour chaque sexe, les fréquences selon l'âge. Un bâton contient trois sous-bâtons représentant les fréquences de l'âge étant donné le sexe. Dans le graphique de droite, un bâton contient plutôt 2 sous-bâtons représentant les fréquences du sexe étant donné l'âge. Il est plus facile de voir dans le graphique de gauche que dans le graphique de

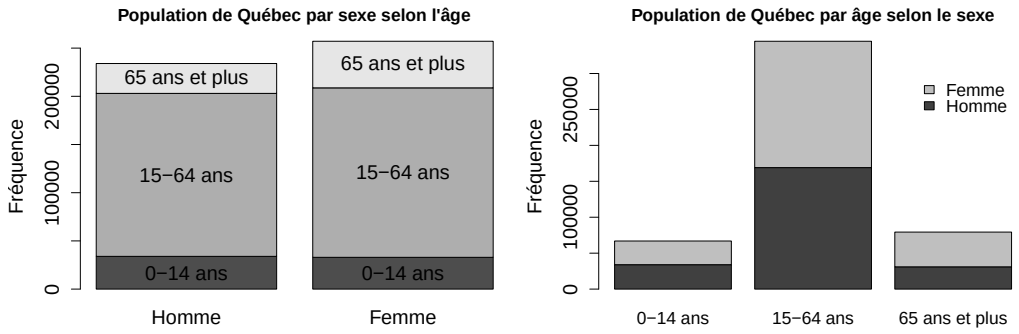


FIGURE 2.3 – Exemples de diagrammes en bâtons empilés.

droite le fait que plus la population est âgée, plus la proportion de femmes est grande.

En plus des fréquences conditionnelles, un diagramme en bâtons empilés sur l'échelle des fréquences illustre la fréquence marginale d'une des deux variables. En effet, les bâtons empilés forment de grands bâtons dont les longueurs sont proportionnelles aux fréquences marginales de la variable par rapport à laquelle on conditionne.

La figure 2.4 reprend les diagrammes de la figure 2.3 en changeant l'échelle des fréquences conditionnelles pour l'échelle des fréquences relatives conditionnelles. Dans ce type de graphique, la longueur des grands bâtons est toujours de 1. On perd donc l'information sur les fréquences marginales de la variable de conditionnement. Par contre, il devient plus facile de juger de l'association possible entre les variables. En effet, pour deux variables non reliées les grands bâtons devraient être tous subdivisés environ aux mêmes endroits. C'est ce qu'on obtient quand les fréquences relatives conditionnelles d'une variable ne varient pas beaucoup selon la valeur de la variable de conditionnement. Au contraire, si les subdivisions ne sont pas aux mêmes endroits dans les grands bâtons, c'est-à-dire si les sous-bâtons associés à une modalité en particulier ne sont pas de mêmes longueurs dans tous les grands bâtons, c'est un signe que les variables sont liées.

Dans l'exemple de la population de la ville de Québec selon l'âge et le

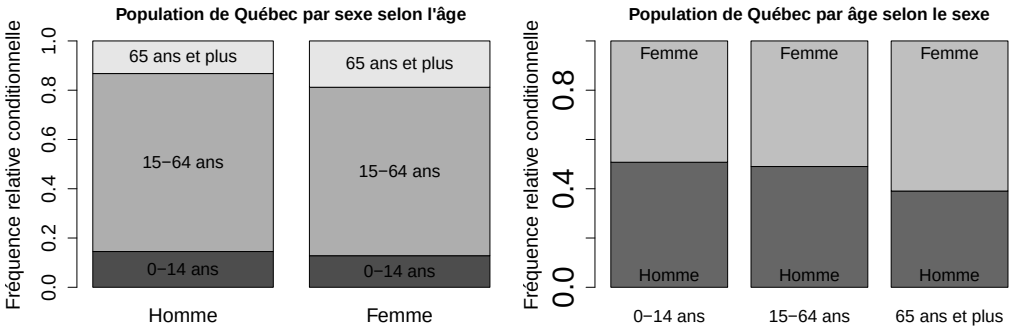


FIGURE 2.4 – Exemples de diagrammes en bâtons empilés sur l'échelle des fréquences relatives.

sexe, on observe encore que plus la population est âgée, plus la proportion de femmes est grande. Il semble donc y avoir une association entre les variables âge et sexe.

Diagramme en mosaïque

Le diagramme en mosaïque (Friendly, 1994) est une extension du diagramme en bâtons empilés sur l'échelle des fréquences relatives conditionnelles. Plutôt que d'utiliser des bâtons de même largeur, la largeur des bâtons est maintenant proportionnelle aux fréquences marginales relatives de la variable par rapport à laquelle on conditionne. La figure 2.5 présente les diagrammes en mosaïque équivalents aux diagrammes en bâtons de la figure 2.4. Typiquement, des espaces sont laissés entre les rectangles, comme dans la figure 2.5, pour rendre le graphique plus lisible.

En conditionnant par rapport au sexe, on ne voit pas beaucoup de différence entre le diagramme en bâton et le diagramme en mosaïque, car les nombres totaux d'hommes et de femmes dans l'échantillon sont pratiquement égaux. Cependant, en conditionnant par rapport à l'âge, on voit clairement l'information ajoutée dans le diagramme en mosaïque concernant la distribution marginale de l'âge.

Les boîtes du diagramme en mosaïque n'ont pas de couleur pour l'instant.

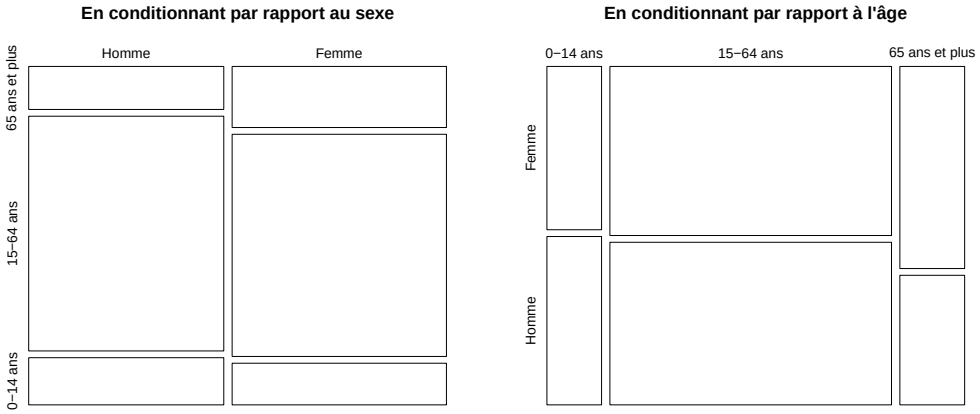


FIGURE 2.5 – Exemples de diagrammes en mosaïque.

On ajoutera plus tard de la couleur ou un motif dans ces boîtes, mais pas de la même façon que dans le diagramme en bâtons. Les couleurs ne vont pas identifier des modalités. Elles vont plutôt aider à identifier les combinaisons des catégories des variables responsables de l'association entre les variables, si une telle association est présente.

Comme dans un diagramme en bâtons, on doit faire des choix concernant les aspects suivants :

- le positionnement des variables (laquelle est en lignes et laquelle est en colonnes) ;
- la variable de conditionnement ;
- l'ordre des modalités des variables.

Si la variable de conditionnement se retrouve en colonnes (comme dans la figure 2.5), les bâtons empilés se retrouveront en colonnes, comme dans un diagramme en bâtons verticaux. À l'inverse, une variable de conditionnement en lignes donnera un diagramme en mosaïque qui est l'extension d'un diagramme en bâtons horizontaux.

Dans un diagramme en mosaïque, comme dans un diagramme en bâtons empilés sur l'échelle des fréquences relatives conditionnelles, des variables non associées présenteraient des coupures entre les sous-blocs bien alignées.

Les sous-blocs seraient donc de longueurs très similaires entre les modalités de la variable selon laquelle on conditionne, comme dans la figure 2.6.

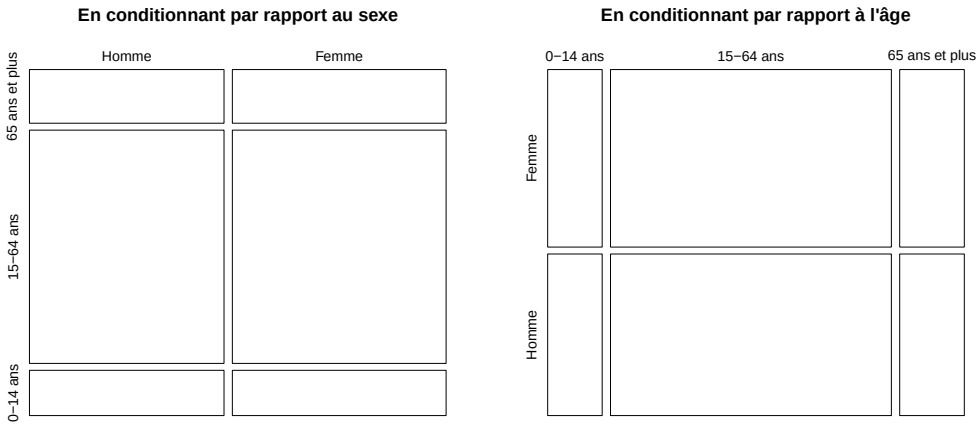


FIGURE 2.6 – Exemples de diagrammes en mosaïque pour deux variables non reliées.

Les diagrammes en mosaïque permettent de représenter facilement jusqu'à quatre variables. On les reverra donc dans le chapitre sur les tableaux de fréquences à trois variables.

2.2 Tests d'association entre deux variables nominales

La question principale d'intérêt lorsque l'on étudie conjointement 2 variables nominales X et Y est de savoir si X et Y sont *associées*.

2.2.1 Test d'indépendance et test d'homogénéité de sous-populations

Selon le mode d'échantillonnage, cette association entre X et Y est représentée par différentes hypothèses.

Échantillonnage multinomial simple :
: On suppose que les totaux des lignes et colonnes du tableau de fréquences croisant X et Y ne sont pas fixés d'avance. Seul le total n est fixé ici. On veut tester si X et Y sont indépendantes. Ce type d'hypothèses requiert un *test d'indépendance*.

Échantillonnage multinomial multiple :
: Supposons que l'échantillonnage multiple a été effectué en stratifiant selon la valeur de la variable X. On considère donc que chaque valeur de X correspond à une sous-population, c'est-à-dire que les totaux des rangées (les $n_{i\bullet}$) sont fixes. On veut tester si la distribution des valeurs de Y est la même (*homogène*) dans chacune des I sous-populations de X. Ce type d'hypothèses requiert un *test d'homogénéité*.

Définition des concepts d'indépendance et d'homogénéité

En termes statistiques, dire que les variables aléatoires discrètes X et Y sont indépendantes signifie que :

$$P(X = m_i^X, Y = m_j^Y) = P(X = m_i^X)P(Y = m_j^Y) \qquad \forall\ i, j$$

Ainsi, la fonction de masse conjointe de X et Y est le produit des fonctions de masse marginales. En utilisant les notations introduites précédemment pour représenter les probabilités d'intérêt, cette définition correspond à :

$$\pi_{ij} = \pi_{i.}\pi_{.j} \quad \text{pour tous } i = 1, \ldots, I \text{ et } j = 1, \ldots, J.$$

L'homogénéité des I sous-populations de X signifie que les I fonctions de masse conditionnelles de Y sachant X sont toutes égales. Ainsi, en se rappelant que $P(Y = m_j^Y | X = m_i^X)$ est noté $\pi_{j|i}$, l'homogénéité des populations signifie que :

$$(\pi_{1|i=1}, \pi_{2|i=1}, \ldots, \pi_{J|i=1}) = \ldots = (\pi_{1|i=I}, \pi_{2|i=I}, \ldots, \pi_{J|i=I}).$$

Cette définition correspond à :

$$\pi_{j|i} = \pi_{j|i'} \quad \text{pour toute paire } (i, i') = 1, \ldots, I \text{ et pour tout } j = 1, \ldots, J.$$

Lien entre les concepts d'indépendance et d'homogénéité

Les deux concepts que l'on vient de définir sont en fait équivalents ! En effet, on a que :

indépendance $\Rightarrow$ homogénéité :

Par définition des probabilités conditionnelles, on a que :

$$P(X = m_i^X, Y = m_j^Y) = P(Y = m_j^Y | X = m_i^X) P(X = m_i^X).$$

Si X et Y sont indépendantes, on a donc :

$$\begin{aligned} P(Y = m_j^Y | X = m_i^X) P(X = m_i^X) &= P(X = m_i^X) P(Y = m_j^Y) \\ P(Y = m_j^Y | X = m_i^X) &= P(Y = m_j^Y) \end{aligned}$$

pour tous $i = 1, \ldots, I$ et $j = 1, \ldots, J$. Si les probabilités conditionnelles sont toutes égales aux probabilités marginales, alors elles sont toutes égales entre elles.

homogénéité $\Rightarrow$ indépendance :

Par la loi des probabilités totales, on a que la probabilité marginale que Y prenne une certaine valeur m_j^Y est égale à :

$$P(Y = m_j^Y) = \sum_{i=1}^{I} P(Y = m_j^Y, X = m_i^X),$$

car $\{m_1^X, m_2^X, \ldots, m_I^X\}$ forment l'ensemble de toutes les valeurs possibles de X. Par la définition des probabilités conditionnelles, on a que :

$$P(Y = m_j^Y) = \sum_{i=1}^{I} P(Y = m_j^Y | X = m_i^X) P(X = m_i^X).$$

Sous l'hypothèse que les I sous-populations formées par X sont homogènes, tous les $P(Y = m_j^Y | X = m_i^X)$ sont égaux. Notons $P(Y = m_j^Y | X \text{quelconque})$ la valeur commune de ces probabilités conditionnelles. On a donc maintenant :

$$\begin{aligned} P(Y = m_j^Y) &= \sum_{i=1}^{I} P(Y = m_j^Y | X \text{quelconque}) P(X = m_i^X) \\ &= P(Y = m_j^Y | X \text{quelconque}) \sum_{i=1}^{I} P(X = m_i^X) \\ &= P(Y = m_j^Y | X \text{quelconque}) \qquad \text{car } \sum_{i=1}^{I} P(X = m_i^X) = 1. \end{aligned}$$

Ainsi, si les vecteurs de probabilités conditionnelles sont tous égaux entre eux, ils sont forcément égaux au vecteur de probabilités marginales de Y. En conséquence, $P(Y = m_j^Y | X = m_i^X) = P(Y = m_j^Y)$ pour tous $i = 1, \ldots, I$ et $j = 1, \ldots, J$, ce qui signifie que X et Y sont indépendants.

Mathématiquement, il s'agit donc du même concept. On utilisera les mêmes statistiques de test pour effectuer ces deux types de test, cependant :

- on suppose que les données proviennent de deux modes d'échantillonnage différents ;
- les hypothèses des tests ne sont pas formulées de la même façon ;
- les conclusions et interprétations ne sont pas non plus formulées de la même façon.

Formulation des hypothèses des tests

Voyons comment formuler les hypothèses d'un test d'indépendance et d'un test d'homogénéité de sous-populations. Ces tests sont ici tous deux bilatéraux.

Test d'indépendance :

$$\begin{aligned} H_0 &: X \text{ et } Y \text{ sont indépendants} \quad \text{ou} \\ &\quad \pi_{ij} = \pi_{i\bullet}\pi_{\bullet j} \quad \forall\ i,\ j \\ H_1 &: X \text{ et } Y \text{ ne sont pas indépendants} \quad \text{ou} \\ &\quad \pi_{ij} \neq \pi_{i\bullet}\pi_{\bullet j} \quad \text{pour au moins un couple } (i, j) \end{aligned}$$

Test d'homogénéité de sous-populations :

$$
\begin{aligned}
H_0 \;:\; & \text{Dans les } I \text{ sous-populations déterminées par } X, \\
& Y \text{ suit la même distribution} \quad \text{ou} \\
& \pi_{j|i} = \pi_{j|i'} \quad \forall\, i \neq i',\ j \quad \text{ou} \\
& \pi_{j|i} = \pi_{\bullet j} \quad \forall\, i,\ j \\
H_1 \;:\; & Y \text{ ne suit pas la même distribution} \\
& \text{dans les } I \text{ sous-populations déterminées par } X \quad \text{ou} \\
& \pi_{j|i} \neq \pi_{j|i'} \quad \text{pour au moins un couple } (i, i') \quad \text{ou} \\
& \pi_{j|i} \neq \pi_{\bullet j} \quad \text{pour au moins un couple } (i, j)
\end{aligned}
$$

Construction des statistiques de test

Pour effectuer ces tests, nous allons utiliser les statistiques X^2 et G^2 définies de façon générale à la section 1.5.1. Ici, on a $I \times J$ classes, les fréquences observées sont notées n_{ij} et les fréquences espérées sous l'hypothèse nulle H_0 doivent être estimées et sont notées $\hat{\mu}_{ij}$. Les statistiques sont donc maintenant notées comme suit :

Statistique du khi-deux de Pearson :

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{n_{ij}^2}{\hat{\mu}_{ij}} - n$$

Statistique du rapport de vraisemblance :

$$G^2 = 2 \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \ln \frac{n_{ij}}{\hat{\mu}_{ij}}$$

Ces statistiques suivent, lorsque les fréquences du tableau sont grandes, une loi du khi-deux. Rappelons que le nombre de degrés de liberté de cette khi-deux, notés d, sont définis de façon générale par :

$$
\begin{aligned}
d \;=\; & \text{dimension de l'espace des paramètres} - \\
& \text{dimension de l'espace des paramètres sous } H_0,
\end{aligned}
$$

Voici comment déduire la valeur de d selon l'hypothèse nulle formulée :

Test d'indépendance :

Ici, on est dans le cas d'un échantillonnage multinomial simple, donc :

$$(n_{ij}, i = 1, \ldots, I; j = 1, \ldots, J) \sim$$
$$Multinomiale(n, \pi_{ij}, i = 1, \ldots, I; j = 1, \ldots, J).$$

Les paramètres sont les probabilités π_{ij} pour $i = 1, \ldots, I$ et $j = 1, \ldots, J$. La dimension de l'espace des paramètres est égale au nombre total de probabilités π_{ij} moins 1 pour la contrainte que ces probabilités somment à 1. Sous H_0, on suppose que les probabilités marginales suffisent à déterminer les valeurs de toutes les probabilités. Les seuls paramètres libres sous H_0 sont donc les $\pi_{i\bullet}$ pour $i = 1, \ldots, I$ et les $\pi_{\bullet j}$ pour $j = 1, \ldots, J$. Ces paramètres sont au nombre de $I+J$. Cependant, elles doivent respecter les contraintes $\sum_{i=1}^{I} \pi_{i\bullet} = 1$ et $\sum_{j=1}^{J} \pi_{\bullet j} = 1$. On se retrouve donc avec $I + J - 2$ paramètres libres sous H_0. Ainsi,

$$\begin{aligned} d &= (IJ - 1) - (I + J - 2) \\ &= IJ - I - J + 1 \\ &= (I - 1)(J - 1). \end{aligned}$$

Test d'homogénéité de sous-populations :

Ici, à cause de l'échantillonnage multinomial multiple, les vecteurs $(n_{i1}, n_{i2}, \ldots, n_{iJ})$ pour $i = 1, \ldots, I$ sont considérés indépendants et suivent une distribution $Multinomiale(n_i, \pi_{1|i}, \pi_{2|i}, \ldots, \pi_{J|i})$. Les paramètres sont les probabilités conditionnelles $\pi_{j|i}$ pour $i = 1, \ldots, I$ et $j = 1, \ldots, J$. La dimension de l'espace des paramètres est égale au nombre total de probabilités $\pi_{j|i}$ moins I parce que chacun de vecteurs $(\pi_{1|i}, \pi_{2|i}, \ldots, \pi_{J|i})$ est soumis à la contrainte que la somme de ses éléments vaille 1. Sous H_0, tous les vecteurs $(\pi_{1|i}, \pi_{2|i}, \ldots, \pi_{J|i})$ sont égaux aux probabilités marginales $(\pi_{\bullet 1}, \pi_{\bullet 2}, \ldots, \pi_{\bullet J})$, aussi soumises à la contrainte de sommer à 1. La dimension de l'espace des paramètres sous H_0 est donc $J - 1$. Ainsi,

$$\begin{aligned} d &= (IJ - I) - (J - 1) \\ &= I(J - 1) - (J - 1) \\ &= (I - 1)(J - 1). \end{aligned}$$

On voit que les deux raisonnements arrivent au même résultat. Les statistiques X^2 et G^2 suivent donc asymptotiquement une distribution $\chi^2_{(I-1)(J-1)}$.

Ainsi, au seuil α, H_0 est rejetée si les valeurs observées de ces statistiques sont grandes, c'est-à-dire si $X^2_{obs} > \chi^2_{\alpha;(I-1)(J-1)}$ ou $G^2_{obs} > \chi^2_{\alpha;(I-1)(J-1)}$.

Voyons maintenant comment estimer les fréquences espérées sous l'hypothèse nulle formulée.

Test d'indépendance :

$$\begin{aligned} \mu_{ij} &= n\pi_{ij} \qquad \text{où } n \text{ est fixe} \\ &= n\pi_{i\bullet}\pi_{\bullet j} \qquad \text{par indépendance sous } H_0 \end{aligned}$$

On estime donc les fréquences espérées par :

$$\hat{\mu}_{ij} = n\hat{\pi}_{i\bullet}\hat{\pi}_{\bullet j} = n\left(\frac{n_{i\bullet}}{n}\right)\left(\frac{n_{\bullet j}}{n}\right) = \frac{n_{i\bullet}n_{\bullet j}}{n}.$$

Test d'homogénéité de sous-populations :

$$\begin{aligned} \mu_{ij} &= n_i\pi_{j|i} \qquad \text{où } n_i \text{ est fixe à cause de l'échantillonnage multiple} \\ &= n_i\pi_{\bullet j} \qquad \text{par homogénéité des sous-populations sous } H_0 \end{aligned}$$

On estime donc les fréquences espérées par :

$$\hat{\mu}_{ij} = n_i\hat{\pi}_{\bullet j} = n_i\left(\frac{n_{\bullet j}}{n}\right) = \frac{n_i n_{\bullet j}}{n}.$$

La seule différence entre les fréquences espérées estimées des deux tests est la façon de percevoir les fréquences marginales de la variable X. Dans le test d'indépendance, elles sont considérées aléatoires à cause de l'échantillonnage multinomial simple et on les note $n_{i\bullet}$. Dans le test d'homogénéité, elles sont considérées fixes à cause de l'échantillonnage multinomial multiple et on les note n_i. La conséquence de cette différence est que les lois exactes des statistiques X^2 et G^2 ne sont pas tout à fait les mêmes selon le test (Conover, 1999, section 4.2). De toute façon, l'important est que la distribution asymptotique des ces statistiques soit la même pour les deux tests. C'est cette distribution que l'on utilise en pratique le plus souvent pour effectuer l'un ou l'autre de ces tests.

On notera les statistiques de test, peu importe que l'on mène un test d'indépendance ou d'homogénéité des populations, par :

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij} - n_{i\bullet}n_{\bullet j}/n)^2}{n_{i\bullet}n_{\bullet j}/n} \qquad G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\ln\frac{n_{ij}}{n_{i\bullet}n_{\bullet j}/n}.$$

Remarques

1. **Mise en garde contre de faibles fréquences**

   Les seuils observés des tests sont basés sur une approximation de la distribution des statistiques X^2 et G^2. Lorsque la taille de l'échantillon n est petite, l'approximation est meilleure pour le test basé sur X^2 que pour celui basé sur G^2 (Agresti, 2007, section 2.4.7). Pour que l'approximation dans ces tests soit raisonnable, il faut encore qu ?au moins 80% des fréquences espérées soient supérieures ou égales à 5.

   Lorsque les fréquences des cellules du tableau croisé sont faibles, il est possible de calculer des seuils observés exacts, en utilisant des algorithmes numériques (Agresti, 1992). On peut aussi effectuer un test exact de Fisher, qui sera présenté dans ce chapitre pour le cas d'un tableau 2×2.

2. **Nature nominale de ces tests**

   Les valeurs des statistiques des tests du khi-deux de Pearson et du rapport de vraisemblance restent inchangées si on permute les lignes du tableau, ou si on permute les colonnes. On suppose donc ici qu'*il n'y a pas d'ordre* dans les modalités des variables X et Y. Si X ou Y sont ordinales, il est plus judicieux d'utiliser d'autres techniques traitées plus loin dans ce chapitre.

3. **Comment choisir en pratique entre un test d'indépendance et un test d'homogénéité**

   On a présenté deux types de test d'association : le test d'indépendance et le test d'homogénéité. Les distinctions entre les deux tests sont motivées par les différences entre deux types d'échantillonnage pour recueillir les données : multinomial simple versus multinomial multiple. En effet, ces deux types d'échantillonnage ont permis de formuler des hypothèses intéressantes à tester et de faire les mathématiques pour dériver des statistiques pour ces tests. On a cependant constaté que le test d'indépendance (motivé par l'échantillonnage multinomial simple) et le test d'homogénéité (motivé par l'échantillonnage multinomial multiple) sont, à un détail près, mathématiquement équivalents. En conséquence, on peut effectuer n'importe lequel des deux types de test sur des données, peu importe la façon dont elles ont été recueillies. L'important est

la cohérence entre la formulation des hypothèses, les interprétations et les conclusions.

Il est intéressant de formuler le test d'association en terme d'homogénéité de populations lorsque l'on soupçonne un lien de causalité entre les variables. Dans ce cas, on tend intuitivement à étiqueter l'une des variables comme étant explicative (celle qui a une influence sur l'autre) et l'autre comme étant une variable réponse. Il est alors naturel de se demander si la distribution de la variable réponse Y varie en fonction de la valeur prise par la variable explicative X, c'est-à-dire de faire un test d'homogénéité des populations formées par la variable X. Ainsi, dans un test d'homogénéité, on sous-entend que c'est la variable X qui influence Y et non l'inverse. Dans un test d'indépendance, on ne suppose aucune direction dans le lien entre X et Y.

Exemple de test d'association : intentions de vote selon le sexe

Dans l'exemple des intentions de vote selon le sexe, l'échantillonnage était multinomial simple. Pourtant, pour tester l'association entre X, le sexe, et Y, l'intention de vote, il est plus naturel de faire un test d'homogénéité de sous-populations qu'un test d'indépendance. Un test d'indépendance répondrait à une question de recherche formulée ainsi :

Question de recherche possible : Aux États-Unis, y a-t-il un lien entre le sexe d'une personne et le parti politique pour lequel elle a l'intention de voter ?

Cependant, il m'apparaît plus simple de comprendre une question de recherche formulée ainsi :

Question de recherche choisie : Aux États-Unis, les intentions de vote diffèrent-elles entre les hommes et les femmes ?

Cette question de recherche mène à la formulation d'une hypothèse d'homogénéité des sous-populations :

H_0 : Les intentions de vote sont les mêmes pour les hommes et les femmes
$(\pi_{1|1}, \pi_{2|1}, \pi_{3|1}) = (\pi_{1|2}, \pi_{2|2}, \pi_{3|2})$

H_1 : Les intentions de vote diffèrent entre les hommes et les femmes
$(\pi_{1|1}, \pi_{2|1}, \pi_{3|1}) \neq (\pi_{1|2}, \pi_{2|2}, \pi_{3|2})$

Pour calculer les statistiques X^2 et G^2, nous devons estimer les fréquences espérées sous l'hypothèse d'indépendance comme suit :

$$\hat{\mu}_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}, \qquad \text{par exemple : } \hat{\mu}_{11} = (444 \times 577)/980 = 261.4$$

Les valeurs suivantes entrent dans nos calculs :

	Démocrate	Indépendant	Républicain	Total
Femme	$n_{11} = 279$ $\hat{\mu}_{11} = 261.4$	$n_{12} = 73$ $\hat{\mu}_{12} = 70.7$	$n_{13} = 225$ $\hat{\mu}_{13} = 244.9$	$n_{1\bullet} = 577$
Homme	$n_{21} = 165$ $\hat{\mu}_{21} = 182.6$	$n_{22} = 47$ $\hat{\mu}_{22} = 49.3$	$n_{23} = 191$ $\hat{\mu}_{23} = 171.1$	$n_{2\bullet} = 403$
Total	$n_{\bullet 1} = 444$	$n_{\bullet 2} = 120$	$n_{\bullet 3} = 446$	$n = 980$

Résultats :
$X^2 = 7.01$ et $G^2 = 7.00$
Sous H_0, ces statistiques suivent une χ^2 à $(I-1)(J-1) = (2-1)(3-1) = 2$ degrés de liberté. Ainsi, le seuil observé est $P(\chi^2_2 \geq 7.00) = 0.03 \leq 0.05$

Conclusion : Nous rejetons H_0 au seuil 5%. Il y a des différences entre les intentions de vote des hommes et celles des femmes. Nous allons voir plus loin comment décrire cette relation.

Ainsi, l'important n'est pas de formuler une hypothèse en accord avec le type d'échantillonnage utilisé pour recueillir les données. On cherche plutôt à être le plus clair possible. Dans l'exemple ici, imaginons que les résultats du test doivent être publiés dans un article de journal grand public. Il devient important de formuler des hypothèses faciles à comprendre, peu importe qu'il

s'agisse d'un test d'indépendance ou d'homogénéité. Pourvu que la question de recherche, les hypothèses et la conclusion concordent, la procédure est correcte d'un point de vue statistique.

2.2.2 Cas particulier des tableaux 2 × 2 : test de comparaison de deux proportions

Un tableau de fréquences 2×2 est un tableau de fréquences pour lequel les variables X et Y n'ont que 2 modalités possibles (variables dichotomiques). C'est donc un cas particulier des tableaux $I \times J$ lorsque $I = 2$ et $J = 2$. Un tableau 2×2 a la forme suivante :

$X \setminus Y$	m_1^Y	m_2^Y	Total
m_1^X	n_{11}	n_{12}	$n_{1\bullet}$
m_2^X	n_{21}	n_{22}	$n_{2\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	n

Exemples de tableaux 2 × 2 :

L'étude expérimentale concernant l'aspirine et l'infarctus du myocarde introduite à la section 2.1.1 a généré un tableau 2×2. Ces données seront utilisées dans le texte qui suit pour illustrer les concepts théoriques présentés.

Les résultats de l'étude cas-témoin concernant la cigarette et l'infarctus du myocarde, aussi introduite à la section 2.1.1, se présentent aussi dans un tableau 2×2. Cet exemple est particulier à cause du conditionnement par rapport à la variable réponse Y dans l'échantillonnage. Il sera traité à la section 2.3.7.

Simplification de la formule de la statistique du khi-deux de Pearson

Pour les tableaux 2×2, la statistique de Pearson peut s'écrire sous la forme suivante :

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n})^2}{\frac{n_{i\bullet} n_{\bullet j}}{n}} = \frac{n[n_{11}n_{22} - n_{12}n_{21}]^2}{n_{1\bullet} n_{2\bullet} n_{\bullet 1} n_{\bullet 2}}.$$

Posons $\Delta = \begin{vmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{vmatrix} = n_{11}n_{22} - n_{12}n_{21}$,

nous avons donc :

$$X^2 = \frac{n\Delta^2}{n_{1\bullet}n_{2\bullet}n_{\bullet 1}n_{\bullet 2}}.$$

Ici, les degrés de liberté de la distribution asymptotique de X^2 valent 1, car $(I-1)\times(J-1) = 1\times 1 = 1$.

Exemple de test d'association pour un tableau 2 × 2 :
étude expérimentale concernant l'aspirine et l'infarctus du myocarde.

Dans cette étude, la grande question de recherche était la suivante :

Question : Est-ce que l'aspirine réduit les risques d'infarctus ?

Avec un test d'homogénéité, en accord avec l'échantillonnage multinomial multiple conditionnel à la prise quotidienne d'aspirine (variable X), on pourrait répondre à une version bilatérale de cette question de recherche :

Est-ce que le risque d'infarctus diffère entre ceux qui ont pris quotidiennement de l'aspirine et ceux qui n'en ont pas pris ?

Les hypothèses du test sont :

$$\begin{aligned} H_0 &: (\pi_{1|i=1}, \pi_{2|i=1}) = (\pi_{1|i=2}, \pi_{2|i=2}) \\ H_1 &: (\pi_{1|i=1}, \pi_{2|i=1}) \neq (\pi_{1|i=2}, \pi_{2|i=2}) \end{aligned}$$

Rappelons que les données sont les suivantes :

X	Y : Infarctus		
Groupe	Oui	Non	Total
Placebo	239	10795	11034
Aspirine	139	10898	11037
Total	378	21693	22071

Utilisons la statistique X^2 de Pearson pour effectuer ce test. On peut calculer sa valeur observée avec la formule originale, ou avec la formule simplifiée pour

le cas 2×2 comme ceci :

$$X^2_{obs} = \frac{22071(239 \times 10898 - 10795 \times 139)^2}{11034 \times 11037 \times 378 \times 21693} = 26.9437.$$

Cette valeur est beaucoup plus grande que la valeur critique du test : $\chi^2_{0.05,1} = 3.84$. On rejette donc de façon convaincante l'hypothèse nulle. Il y a une grande différence de risque d'infarctus entre ceux qui prennent quotidiennement de l'aspirine et ceux qui n'en prennent pas. En fait, on estime $\pi_{1|i=1} = P(\text{infarctus} \mid \text{placebo})$ par la valeur $n_{11}/n_{1\bullet} = 239/11034 = 0.02166032$, et on estime $\pi_{1|i=2} = P(\text{infarctus} \mid \text{aspirine})$ par la valeur $n_{21}/n_{2\bullet} = 139/10898 = 0.01259400$. Ces estimations sont bonnes, car les sous-populations pour l'échantillonnage multiple sont formées à partir de la variable X. Elles ne seraient pas adéquates si les sous-populations avaient été formées à partir de la variable Y, comme dans une étude cas-témoin. Étant donné que $\pi_{1|i=1} > \pi_{1|i=2}$, on peut conclure que le risque d'infarctus est plus faible pour ceux qui prennent quotidiennement de l'aspirine.

Test de comparaison de deux proportions

Pour un tableau 2×2, l'hypothèse nulle du test d'homogénéité des deux sous-populations formées par la variable X s'écrit :

$$H_0 : (\pi_{1|i=1}, \pi_{2|i=1}) = (\pi_{1|i=2}, \pi_{2|i=2}).$$

Puisque $\pi_{1|i} + \pi_{2|i} = 1$ pour $i = 1, 2$, cette hypothèse est équivalente à :

$$H_0 : \pi_{1|i=1} = \pi_{1|i=2}.$$

Il s'agit donc en fait d'un simple test de comparaison de deux proportions, aussi appelé test d'égalité de deux proportions. On est bien dans le contexte d'un tel test. En effet, on a deux sous-populations postulées indépendantes. Le modèle statistique de l'échantillonnage multinomial multiple revient à dire qu'on a deux variables aléatoires binomiales indépendantes. En effet, définissons un succès comme étant $Y = m_1^Y$. On a donc n_{11} le nombre de succès dans la première population (soit la sous-population définie par $X = m_1^X$) et n_{21} le nombre de succès dans la deuxième population (soit la sous-

population définie par $X = m_2^X$) tels que :

$$\begin{aligned} n_{11} &\sim Bin(n_1 = n_{1\bullet}, \pi_1 = \pi_{1|i=1}), \\ n_{21} &\sim Bin(n_2 = n_{2\bullet}, \pi_2 = \pi_{1|i=2}). \end{aligned}$$

Ainsi, on estime les paramètres inconnus de ces distributions par $\hat{\pi}_1 = n_{11}/n_1$ et $\hat{\pi}_2 = n_{21}/n_2$.

Dans un test de comparaison de proportions, étant donné que l'on compare seulement deux paramètres, il est possible de faire un test unilatéral. Les hypothèses du test sont formulées ainsi :

$$H_0 : \pi_1 = \pi_2 \quad \text{versus} \quad H_1 : \begin{array}{llll} \pi_1 \neq \pi_2 & \text{ou} & & \text{test biltéral} \\ \pi_1 > \pi_2 & \text{ou} & \pi_1 < \pi_2 & \text{test unilatéral.} \end{array}$$

Voyons deux versions du test de comparaison de deux proportions : un test de Wald et un test score.

Test de Wald de comparaison de deux proportions

La statistique du test de Wald de comparaison de deux proportions est la suivante (Agresti, 2002, section 3.1.3) :

$$Z_w = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\dfrac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \dfrac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}} \xrightarrow[H_0]{\text{asymp.}} \mathcal{N}(0,1),$$

où $\hat{\pi}_i = n_{i1}/n_i$ pour $i = 1, 2$.

Exemple de test de Wald de comparaison de deux proportions : étude expérimentale concernant l'aspirine et l'infarctus du myocarde.

Afin de répondre à la question « Est-ce que l'aspirine réduit les risques d'infarctus ? », nous pouvons tester l'égalité des proportions :
$\pi_1 = \pi_{1|i=1} = P(\text{infarctus} \mid \text{placebo})$ et
$\pi_2 = \pi_{1|i=2} = P(\text{infarctus} \mid \text{aspirine})$

contre une hypothèse alternative unilatérale à droite :

$$\begin{array}{lcll} H_0 & : & \pi_1 = \pi_2 & \Leftrightarrow \quad \text{l'aspirine n'a pas d'effet} \\ H_1 & : & \pi_1 > \pi_2 & \Leftrightarrow \quad \text{l'aspirine réduit le risque d'infarctus.} \end{array}$$

La statistique du test de Wald pour confronter ces hypothèses prend la valeur observée suivante :

$$z_w = \frac{0.02166 - 0.01259}{\sqrt{\dfrac{0.02166(1-0.02166)}{11304} + \dfrac{0.01259(1-0.01259)}{11307}}} = 5.193717.$$

Cette valeur est beaucoup plus grande que la valeur critique du test : $z_{0.05} = 1.645$. On peut donc encore conclure que l'aspirine réduit les risques d'infarctus.

Test score de comparaison de deux proportions

Le test de comparaison de proportions le plus souvent présenté dans un cours d'introduction à la statistique (Hines *et al.* (2012, section 11.3.5), Agresti (2002, exercice 3.30)) est le test score se basant sur la statistique suivante :

$$Z_s = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \xrightarrow[H_0]{\text{asymp.}} \mathcal{N}(0,1).$$

où $\hat{\pi}_i = n_{i1}/n_i$ pour $i = 1, 2$ et $\hat{\pi} = \frac{n_1\hat{\pi}_1 + n_2\hat{\pi}_2}{n_1+n_2} = \frac{n_{11}+n_{21}}{n_1+n_2}$ est utilisé pour calculer une variance groupée (en anglais pooled variance).

Exemple de test score de comparaison de deux proportions :
étude expérimentale concernant l'aspirine et l'infarctus du myocarde.

Une statistique similaire à la statistique de Wald que l'on vient de calculer

est la statistique score suivante :

$$z_s = \frac{0.02166 - 0.01259}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{11304} + \frac{1}{11307}\right)}} = 5.190729$$

$$\text{car } \hat{\pi} = \frac{11304 \times 0.02166 + 11307 \times 0.01259}{11304 + 11307} = \frac{378}{22071} = 0.01712655.$$

La valeur observée de la statistique du test score $z_s = 5.190729$ est très proche, mais pas tout à fait égale à celle de la statistique du test de Wald $z_w = 5.193717$. Remarquez qu'en élevant au carré z_s, on retombe exactement sur la statistique X^2_{obs} de Pearson ($z_s^2 = 5.190729^2 = 26.94367 = X^2_{obs}$).

L'équivalence que l'on vient de voir dans l'exemple n'est pas uniquement vraie pour ces données. On peut prouver théoriquement que Z_s, la statistique du test score de comparaison de deux proportions, élevée au carré, est égale à X^2, la statistique du khi-deux de Pearson d'un test d'homogénéité de populations dans un tableau de fréquences 2×2.

Calcul de tailles d'échantillons

Lors de la planification d'une expérience visant à comparer les probabilités de succès dans deux expériences binomiales, il est parfois utile de déterminer les tailles d'échantillons permettant d'atteindre une puissance prédéterminée pour une certaine hypothèse alternative. Des tailles d'échantillon $n_1 = n_2$ qui donnent au test fait à un seuil α une puissance de $1 - \beta$ pour l'hypothèse alternative $H_1 : \pi_1 = \pi_1^{H_1}$ et $\pi_2 = \pi_2^{H_1}$ sont

$$n_1 = n_2 = \frac{(z_{\alpha/2} + z_\beta)^2\{\pi_1^{H_1}(1-\pi_1^{H_1}) + \pi_2^{H_1}(1-\pi_2^{H_1})\}}{(\pi_1^{H_1} - \pi_2^{H_1})^2},$$

où $z_{\alpha/2}$ et z_β sont des quantiles d'une loi normale standard.

Notons que cette formule s'appuie sur la distribution normale qui n'est valide que pour de grands échantillons (Agresti, 2002, section 6.5.1). Une formule alternative est présentée dans Fleiss *et al.* (2003, chapitre 4), qui traite aussi du problème sans postuler l'égalité des deux tailles d'échantillons n_1 et n_2.

Exemple de calcul de tailles d'échantillons :

Supposons que $\pi_1^{H_1} = 0.2$ et $\pi_2^{H_1} = 0.3$. Pour que la puissance du test au seuil 5% soit 90%, il faut prendre

$$n_1 = n_2 = \frac{(1.96 + 1.28)^2(0.2 \times 0.8 + 0.3 \times 0.7)}{(0.2 - 0.3)^2} = 389.$$

Ainsi, si nous nous assurons de faire 389 essais dans les deux expériences binomiales à l'étude, le test d'homogénéité au seuil 5% aura une puissance de 90%.

2.2.3 Petits échantillons : test de Fisher

Tous les tests vus jusqu'à maintenant pour un tableau de fréquences à deux variables sont asymptotiques. Les distributions sous H_0 de leurs statistiques de test sont donc valides à la condition que n (ou encore les n_i ou les n_j) soit assez grand. Dans le cas de petits échantillons, voici les solutions qui s'offrent à nous.

Correction pour la continuité (de Yates) de la statistique de Pearson

Ajouter une correction pour la continuité à la formule de la statistique X^2 de Pearson améliore la validité de la loi asymptotique dans le cas de fréquences faibles. Avec une telle correction, la statistique devient :

$$X^2_{corr} = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(|n_{ij} - \hat{\mu}_{ij}| - 0.5)^2}{\hat{\mu}_{ij}}$$

où $|\cdot|$ représente la valeur absolue. Certains logiciels modifient un peu cette formule ramenant à zéro les termes $(|n_{ij} - \hat{\mu}_{ij}| - 0.5)$ négatifs.

Dans le cas particulier d'un tableau 2×2, la formule de la valeur observée de cette statistique se simplifie à :

$$X^2_{corr} = \frac{n(|\Delta| - \frac{n}{2})^2}{n_{1\bullet} n_{2\bullet} n_{\bullet 1} n_{\bullet 2}}.$$

Tout comme c'était le cas pour X^2, sous H_0 (indépendance entre les deux variables du tableau), X^2_{corr} suit asymptotiquement une $\chi^2_{(I-1)(J-1)}$.

Calcul de seuils observés exacts

Il est possible de calculer numériquement des seuils observés se basant sur les distributions exactes de plusieurs statistiques de test, notamment X^2 et G^2. La théorie derrière ces calculs est présentée dans Agresti (1992). On ne peut pas dériver de formules algébriques pour ces calculs, ils doivent être faits numériquement, donc à l'aide d'un ordinateur. La procédure `FREQ` de SAS offre ces calculs.

Test exact de Fisher pour un tableau 2 × 2

Ce test permet de tester si deux variables dichotomiques X et Y sont dépendantes. La statistique du test est simplement n_{11}, soit la fréquence théorique que $X = m_1^X$ et $Y = m_1^Y$ simultanément. En postulant que les fréquences des deux marges sont fixes, on peut trouver la loi exacte de n_{11} sous l'hypothèse d'indépendance entre X et Y. Il s'agit d'une distribution hypergéométrique.

Formulation des hypothèses du test de Fisher : Un test de Fisher peut être bilatéral ou unilatéral. L'hypothèse nulle est H_0 : X et Y sont indépendantes. L'hypothèse alternative H_1 peut être l'une des suivantes :

X et Y ne sont pas indépendantes
n_{11} est plus petit qu'attendu sous indépendance
n_{11} est plus grand qu'attendu sous indépendance

Description de la distribution hypergéométrique : La distribution hypergéométrique est discrète et dépend de 3 paramètres qui sont des entiers positifs, notés ici a, b, c. Sa fonction de masse s'écrit :

$$P(W = w) = \frac{\binom{b}{w}\binom{c-b}{a-w}}{\binom{c}{a}} \quad \text{où} \quad \max(0, a+b-c) \leq w \leq \min(a, b).$$

Si la variable aléatoire W suit une distribution hypergéométrique, nous pouvons montrer que ses deux premiers moments sont donnés par :

$$E(W) = \frac{ab}{c} \text{ et } Var(W) = \frac{ab(c-a)(c-b)}{c^2(c-1)}.$$

Nous obtenons une distribution hypergéométrique lorsque nous tirons, sans remise, des boules d'une urne. Si l'urne contient b boules blanches et $c-b$ boules noires, appelons W le nombre de boules blanches obtenues après avoir effectué a tirages. La variable aléatoire W suit alors une distribution hypergéométrique de paramètres a, b, c.

Notons que l'expérience qui permet de construire la distribution hypergéométrique est semblable à une expérience binomiale. En effet, elle se compose de a essais (tirages), qui ont chacun 2 résultats possibles : blanc (succès) ou noir (échec). Cependant, dans une expérience binomiale, les essais sont indépendants, comme s'il s'agissait de tirages avec remise. Ainsi, la probabilité de succès (tirer une boule blanche) est toujours la même. Cependant, si les tirages sont effectués sans remise, les essais ne sont pas indépendants. La probabilité d'obtenir blanc comme résultat dépend des essais précédents. Puisque les balles ne sont pas remises dans l'urne après chaque tirage, la probabilité d'obtenir une boule blanche à un tirage est plus faible si tous les tirages précédents ont donné des boules blanches.

Distribution sous H_0 de la statistique du test de Fisher : Sous H_0 : X et Y sont indépendantes, la loi conditionnelle de n_{11} étant donné que les fréquences sont fixes dans les deux marges est la distribution hypergéométrique de paramètres $a = n_{\bullet 1}, b = n_{1\bullet}, c = n$. Ainsi, on a :

$$P(n_{11} = w \mid H_0, n_{\bullet 1}, n_{1\bullet} \text{ et } n) = \frac{\binom{n_{1\bullet}}{w}\binom{n_{2\bullet}}{n_{\bullet 1} - w}}{\binom{n}{n_{\bullet 1}}}$$

pour $w \in [\max(0, n_{\bullet 1} + n_{1\bullet} - n), \min(n_{\bullet 1}, n_{1\bullet})]$.

Afin de justifier cette affirmation, faisons d'abord le parallèle entre le contexte de la distribution hypergéométrique et le contexte d'un tableau 2×2. Disons que les boules représentent les individus de la population à l'étude.

On mesure deux variables : X = la couleur des boules et Y = l'état tirée ou non tirée des boules. La variable à laquelle on s'intéresse ici (la statistique du test de Fisher) est n_{11}, soit le nombre de boules blanches parmi les boules tirées. On obtient le tableau suivant :

$X \setminus Y$	m_1^Y =tirée	m_2^Y =non tirée	Total
m_1^X = blanche	n_{11}	n_{12}	$n_{1\bullet} = b$
m_2^X = noire	n_{21}	n_{22}	$n_{2\bullet} = c - b$
Total	$n_{\bullet 1} = a$	$n_{\bullet 2} = c - a$	$n = c$

Le nombre de boules tirées (paramètre hypergéométrique a) vaut donc $n_{\bullet 1}$, le nombre de boules blanches (paramètre hypergéométrique b), vaut $n_{1\bullet}$ et le nombre total de boules (paramètre hypergéométrique c) vaut n.

La distribution de n_{11} en conditionnant sur la valeur des marges est bien hypergéométrique sous H_0. En effet,

$$\begin{aligned}
&P(n_{11} = w \mid (n_{1\bullet}, n_{2\bullet}, n_{\bullet 1}, n_{\bullet 2}) = (b, c-b, a, c-a) \\
&= \frac{P(n_{11} = w \ , \ (n_{1\bullet}, n_{2\bullet}, n_{\bullet 1}, n_{\bullet 2}) = (b, c-b, a, c-a))}{P((n_{1\bullet}, n_{2\bullet}, n_{\bullet 1}, n_{\bullet 2}) = (b, c-b, a, c-a))} \\
&= \frac{P(n_{11} = w \ , \ n_{12} = b - w \ , \ n_{21} = a - w \ , \ n_{22} = c - b - (a - w))}{P(n_{1\bullet} = b \ , \ n_{\bullet 1} = a)} \\
&= \frac{P((n_{11}, n_{12}, n_{21}, n_{11}) = (w, b-w, a-w, c-b-a+w))}{P(n_{1\bullet} = b) \ P(n_{\bullet 1} = a)}
\end{aligned}$$

La probabilité multinomiale au numérateur est égale à :

$$\frac{c!}{w!(b-w)!(a-w)!(c-b-a+w)!} \pi_{11}^w \pi_{12}^{b-w} \pi_{21}^{a-w} \pi_{22}^{c-b-a+w}$$

Sous l'hypothèse nulle d'indépendance entre X et Y, on a que $\pi_{ij} = \pi_{i\bullet}\pi_{\bullet j}$. Cette probabilité devient donc :

$$\frac{c!(\pi_{1\bullet}\pi_{\bullet 1})^w (\pi_{1\bullet}\pi_{\bullet 2})^{b-w} (\pi_{2\bullet}\pi_{\bullet 1})^{a-w} (\pi_{2\bullet}\pi_{\bullet 2})^{c-b-a+w}}{w!(b-w)!(a-w)!(c-b-a+w)!}$$

Le produit des probabilités binomiales au dénominateur est égal à :

$$\frac{c!}{b!(c-b)!} \pi_{1\bullet}^b \pi_{2\bullet}^{c-b} \times \frac{c!}{a!(c-a)!} \pi_{\bullet 1}^a \pi_{\bullet 2}^{c-a}$$

Après simplification et réarrangement des termes, on obtient que sous H_0 :

$$P(n_{11} = w \mid (n_{1\bullet}, n_{2\bullet}, n_{\bullet 1}, n_{\bullet 2}) = (b, c-b, a, c-a)) = \frac{\binom{b}{w}\binom{c-b}{a-w}}{\binom{c}{a}}.$$

Comment mener un test de Fisher : La méthode du test de Fisher est la suivante :

1. Pour chaque valeur possible de n_{11}, notée w, on calcule la fonction de masse de la statistique de test n_{11} sous H_0, considérant fixes les sommes des lignes et des colonnes :

   $$P(n_{11} = w \mid n_{11} \sim Hypergeometrique(n_{\bullet 1}, n_{1\bullet}, n))$$

   pour $w \in [\max(0, n_{\bullet 1} + n_{1\bullet} - n), \min(n_{\bullet 1}, n_{1\bullet})]$.

2. Le seuil du test est obtenu ainsi ($n_{11_{obs}}$ est la valeur observée de n_{11} dans l'échantillon) :

   - Pour H_1 : X et Y ne sont pas indépendantes :
     $\sum_{w \in A} P(n_{11} = w \mid n_{11} \sim Hypergeometrique(n_{\bullet 1}, n_{1\bullet}, n))$,
     où A est l'ensemble $\{w$ tel que $P(n_{11} = w | \ldots) \leq P(n_{11} = n_{11_{obs}} | \ldots)\}$;

   - Pour H_1 : n_{11} plus petit qu'attendu sous H_0 :
     $P(n_{11} \leq n_{11_{obs}} \mid n_{11} \sim Hypergeometrique(n_{\bullet 1}, n_{1\bullet}, n))$;

   - Pour H_1 : n_{11} plus grand qu'attendu sous H_0 :
     $P(n_{11} \geq n_{11_{obs}} \mid n_{11} \sim Hypergeometrique(n_{\bullet 1}, n_{1\bullet}, n))$.

Remarque : En pratique, il est très rare que les deux marges soient réellement fixes. On peut effectuer le test de Fisher sur des données où les marges ne sont pas fixées d'avance. Le test est tout à fait valide. Cependant, il pourrait être trop conservateur comparativement à un test asymptotique ne supposant pas les deux marges fixes. L'utilisation de « mid p-value » (voir section 1.3.3) au lieu de seuil observé ordinaire vise à corriger ce côté trop conservateur en augmentant la puissance du test. Rappelons que le « mid

p-value » est la moitié de la probabilité d'un résultat aussi probable sous H_0 que celui observé, plus la probabilité d'un résultat moins probable sous H_0, tout en respectant la direction de H_1.

Exemple 1 de test de Fisher :

Une collègue de Sir Ronald Fisher affirmait qu'en buvant un thé, elle pouvait dire lequel du thé ou du lait avait été versé en premier dans la tasse. Elle considérait que verser le lait en premier produisait du meilleur thé. Pour tester son affirmation, Fisher lui fit goûter 8 tasses de thé. Quatre thés avaient été préparés selon la façon anglaise en versant le lait en premier et les quatre autres étaient ordinaires. Elle savait au départ qu'il y avait 4 thés de chaque type.

Question : La façon anglaise de faire du thé en versant le lait en premier fait-elle réellement du meilleur thé selon le goût de cette collègue ?

Données :

	Évaluation de la collègue		
Type réel	bon	moins bon	Total
anglais	3	1	4
ordinaire	1	3	4
Total	4	4	

Test d'hypothèses :
H_0 : Il n'y a pas de différence entre un thé anglais et un thé ordinaire
H_1 : Il y effectivement une différence entre un thé anglais et un thé ordinaire $\longrightarrow$ Test bilatéral
ou bien
$H_1 : n_{11} >$ sous H_0 : Le thé anglais est meilleur que l'ordinaire
$\longrightarrow$ Test unilatéral.

La loi conditionnelle de n_{11}, sous H_0, est la suivante :

valeurs possibles de $n_{11} : w$	0	1	2	3	4
$P(n_{11} = w \mid n_{\bullet 1} = 4, n_{1\bullet} = 4 \; et \; n = 8)$	0.014	0.229	0.514	0.229	0.014
X^2	8	2	0	2	8

La dernière ligne de ce tableau donne la valeur observée que prendrait la statistique X^2 de Pearson pour chaque valeur possible de n_{11}.

Résultats :
mid p-value unilatéral $= P(n_{11} = 3)/2 + P(n_{11} = 4) = 0.129 > 5\%$
mid p-value bilatéral exact $= P(X^2 = 2)/2 + P(X^2 = 8) = 0.258 > 5\%$
Seuil observé bilatéral asymptotique $= P(\chi_1^2 \geq 2) = 0.157 > 5\%$

Conclusion : Nous ne démontrons pas que le thé anglais est meilleur que l'ordinaire.

Exemple 2 de test de Fisher :

Une étude a été réalisée sur 95 souris. Seize de ces souris ont reçu un traitement particulier. Les souris ont ensuite été observées pendant un certain laps de temps afin de savoir si elles développaient un cancer.

Question : Les souris ayant reçu un traitement sont-elles plus sujettes au cancer que des souris témoins ?

Données :

	Tumeurs		
	Présentes	Absentes	Total
Traitée	4	12	16
Témoin	5	74	79
Total	9	86	95

Test d'hypothèses :
H_0 : Il n'y a pas de différence entre les souris traitées et les souris témoin
H_1 : Il y effectivement une différence entre les souris traitées et les souris témoin $\longrightarrow$ Test bilatéral
ou bien
$H_1 : n_{11} >$ sous H_0 : Les souris traitées sont plus sujettes au cancer que les souris témoin $\longrightarrow$ Test unilatéral.

La distribution de n_{11} sous l'hypothèse d'indépendance, et étant donné que $n_{1\bullet} = 16, n_{\bullet 1} = 9$ *et* $n = 95$, est la suivante :

w	0	1	2	3	4	5	6+
$P(n_{11} = w \vert \ldots)$	0.175	0.355	0.296	0.132	0.035	0.006	0.001

Le seuil observé du test unilatéral calculé avec le « mid p-value » est donné par $0.035/2 + 0.006 + 0.001 = 0.024$.

Pour calculer le seuil observé du test bilatéral sans calculer la statistique X^2, il faut sommer les probabilités de toutes les valeurs possibles de n_{11} qui ont une probabilité inférieure ou égale à celle de la fréquence 4, soit la valeur observée n_{11}. Ces valeurs sont les probabilités des modalités 4, 5 et 6+. En accordant un poids de 1/2 aux valeurs ayant des probabilités égales à celle de 4, nous obtenons comme « mid p-value » du test bilatéral $0.035/2 + 0.006 + 0.001 = 0.024$. Ici, étonnement, les seuils observés sont les mêmes pour le test unilatéral et pour le test bilatéral.

Nous calculons $X^2 = 5.41$. Le seuil observé du test du khi-deux basé sur une approximation asymptotique est $P(\chi^2_1 \geq 5.41) = 0.02$.

Conclusion : Nous pouvons affirmer que les souris traitées sont plus sujettes au cancer que les souris témoin.

Généralisation du test de Fisher à un tableau de dimension quelconque $I \times J$: Le test de Fisher a été généralisé au cas général $I \times J$ par Freeman et Halton (1951). Le test peut alors uniquement être bilatéral. Il ne sera pas vu dans ce cours.

2.3 Décrire et mesurer l'association entre deux variables nominales

Si un test d'association mène au rejet de l'hypothèse nulle d'absence d'association, on va vouloir décrire l'association détectée entre les variables. Nos outils pour formuler cette description sont les suivants.

2.3.1 Probabilités conditionnelles

Rappelons que s'il y a indépendance entre les variables X et Y, les probabilités des modalités d'une variable, disons Y, conditionnelles à la valeur de l'autre variable, disons X, seront égales aux probabilités marginales en Y. Des probabilités conditionnelles très différentes entre elles, donc différentes des probabilités marginales, causent une dépendance.

Afin de décrire une association entre deux variables nominales, on peut donc estimer les probabilités conditionnelles à partir des fréquences relatives conditionnelles et les probabilités marginales par les fréquences relatives marginales. Attention, il faut d'abord s'assurer que le mode d'échantillonnage permet de faire ces estimations (voir section 2.1.2). Si ce n'est pas le cas, cet outil ne peut pas être utilisé. Si les estimations sont bonnes, il suffit de cibler les différences et de les présenter. Il faut donc cerner quelles fréquences relatives en conditionnant par rapport aux lignes diffèrent le plus des fréquences relatives dans la marge en bas du tableau, ou encore quelles fréquences relatives en conditionnant par rapport aux colonnes diffèrent le plus des fréquences relatives dans la marge de droite. Souvent, ce travail a déjà été fait à l'étape de la visualisation des données, avant même de tester l'association.

Exemple de calcul de probabilités conditionnelles :
intentions de vote selon le sexe

Lorsque nous avions fait une analyse exploratoire des données, nous avions obtenu les fréquences relatives conditionnelles à X et les fréquences relatives marginales de Y suivantes :

	Démocrate	Indépendant	Républicain
Femme	$\hat{\pi}_{1\|i=1} = f_{1\|i=1} = 0.4835$	$\hat{\pi}_{2\|i=1} = f_{2\|i=1} = 0.1265$	$\hat{\pi}_{3\|i=1} = f_{3\|i=1} = 0.3899$
Homme	$\hat{\pi}_{1\|i=2} = f_{1\|i=2} = 0.4094$	$\hat{\pi}_{2\|i=2} = f_{2\|i=2} = 0.1166$	$\hat{\pi}_{3\|i=2} = f_{3\|i=2} = 0.4739$
Total	$\hat{\pi}_{\bullet 1} = f_{\bullet 1} = 0.4531$	$\hat{\pi}_{\bullet 2} = f_{\bullet 2} = 0.1224$	$\hat{\pi}_{\bullet 3} = f_{\bullet 3} = 0.4245$

Sous indépendance, $\pi_{1|i=1}$ et $\pi_{1|i=2}$ seraient égaux à $\pi_{\bullet 1}$. Cependant, dans cet échantillon, la proportion d'hommes ayant l'intention de voter démocrate est plus faible que la proportion totale de démocrates et, à l'inverse, la même proportion pour les femmes est plus élevée que la proportion totale. On observe le phénomène inverse chez les républicains. Ainsi, l'association entre les intentions de vote et le sexe vient d'une popularité un peu plus grande du parti démocrate chez les femmes comparativement aux hommes et d'une popularité un peu plus grande du parti républicain chez les hommes comparativement aux femmes.

2.3.2 Résidus

Il est aussi possible de définir des résidus associés à chaque cellule d'un tableau de fréquences. Ces résidus sont en lien avec les modèles loglinéaires pour tableaux de fréquences. Nous n'avons pas besoin ici de décrire ce que sont ces modèles. Il suffit de savoir que les résidus présentés proviennent d'un modèle représentant l'indépendance entre X et Y. Ainsi, si un résidu est petit, le modèle d'indépendance s'ajuste bien, s'il est grand, il ne s'ajuste pas bien. Les cases associées à de grands résidus sont donc celles responsables de la dépendance entre les variables, si présente.

Plusieurs types de résidus existent, nous verrons ici les résidus bruts et ceux de Pearson, ajustés ou non. Ces résidus peuvent être utilisés pour ajouter de l'information à un diagramme en mosaïque. Par exemple, plus le résidu d'une case est grand, plus la couleur de la case peut être foncée. Les couleurs foncées permettent ainsi d'identifier rapidement les cases responsables d'un écart à l'hypothèse d'indépendance.

Résidus bruts

Il s'agit simplement de la différence entre une fréquence et l'estimation de sa valeur espérée sous H_0 :

$$RB_{ij} = n_{ij} - \hat{\mu}_{ij} = n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n}.$$

Résidus de Pearson

Les résidus de Pearson représentent les parts de chaque case du tableau à la statistique X^2 de Pearson. Ces résidus sont les éléments de la somme dans cette statistique, mais non élevés au carré :

$$RP_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}}} = \frac{n_{ij} - n_{i\bullet} n_{\bullet j}/n}{\sqrt{n_{i\bullet} n_{\bullet j}/n}}.$$

On voit bien que $\sum_{i,j} RP_{ij}^2 = X^2$. Sous l'hypothèse nulle d'indépendance, ces résidus suivent asymptotiquement une loi normale d'espérance nulle, mais de variance pas obligatoirement égale à 1.

Résidus de Pearson ajustés

Afin de définir des résidus de loi asymptotique normale standard sous H_0, il suffit d'ajuster les résidus de Pearson comme suit :

$$RAP_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1 - \hat{\pi}_{i\bullet})(1 - \hat{\pi}_{\bullet j})}} = \frac{n_{ij} - n_{i\bullet} n_{\bullet j}/n}{\sqrt{(n_{i\bullet} n_{\bullet j}/n)(1 - n_{i\bullet}/n)(1 - n_{\bullet j}/n)}}.$$

Puisque sous l'hypothèse d'indépendance les RAP_{ij} suivent asymptotiquement une $N(0,1)$, alors au seuil α une cellule pour laquelle la valeur observée de RAP_{ij} est supérieur à $z_{\alpha/2}$ est une cellule où l'hypothèse d'indépendance est violée.

Exemple de calcul de résidus : intentions de vote selon le sexe

Dans cet exemple, nous calculons les valeurs observées suivantes pour les différents résidus que l'on vient de définir.

	Démocrate	Indépendant	Républicain	Total
Femme	$n_{11} = 279$ $\hat{\mu}_{11}^{H_0} = 261.4$ $rb_{11} = 17.6$ $rp_{11} = 1.09$ $rap_{11} = 2.29$	$n_{12} = 73$ $\hat{\mu}_{12}^{H_0} = 70.7$ $rb_{12} = 2.3$ $rp_{12} = 0.28$ $rap_{12} = 0.46$	$n_{13} = 225$ $\hat{\mu}_{13}^{H_0} = 244.9$ $rb_{13} = -19.9$ $rp_{13} = -1.27$ $rap_{13} = -2.62$	$n_{1\bullet} = 577$
Homme	$n_{21} = 165$ $\hat{\mu}_{21}^{H_0} = 182.6$ $rb_{21} = -17.6$ $rp_{21} = -1.30$ $rap_{21} = -2.29$	$n_{22} = 47$ $\hat{\mu}_{22}^{H_0} = 49.3$ $rb_{22} = -2.3$ $rp_{22} = -0.33$ $rap_{22} = -0.46$	$n_{23} = 191$ $\hat{\mu}_{23}^{H_0} = 171.1$ $rb_{23} = 19.9$ $rp_{23} = 2.62$ $rap_{23} = 1.52$	$n_{1\bullet} = 403$
Total	$n_{\bullet 1} = 444$	$n_{\bullet 2} = 120$	$n_{\bullet 3} = 416$	$n = 980$

Par exemple, la valeur observée du résidu RAP_{11} est calculée par :

$$rap_{11} = \frac{279 - 261.4}{\sqrt{261.4(1 - (577/980))(1 - (444/980)}} = 2.29.$$

Ce résidu est positif et supérieur à $z_{0.05/2} = 1.96$. Il met donc en évidence le fait que les femmes s'identifient au parti démocrate dans une proportion plus grande que les hommes. Les résidus observés supérieurs, en valeur absolue, à 1.96 sont ceux des partis démocrate et républicain. Ce sont donc ces partis qui sont responsables de la dépendance entre le sexe et les intentions de vote.

La figure 2.7 présente le diagramme en mosaïque pour ces données. De la couleur a été ajoutée aux cases en fonction des résidus de Pearson des cases du tableau. Le bleu est associé aux résidus positifs et le rouge aux négatifs. De la couleur apparaît dans une case uniquement si la valeur observée du résidu est supérieure à 1 en valeur absolue.

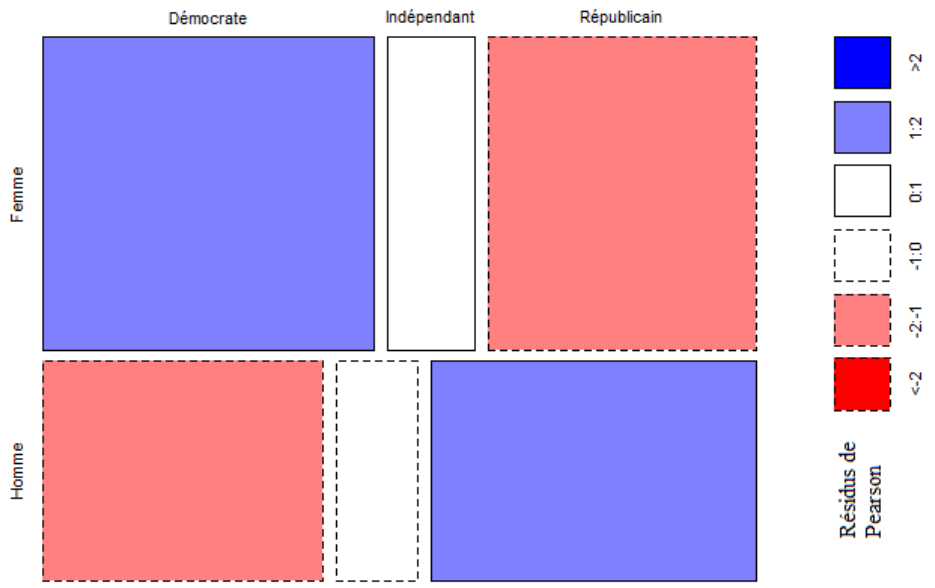


FIGURE 2.7 – Diagrammes en mosaïque pour les données des intentions de vote selon le sexe, avec de la couleur selon la valeur des résidus de Pearson.

2.3.3 Coefficient de Cramer

Il existe quelques mesures d'association basées sur la statistique X^2 de Pearson. À la manière d'un coefficient de corrélation, ces mesures sont utilisées pour juger de la force de l'association entre les variables catégoriques nominales X et Y. Une des plus populaires de ces mesures, particulièrement utilisée en psychologie, est le coefficient de Cramer. Il porte aussi parfois le nom V de Cramer. Ce coefficient est défini ainsi :

$$V = \sqrt{\frac{X^2/n}{\min\left(I-1, J-1\right)}}.$$

Le but de ce coefficient est d'avoir une mesure qui ne dépend pas de l'échelle des données (grandes ou petites fréquences). Cette mesure contient uniquement de l'information sur la force de l'association entre les variables. Elle permet donc de comparer la force de différentes associations, qui ne réfèrent pas nécessairement aux mêmes variables.

Dans le cas général d'un tableau $I \times J$, les valeurs possibles de V sont entre 0 et 1 inclusivement. Ce résultat s'explique par le fait que $n \times \min(I-1, J-1)$ est la valeur maximale de X^2. Un V prenant la valeur de 0 correspond à une statistique X^2 de Pearson nulle, donc à l'indépendance entre X et Y. À l'inverse, un V ayant une valeur observée près de 1 signifie que X et Y sont fortement dépendantes.

Pour un tableau 2×2, on permet parfois à V de prendre des valeurs négatives en changeant légèrement sa définition. Il devient :

$$V = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1\bullet}n_{2\bullet}n_{\bullet 1}n_{\bullet 2}}}.$$

Dans ce cas, V prend une valeur entre -1 et 1. À la manière d'un coefficient de corrélation, une valeur proche de 1 ou -1 représente une forte association entre X et Y et une valeur proche de 0 représente une absence d'association. Une valeur positive de V signifie que si X prend la modalité m_1^X alors Y a tendance a prendre aussi sa première modalité, soit m_1^Y. C'est aussi vrai pour la deuxième modalité des variables. Une valeur négative de V signifie plutôt que si X prend sa première modalité m_1^X, alors Y a tendance a prendre sa deuxième modalité, soit m_2^Y.

D'autres coefficients similaires existent, mais ne seront pas décrits ici. Il s'agit notamment du coefficient Phi et du coefficient de contingence (voir Conover, 1999, section 4.4).

2.3.4 Cas particulier des tableaux 2×2 : différence de proportions

Si les variables X et Y ont chacune seulement deux modalités, on a mentionné précédemment que l'homogénéité des sous-populations était en fait l'égalité entre les probabilités que Y prenne une valeur choisie, conditionnement à ce que X prenne sa première ou sa deuxième modalité. On a aussi montré que l'indépendance entre X et Y était l'équivalent de l'homogénéité des sous-populations. Alors les hypothèses nulles suivantes sont toutes équivalentes entre elles :

$$H_0 : \pi_{ij} = \pi_{i\bullet}\pi_{\bullet j} \ \forall i, j \Leftrightarrow (\pi_{1|i=1}, \pi_{2|i=1}) = (\pi_{1|i=2}, \pi_{2|i=2}) \Leftrightarrow \pi_{1|i=1} = \pi_{1|i=2}.$$

La dernière hypothèse revient à dire $\pi_{1|i=1} - \pi_{1|i=2} = 0$. Ainsi, la différence entre les proportions $\pi_1 = \pi_{1|i=1}$ et $\pi_2 = \pi_{1|i=2}$ est une mesure de l'association

entre X et Y. Si la différence est proche de 0, c'est le signe que les variables sont indépendantes. Si au contraire cette différence est grande, c'est le signe que les variables ne sont pas indépendantes.

Estimation ponctuelle : Un estimateur de la différence de proportion $\pi_1 - \pi_2$ est bien sûr

$$\hat{\pi}_1 - \hat{\pi}_2 = n_{11}/n_{1\bullet} - n_{21}/n_{2\bullet}.$$

Intervalle de confiance : Du test de Wald de comparaison de ces deux proportions, on peut déduire l'intervalle de confiance à $(1-\alpha)\%$ pour la différence entre les proportions suivant :

$$\pi_1 - \pi_2 \ \epsilon \ \hat{\pi}_1 - \hat{\pi}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_{1\bullet}} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_{2\bullet}}}.$$

Étant donné l'équivalence entre les tests et les intervalles de confiance, on peut affirmer que : si la valeur 0 n'est pas incluse dans cet intervalle de confiance, alors X et Y ne sont pas indépendantes.

Exemple d'estimation d'une différence de proportions :
étude expérimentale concernant l'aspirine et l'infarctus du myocarde.

Dans cet exemple, on travaille avec de faibles proportions. La différence entre $\pi_1 = \pi_{1|i=1}$ et $\pi_2 = \pi_{1|i=2}$ est estimée à partir de l'échantillon observé par :

$$\hat{\pi}_1 - \hat{\pi}_2 = \frac{n_{11}}{n_{1\bullet}} - \frac{n_{21}}{n_{2\bullet}} = 0.02166 - 0.01259 = 0.00907.$$

Cette valeur est petite, mais ça ne signifie pas pour autant qu'elle n'est pas significativement différente de 0. Calculons un intervalle de confiance à 95% de cette différence de proportions pour s'en convaincre :

$$\pi_1 - \pi_2 \ \epsilon \ 0.00907 \pm 1.96\sqrt{\frac{0.02166(1-0.02166)}{11304} + \frac{0.01259(1-0.01259)}{11307}}.$$

Donc $\pi_1 - \pi_2 \ \epsilon \ [0.0056, 0.0125]$. Cet intervalle de confiance ne contient pas la valeur 0. Ainsi, bien que la différence entre les proportions soit petite, $\pi_1 - \pi_2$

est significativement différente de 0. En d'autres mots, l'aspirine a un impact significatif sur la survenue d'un infarctus du myocarde. De plus, la différence $\pi_1 - \pi_2$ étant positive, nous avons $\pi_1 > \pi_2$ donc l'aspirine semble réduire les risques d'infarctus.

2.3.5 Cas particulier des tableaux 2×2 : risque relatif

Étant donné que la différence des proportions est dépendante de l'échelle de grandeur des proportions comparées, il est difficile de juger à partir de sa valeur de la force de l'association entre X et Y. Nous n'aurions pas ce problème si nous calculions le ratio des deux probabilités plutôt que la différence entre les deux. En effet, dans le ratio π_1/π_2, la division entre les deux proportions permet de se débarrasser de l'échelle de grandeur de celles-ci. Ce ratio est appelé risque relatif (en anglais relative risk). Il porte ce nom parce que dans une étude épidémiologique Y est typiquement une indicatrice de l'apparition d'une maladie alors que X définit des groupes d'exposition à un certain facteur de risque de développer la maladie. Alors, si la première modalité de Y représente bien l'apparition de la maladie, alors $\pi_1 = \pi_{1|i=1}$ est la probabilité d'avoir la maladie pour les individus du premier groupe, alors que $\pi_2 = \pi_{1|i=2}$ est la même probabilité, mais pour les individus du deuxième groupe. Ainsi, $RR = \pi_1/\pi_2$ est le risque relatif du premier groupe par rapport au deuxième de développer la maladie.

Interprétation : Remarquez que l'hypothèse d'homogénéité $H_0 : \pi_1 = \pi_2$ est équivalente à $H_0 : RR = \pi_1/\pi_2 = 1$. Ainsi, on interprète le risque relatif ainsi :

- si $RR = 1$: π_1 est égale à π_2
- si $RR > 1$: π_1 est RR fois plus grande que π_2 ;
- si $RR < 1$: π_1 est 1/RR fois plus petite que π_2.

Dans le cas où RR se situe entre 0 et 2, on peut aussi parler en terme de pourcentage comme ceci :

- si $RR > 1$: π_1 est (RR-1)×100% plus grande que π_2 ;
- si $RR < 1$: π_1 est (1-RR)×100% plus petite que π_2.

Estimation ponctuelle : Le risque relatif est estimé par :

$$\widehat{RR} = \frac{\hat{\pi}_1}{\hat{\pi}_2} = \frac{n_{11}/n_{1\bullet}}{n_{21}/n_{2\bullet}}.$$

Intervalle de confiance : Pour construire un intervalle de confiance du risque relatif, on passe par le logarithme de celui-ci. On peut montrer lorsque n est grand (Agresti, 2002, section 3.1.4) que $\ln(\widehat{RR})$ suit une loi normale de paramètres suivants :

$$\begin{aligned} E[\ln(\widehat{RR})] &\approx \ln(RR) \\ Var[\ln(\widehat{RR})] &\approx \frac{1-\pi_1}{\pi_1 n_{1\bullet}} + \frac{1-\pi_2}{\pi_2 n_{2\bullet}} \end{aligned}$$

Dans l'intervalle de confiance, on va utiliser un estimateur de la racine de cette variance, appelée erreur-type asymptotique :

$$\begin{aligned} \hat{\sigma}(\ln(\widehat{RR})) = \sqrt{\widehat{Var}[\ln(\widehat{RR})]} &= \sqrt{\frac{1-\hat{\pi}_1}{\hat{\pi}_1 n_{1\bullet}} + \frac{1-\hat{\pi}_2}{\hat{\pi}_2 n_{2\bullet}}} \\ &= \sqrt{\frac{1-n_{11}/n_{1\bullet}}{n_{11}} + \frac{1-n_{11}/n_{2\bullet}}{n_{21}}} \\ &= \sqrt{\frac{1}{n_{11}} - \frac{1}{n_{1\bullet}} + \frac{1}{n_{21}} - \frac{1}{n_{2\bullet}}}. \end{aligned}$$

Ainsi, un intervalle de confiance asymptotique de niveau $(1-\alpha)\%$ pour $\ln(RR)$ est :

$$\ln(\widehat{RR}) \pm z_{\alpha/2}\hat{\sigma}(\ln(\widehat{RR})).$$

Étant donné que $L < \ln(RR) < U$ si et seulement si $\exp(L) < RR < \exp(U)$, on obtient l'intervalle de confiance asymptotique de niveau $(1-\alpha)\%$ pour RR suivant :

$$\left[e^{\ln(\widehat{RR}) - z_{\alpha/2}\hat{\sigma}(\ln(\widehat{RR}))},\ e^{\ln(\widehat{RR}) + z_{\alpha/2}\hat{\sigma}(\ln(\widehat{RR}))}\right]$$
$$\left[\widehat{RR}\ e^{-z_{\alpha/2}\hat{\sigma}(\ln(\widehat{RR}))},\ \widehat{RR}\ e^{z_{\alpha/2}\hat{\sigma}(\ln(\widehat{RR}))}\right].$$

Rappelons que l'hypothèse d'homogénéité $\pi_1 = \pi_2$ est équivalente à $RR = \pi_1/\pi_2 = 1$. Ainsi, si la valeur 1 n'est pas incluse dans l'intervalle de confiance du risque relatif, alors π_1 est significativement différent de π_2, ce qui est équivalent à dire que X et Y ne sont pas indépendantes.

Exemple d'estimation du risque relatif :
étude expérimentale concernant l'aspirine et l'infarctus du myocarde.

Ici, on estime le risque relatif par :

$$\widehat{RR}_{obs} = \frac{\hat{\pi}_1}{\hat{\pi}_2} = \frac{n_{11}/n_{1\bullet}}{n/n_{2\bullet}} = \frac{239/11034}{139/11037} = 1.719892$$

Ainsi, le risque d'infarctus du myocarde est 72% fois plus élevé chez le groupe prenant le placebo que chez le groupe prenant l'aspirine. Ce pourcentage est obtenu en prenant $(\widehat{RR}_{obs} - 1) \times 100\% = (1.719892 - 1) \times 100\% \approx 72\%$.

La valeur observée de l'erreur-type de $\ln(\widehat{RR})$ est :

$$\begin{aligned} \hat{\sigma}(\ln(\widehat{RR}))_{obs} &= \sqrt{\frac{1}{n_{11}} - \frac{1}{n_{1\bullet}} + \frac{1}{n_{21}} - \frac{1}{n_{2\bullet}}} \\ &= \sqrt{\frac{1}{239} - \frac{1}{11034} + \frac{1}{139} - \frac{1}{11037}} = 0.1058164 \end{aligned}$$

Ainsi, un intervalle de confiance à 95% de RR est :

$$[1.719892\ e^{-1.96\times 0.1058164},\ 1.719892\ e^{1.96\times 0.1058164}]$$
$$[1.397747, 2.116284]\,.$$

Cet intervalle de confiance est conforme à tous les résultats précédents pour cet exemple. Encore une fois, étant donné que 1 n'est pas dans l'intervalle de confiance, et que l'estimation du risque relatif est positive, on peut affirmer que l'aspirine réduit les risques d'infarctus.

2.3.6 Cas particulier des tableaux 2×2 : rapport de cotes (odds ratio)

Une autre mesure de l'association entre les variables X et Y lorsque les deux variables sont dichotomiques nominales est le rapport de cotes. Cette mesure est en lien avec les modèles de régression logistique que nous verrons

plus loin.

La cote d'un événement A est définie comme étant le rapport $\frac{\pi}{1-\pi}$ où $\pi = P(A)$. Tout comme sa probabilité, la cote d'un événement est une mesure de sa vraisemblance. Une probabilité prend des valeurs entre 0 et 1, alors qu'une cote prend des valeurs entre 0 et l'infini. Si $\pi > 0.5$, ou de façon équivalente $\frac{\pi}{1-\pi} > 1$, on dira que l'événement A a plus de chance de survenir que de ne pas survenir.

Pour un tableau de fréquence 2×2, un rapport de cotes est défini comme suit :

$$RC = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}$$

où $\pi_i = \pi_{1|i} = P(Y = m_1^Y | X = m_i^X)$ pour $i = 1, 2$. Il s'agit donc du rapport entre les cotes de l'événement $Y = m_1^Y$ pour les deux groupes ou sous-populations formés par X. Cette formule du rapport de cote fait intervenir des probabilités conditionnelles. On peut aussi définir le rapport de cotes en terme des probabilités conjointes comme suit :

$$RC = \frac{\pi_{1|i=1}(1-\pi_{1|i=2})}{\pi_{1|i=2}(1-\pi_{1|i=1})} = \frac{\pi_{1|i=1}\pi_{2|i=2}}{\pi_{1|i=2}\pi_{2|i=1}} = \frac{(\pi_{11}/\pi_{1\bullet})(\pi_{22}/\pi_{2\bullet})}{(\pi_{21}/\pi_{2\bullet})(\pi_{12}/\pi_{1\bullet})} = \frac{\pi_{11}\pi_{22}}{\pi_{21}\pi_{12}}.$$

Notez que si on interchange les rôles de X et Y, donc si on travaille avec la cote de l'événement $X = m_1^X$ conditionnement à la valeur de Y, on obtient le même rapport de cotes puisque :

$$\frac{\pi_{1|j=1}(1-\pi_{1|j=2})}{\pi_{1|j=2}(1-\pi_{1|j=1})} = \frac{\pi_{1|j=1}\pi_{2|j=2}}{\pi_{1|j=2}\pi_{2|j=1}} = \frac{(\pi_{11}/\pi_{\bullet 1})(\pi_{22}/\pi_{\bullet 2})}{(\pi_{12}/\pi_{\bullet 2})(\pi_{21}/\pi_{\bullet 1})} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}.$$

Interprétation : L'indépendance entre X et Y, qui est équivalente à $\pi_1 = \pi_2$, est aussi équivalente à $RC = 1$ (à prouver en exercice). On interprète le rapport de cotes ainsi :

- si $RC = 1$: π_1 est égale à π_2
- si $RR > 1$: π_1 est plus grande que π_2 ;
- si $RR < 1$: π_1 est plus petite que π_2.

De plus, lorsque les probabilités π_1 et π_2 sont faibles, le rapport de cotes est approximativement égal au risque relatif. En effet, on a que :

$$RC = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)} = RR\frac{(1-\pi_2)}{(1-\pi_1)}.$$

Si π_1 et π_2 sont petits, alors $1-\pi_1$ et $1-\pi_2$ valent approximativement 1 et $(1-\pi_1)/(1-\pi_2) \approx 1$. Ainsi, dans ce cas, le rapport de cotes peut remplacer le risque relatif et permettre de quantifier de combien le risque est plus grand pour le premier groupe ($X = m_1^X$) comparativement au deuxième groupe ($X = m_2^X$).

Estimation ponctuelle : Le rapport de cotes est estimé par :

$$\widehat{RC} = \frac{\hat{\pi}_{11}\hat{\pi}_{22}}{\hat{\pi}_{21}\hat{\pi}_{12}} = \frac{(n_{11}/n)(n_{22}/n)}{(n_{21}/n)(n_{12}/n)} = \frac{n_{11}n_{22}}{n_{21}n_{12}}.$$

Intervalle de confiance : Pour construire un intervalle de confiance du rapport de cotes, on passe par le logarithme comme pour le risque relatif. On peut montrer que $\ln(\widehat{RC})$ suit asymptotiquement une loi normale de paramètres suivants (Agresti, 2002, section 3.1.1) :

$$\begin{aligned} E[\ln(\widehat{RC})] &\approx \ln(RC) \\ Var[\ln(\widehat{RC})] &\approx \frac{1}{n\pi_{11}} + \frac{1}{n\pi_{12}} + \frac{1}{n\pi_{21}} + \frac{1}{n\pi_{22}} \end{aligned}$$

Dans l'intervalle de confiance, on va utiliser un estimateur de la racine de cette variance, appelée erreur-type asymptotique :

$$\begin{aligned} \hat{\sigma}(\ln(\widehat{RC})) &= \sqrt{\frac{1}{n\hat{\pi}_{11}} + \frac{1}{n\hat{\pi}_{12}} + \frac{1}{n\hat{\pi}_{21}} + \frac{1}{n\hat{\pi}_{22}}} \\ &= \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}. \end{aligned}$$

Ainsi, un intervalle de confiance asymptotique de niveau $(1-\alpha)\%$ pour $\ln(RC)$ est :

$$\ln(\widehat{RC}) \pm z_{\alpha/2}\hat{\sigma}(\ln(\widehat{RC})).$$

Comme pour le risque relatif, l'intervalle de confiance asymptotique de niveau $(1-\alpha)\%$ pour RC est obtenu en appliquant la fonction exponentielle aux deux bornes de l'intervalle de confiance du logarithme du rapport de cotes. On obtient :

$$\left[\widehat{RC}\ e^{-z_{\alpha/2}\hat{\sigma}(\ln(\widehat{RC}))},\ \widehat{RC}\ e^{z_{\alpha/2}\hat{\sigma}(\ln(\widehat{RC}))}\right].$$

Si la valeur 1 n'est pas incluse dans l'intervalle de confiance du rapport de cotes, alors on peut conclure que π_1 est significativement différent de π_2.

Exemple de calcul de rapport de cotes :
étude expérimentale concernant l'aspirine et l'infarctus du myocarde.

Ici, on estime le rapport de cotes par :

$$\widehat{RC}_{obs} = \frac{\hat{\pi}_{11}\hat{\pi}_{22}}{\hat{\pi}_{21}\hat{\pi}_{12}} = \frac{(n_{11}/n)(n_{22}/n)}{(n_{21}/n)(n_{12}/n)} = \frac{n_{11}n_{22}}{n_{21}n_{12}} = \frac{239 \times 10898}{139 \times 10795} = 1.73583.$$

Ainsi, le risque d'infarctus du myocarde est plus élevé chez le groupe prenant le placebo que chez le groupe prenant l'aspirine. Mais est-il significativement plus élevé ? Pour le savoir, calculons un intervalle de confiance du rapport de cotes et voyons si la valeur 1 est dans cet intervalle. La valeur observée de l'erreur-type de $\ln(\widehat{RC})$ est :

$$\begin{aligned} \hat{\sigma}(\ln(\widehat{RC}))_{obs} &= \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} \\ &= \sqrt{\frac{1}{239} + \frac{1}{10795} + \frac{1}{139} + \frac{1}{10898}} = 0.1075302 \end{aligned}$$

Ainsi, un intervalle de confiance à 95% de RC est :

$$[1.73583\ e^{-1.96\times 0.1075302},\ 1.73583\ e^{1.96\times 0.1075302}]$$
$$[1.405969, 2.143082]\,.$$

Puisque la valeur 1 n'est pas dans l'intervalle de confiance, et que l'estimation du rapport de cotes est positive, on peut affirmer que l'aspirine réduit les risques d'infarctus.

Remarque : Ici, la probabilité d'infarctus étant faible dans les deux sous-populations, l'estimation du rapport de cotes est pratiquement identique à celle du risque relatif.

2.3.7 Quelles mesures demeurent utilisables lorsque l'échantillonnage est multiple ?

Les mesures d'association présentées dans cette section ne sont pas toujours estimables adéquatement à partir des données. Il faut toujours se préoccuper de la façon dont les données ont été recueillies, c'est-à-dire du type d'échantillonnage, avant d'utiliser une mesure.

Comme il a été expliqué à la section 2.1.2, si l'échantillonnage est simple, toutes les probabilités d'intérêt sont estimables sans problèmes. Cependant, si l'échantillonnage est multiple, on peut seulement être assuré d'estimer correctement les probabilités conditionnelles à la variable utilisée pour former les sous-populations. On estime aussi correctement les probabilités marginales à l'autre variable, celle non utilisée pour former les sous-populations, mais ces probabilités n'interviennent pas dans les mesures présentées ici. Il faut donc faire attention avec trois des mesures présentées ici :

- les probabilités conditionnelles ;
- la différence de proportions ;
- le risque relatif.

Ces mesures font toutes intervenir des probabilités conditionnelles à X. Elles sont donc estimables avec un échantillonnage multiple uniquement si les sous-populations ont été formées à partir des modalités de X. Dans une étude cas-témoins, ce n'est pas le cas.

Le rapport de cote a lui aussi été défini à partir de probabilités conditionnelles. Cependant, on a montré que, en réalité, il ne dépendait pas de la variable de conditionnement. Ainsi, il est correctement estimable à partir des données provenant d'un échantillonnage multiple, peu importe la variable formant les sous-populations. En fait, on pourrait se fier à la règle suivante pour savoir si une statistique est correctement estimable, peu importe le type d'échantillonnage : si la statistique ne change pas en inversant l'ordre des variables, alors on peut l'estimer correctement. Ainsi, les statistiques suivantes sont toujours estimables :

- la statistique X^2 de Pearson du test d'indépendance ou d'homogénéité de sous-populations ;
- la statistique G^2 du test d'indépendance ou d'homogénéité de sous-populations ;
- la statistique du test exact de Fisher (n_{11}) ;

– les résidus bruts ou de Pearson, ajustés ou non ;
– le coefficient de Cramer ;
– le rapport de cotes.

Remarque 1 : Même si le test score de comparaison de proportions est équivalent au test du khi-deux de Pearson, on évitera d'utiliser ce test si l'échantillonnage est multiple conditionnel à Y puisque les proportions elles-mêmes sont non-estimables.

Remarque 2 : Lorsque les données ont été recueillies en fixant préalablement le nombre d'individus pour lesquelles $Y = m_1^Y$ et $Y = m_2^Y$, comme dans une étude cas-témoins, l'échantillonnage est multiple conditionnel à Y. Dans ce cas, on ne peut pas utiliser le risque relatif puisqu'il n'est pas estimable. C'est problématique puisque cette mesure est souvent au coeur des questions de recherche. On veut fréquemment savoir de combien le risque d'un groupe est plus grand ou plus petit par rapport à un autre groupe. Si on peut supposer que la probabilité que $Y = m_1^Y$ soit petite, alors une solution s'offre à nous. C'est souvent le cas dans une étude cas-témoins, où $P(Y = m_1^Y)$ est typiquement la probabilité de développement d'une maladie rare. Dans ce cas, on peut interpréter le rapport de cotes comme un risque relatif.

Exemple 1 de données provenant d'un échantillonnage multiple : étude expérimentale concernant l'aspirine et l'infarctus du myocarde

Dans cet exemple, l'échantillonnage est multiple conditionnel à X. Ainsi, on pouvait utiliser toutes les mesures vues jusqu'ici dans ce chapitre.

Exemple 2 de données provenant d'un échantillonnage multiple : étude cas-témoin concernant la cigarette et l'infarctus du myocarde

Dans cet exemple, l'échantillonnage est multiple conditionnel à Y, une indicatrice du fait d'avoir été victime d'un infarctus du myocarde. En conséquence, on ne peut pas estimer de proportions conditionnelles à X, une indicatrice d'avoir déjà fumé. On veut répondre à la question de recherche

suivante :

Question : Est-ce que la cigarette augmente le risque d'infarctus du myocarde ?

Il serait inapproprié de formuler les hypothèses mathématiques d'un test en fonction de proportions conditionnelles à X puisqu'elles sont non estimables ici. On va donc plutôt tester l'indépendance entre X et Y, et ensuite utiliser l'équivalence mathématique entre l'indépendance et l'égalité des proportions $\pi_{1|i=1} = P(\text{infarctus} \mid \text{déjà fumé})$ et $\pi_{1|i=2} = P(\text{infarctus} \mid \text{jamais fumé})$. On pourrait faire ce test avec la statistique X^2 ou G^2. On peut aussi utiliser l'intervalle de confiance du rapport de cotes pour faire ce test. C'est ce que nous ferons ici.

Test d'hypothèses :
$H_0 : RC = 1$: il y a indépendance entre la consommation de cigarette et les infarctus du myocarde
$H_1 : RC \neq 1$

Résultats : Le rapport de cotes est estimé par :

$$\widehat{RC}_{obs} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{172 \times 346}{173 \times 90} = 3.822222.$$

Un intervalle de confiance à 95% de RC est :

$$\left[3.822222\ e^{-1.96\times\sqrt{\frac{1}{172}+\frac{1}{173}+\frac{1}{90}+\frac{1}{346}}},\ 3.822222\ e^{1.96\times\sqrt{\frac{1}{172}+\frac{1}{173}+\frac{1}{90}+\frac{1}{346}}}\right]$$
$$[2.793399, 5.229967]\,.$$

Ainsi, on peut rejeter H_0 au seuil de 5% puisque la valeur 1 n'est pas dans cet intervalle de confiance. Il y a une dépendance entre la consommation de cigarette et les infarctus du myocarde. Étant donné que l'indépendance implique que $\pi_{1|i=1} \neq \pi_{1|i=2}$ et que l'estimation du rapport de cotes est supérieur à 1, on peut même conclure que la consommation de cigarettes augmente le risque d'infarctus.

Estimation du risque relatif : On souhaite maintenant décrire l'association testée significative. Dans un cas comme celui-ci, il serait informatif

d'estimer de combien le risque d'infarctus est plus grand pour ceux qui ont déjà fumé versus ceux qui n'ont jamais fumé. C'est le risque relatif qui nous donne cette information. Cependant, on ne peut l'estimer directement ici à partir des données. Heureusement, on sait que les probabilités $\pi_{1|i=1}$ et $\pi_{1|i=2}$ sont faibles. Comme mentionné précédemment, la probabilité de subir un infarctus est de l'ordre de 1% ou 2% dans la population générale. Nous pouvons donc estimer le risque relatif par le rapport de cotes. Ainsi, $\widehat{RR}_{obs} \approx \widehat{RC}_{obs} = 3.822222$. Le risque d'infarctus du myocarde est donc environ 4 fois plus élevé pour ceux qui ont déjà fumé versus ceux qui n'ont jamais fumé. On pourrait aussi formuler cette conclusion ainsi : la consommation de cigarette multiplie par 4 le risque d'infarctus.

Remarquez que si, par erreur, on avait estimé le risque relatif par $\widehat{RR} = \frac{\hat{\pi}_1}{\hat{\pi}_2} = \frac{n_{11}/n_{1\bullet}}{n_{21}/n_{2\bullet}}$, on aurait obtenu l'estimation $\widehat{RR}_{obs} = \frac{172/345}{90/436} = 2.415201$. On aurait donc sous-estimé le véritable risque relatif.

2.4 Cas particulier des variables ordinales

Considérons maintenant le cas où la variable X ou Y (ou bien X et Y) est ordinale. Les tests d'association que nous avons vus jusqu'à présent peuvent en théorie s'appliquer, mais nous perdons de la puissance en n'exploitant pas la structure ordinale des données. Il est donc préférable d'utiliser des tests et mesures d'association prenant en compte l'ordinalité des données.

Exemple de variables ordinales :
malformations chez les nouveau-nés et alcool pendant la grossesse

Lors d'une étude longitudinale de cohorte avec échantillonnage multinomial simple, 32574 femmes enceintes ont été suivies pendant leur grossesse. Par un questionnaire, leur consommation d'alcool pendant le premier trimestre de leur grossesse a été évaluée. Après la naissance des enfants, des médecins ont vérifié si ceux-ci présentaient des malformations.

Question : Les mères qui consomment de l'alcool durant la grossesse augmentent-elles le risque de malformations de leurs enfants ?

Variable réponse Y : présence de malformations chez l'enfant à la naissance
Variable explicative X : nombre moyen de verres d'alcool consommés par jour par la mère

Les observations recueillies sont les suivantes :

X	Y : malformations			
alcool	absence	présence	Total	$P(\text{malformations} \mid X = m_i^X)$
0	17066	48	17114	$n_{11}/n_{1\bullet} = 0.0028$
< 1	14464	38	14502	$n_{21}/n_{2\bullet} = 0.0026$
1-2	788	5	793	$n_{31}/n_{3\bullet} = 0.0063$
3-5	126	1	127	$n_{41}/n_{4\bullet} = 0.0079$
≥ 6	37	1	38	$n_{51}/n_{5\bullet} = 0.0263$
Total	32481	93	32574	

Pour répondre à la question, on pourrait d'abord tester l'association entre les variables. On peut faire un test d'indépendance ou d'homogénéité avec la

statistique X^2 ou G^2. Sous H_0, les probabilités de malformations conditionnelles à la valeur de X sont toutes égales. On peut voir les estimations de ces probabilités dans le tableau ci-dessus. Cette probabilité semble augmenter avec une augmentation de la consommation d'alcool de la mère, mais est-ce de façon significative ? On obtient :

$X^2_{obs} = 12.1$, *seuil observé* $= P(\chi^2_4 > 12.1) = 0.02$.

$G^2_{obs} = 6.2$, *seuil observé* $= P(\chi^2_4 > 6.2) = 0.19$.

Ainsi, le X^2_{obs} nous pousserait à rejeter H_0, mais pas le G^2_{obs}. Ici, on ne peut avoir confiance en la distribution asymptotique khi-carré de ces statistiques en raison de quelques faibles fréquences pour les consommations d'alcool élevées. D'ailleurs, ces faibles fréquences expliquent la disparité entre X^2_{obs} et G^2_{obs}. On a déjà mentionné une solution à ce problème : regrouper des classes. Cependant, cette solution nous ferait perdre de l'information importante. Une meilleure solution ici serait de tenir compte du caractère ordinal des variables.

2.4.1 Association entre deux variables ordinales : coefficients de corrélation

Lorsque deux variables X et Y sont numériques, on peut facilement mesurer et tester l'association entre elles avec des méthodes classiques tels le coefficient de corrélation de Pearson et la régression linéaire simple. Avec deux variables catégoriques ordinales, on peut utiliser ces méthodes en attribuant un score numérique aux modalités des variables. Même une variable catégorique nominale comprenant seulement deux modalités peut être traitée de façon ordinale. Il suffit d'attribuer la valeur 1 à une modalité et 0 à l'autre.

Coefficient de corrélation de Pearson

Théoriquement, la corrélation entre deux variables aléatoires X et Y est définie par :

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

Cette statistique mesure l'association linéaire entre X et Y. Ce concept n'est pas tout à fait équivalent à l'indépendance entre X et Y. En effet, l'indépen-

dance entre X et $Y \Rightarrow$ l'absence d'association linéaire entre X et Y ($\rho = 0$), mais l'implication inverse n'est pas vraie.

L'estimateur le plus connu de ρ est le coefficient de corrélation de Pearson. Pour des variables numériques X et Y et des données sous le format individus (voir section 1.1.1), ce coefficient est le suivant :

$$\begin{aligned} r_p &= \frac{\sum_{u=1}^{n}(X_u - \bar{X})(Y_u - \bar{Y})}{\sqrt{\left(\sum_{u=1}^{n}(X_u - \bar{X})^2\right)\left(\sum_{u=1}^{n}(Y_u - \bar{Y})^2\right)}} \\ &= \frac{\sum_{u=1}^{n} X_u Y_u - n\bar{X}\bar{Y}}{\sqrt{\left(\sum_{u=1}^{n} X_u^2 - n\bar{X}^2\right)\left(\sum_{u=1}^{n} Y_u^2 - n\bar{Y}^2\right)}} \end{aligned}$$

Plaçons-nous maintenant dans un contexte de tableaux de fréquences $I \times J$. Supposons donc que les données sont sous un format fréquences. Supposons aussi que les variables X et Y sont catégoriques ordinales. Ainsi, leurs modalités sont ordonnables. Nous allons remplacer ces modalités par des scores numériques. Ainsi, les modalités $\{m_1^X, \ldots, m_I^X\}$ de X sont représentées par les scores $\{s_1^X, \ldots, s_I^X\}$, et les modalités $\{m_1^Y, \ldots, m_J^Y\}$ de Y sont représentées par les scores $\{s_1^Y, \ldots, s_J^Y\}$. La formule pour r_p devient :

$$\begin{aligned} r_p &= \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}(s_i^X - \overline{s^X})(s_j^Y - \overline{s^Y})}{\sqrt{\left(\sum_{i=1}^{I} n_{i\bullet}(s_i^X - \overline{s^X})^2\right)\left(\sum_{j=1}^{J} n_{\bullet j}(s_j^Y - \overline{s^Y})^2\right)}} \\ &= \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} s_i^X s_j^Y - \frac{1}{n}\left(\sum_{i=1}^{I} n_{i\bullet} s_i^X\right)\left(\sum_{j=1}^{J} n_{\bullet j} s_j^Y\right)}{\sqrt{\left(\sum_{i=1}^{I} n_{i\bullet}(s_i^X)^2 - \frac{1}{n}\left(\sum_{i=1}^{I} n_{i\bullet} s_i^X\right)^2\right)\left(\sum_{j=1}^{J} n_{\bullet j}(s_i^Y)^2 - \frac{1}{n}\left(\sum_{j=1}^{J} n_{\bullet j} s_j^Y\right)^2\right)}} \end{aligned}$$

Notons que le choix des scores numériques est arbitraire et la valeur de r_p changera si on change ces valeurs. Ceci arrivera même si, en théorie, le niveau d'association ne dépend pas des valeurs choisies. Alors s'il est possible de définir des scores représentatifs de la réalité, cette corrélation est adéquate. Si ce n'est pas le cas, il vaut mieux se tourner vers une statistique qui ne dépend pas de scores arbitraires, par exemple le coefficient de corrélation de Spearman.

Exemple de calcul d'un coefficient de corrélation de Pearson :

Voici un jeu de données fictif, sous le format individus et sous le format fréquences. La première variable X, prend les modalités 0, 1 et 2. La deuxième variable Y prend les modalités 0 et 1. Ces variables sont numériques discrètes avec peu de modalités possibles. Voici les observations recueillies :

Format individus :

Individu	X	Y	X^2	Y^2	$X \times Y$
1	0	0	0	0	0
2	0	0	0	0	0
3	0	1	0	1	0
4	1	0	1	0	0
5	1	0	1	0	0
6	1	1	1	1	1
7	1	1	1	1	1
8	2	0	4	0	0
9	2	1	4	1	2
10	2	1	4	1	2
Total	10	5	16	5	6

Format fréquences :

$X \setminus Y$	0	1	Total
0	2	1	3
1	2	2	4
2	1	2	3
Total	5	5	10

Avec ces données, on n'a pas besoin de créer des scores. Les valeurs sont déjà sur une échelle numérique (donc $s_i^X = m_i^X$ et $s_j^Y = m_j^Y$). Les trois dernières colonnes du tableau sous format individus servent à calculer le coefficient de corrélation de Pearson. La ligne « Total » de ce tableau nous informe que $\bar{x} = 10/10 = 1$, $\bar{y} = 5/10 = 0.5$, $\sum x_u^2 = 16$, $\sum y_u^2 = 5$ et $\sum x_u y_u = 6$.

Ainsi, avec la première formule énoncée pour le coefficient de corrélation de Pearson, on obtient :

$$r_{p,obs} = \frac{6 - 10 \times 1 \times 0.5}{\sqrt{(16 - 10 \times 1^2)\,(5 - 10 \times 0.5^2)}} = \frac{1}{\sqrt{15}} = 0.2581989.$$

Ce jeu de données est très petit dans l'unique but de bien illustrer les formules. En pratique, avec des données catégoriques et un jeu de données de grande taille, le calcul à la main est irréaliste avec la formule sous le format individus. Par contre, si le nombre de modalités des variables est petit, le calcul à la main avec la formule sous le format fréquences est tout à fait réalisable. Voici ce calcul pour ce jeu de données fictif. Calculons d'abord les sommes intervenant dans la formule :

$$\begin{aligned}
\sum_{i=1}^{I} n_{i\bullet} s_i^X &= 3 \times 0 + 4 \times 1 + 3 \times 2 = 10 \\
\sum_{j=1}^{J} n_{\bullet j} s_j^Y &= 5 \times 0 + 5 \times 1 = 5 \\
\sum_{i=1}^{I} n_{i\bullet} (s_i^X)^2 &= 3 \times 0^2 + 4 \times 1^2 + 3 \times 2^2 = 16 \\
\sum_{j=1}^{J} n_{\bullet j} (s_j^Y)^2 &= 5 \times 0^2 + 5 \times 1^2 = 5 \\
\sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} s_i^X s_i^Y &= 2 \times 0 \times 0 + 1 \times 0 \times 1 + 2 \times 1 \times 0 + 2 \times 1 \times 1 + \\
&\quad 1 \times 2 \times 0 + 2 \times 2 \times 1 = 6
\end{aligned}$$

On a donc

$$r_{p,obs} = \frac{6 - \frac{1}{10} \times 10 \times 5}{\sqrt{\left(16 - \frac{1}{10} \times 10^2\right)\left(5 - \frac{1}{10} \times 5^2\right)}} = \frac{1}{\sqrt{15}} = 0.2581989.$$

Coefficient de corrélation de Spearman

Afin d'éviter d'avoir à créer des scores numériques, on peut travailler uniquement avec les rangs des observations. C'est ce que fait la corrélation de Spearman. Soit rm_u^X le rang moyen de la $u^{\text{ième}}$ observation de X et rm_u^Y le rang moyen de la $u^{\text{ième}}$ observation de Y. Le coefficient de Spearman entre X et Y est défini comme étant le coefficient de corrélation de Pearson entre les couples de rangs moyens (rm_u^X, rm_u^Y) pour $u = 1, \ldots, n$. On le notera r_s. La formule pour calculer ce coefficient à partir de données sous le format individus est donc la suivante :

$$r_s = \frac{\displaystyle\sum_{u=1}^{n} rm_u^X \times rm_u^Y - n \times \overline{rm^X} \times \overline{rg^Y}}{\sqrt{\left(\displaystyle\sum_{u=1}^{n} (rm_u^X)^2 - n(\overline{rm^X})^2\right)\left(\displaystyle\sum_{u=1}^{n} (rm_u^Y)^2 - n(\overline{rm^Y})^2\right)}}.$$

Afin d'utiliser correctement cette formule, il faut d'abord savoir ce qu'est un rang moyen. Le rang d'une observation est le numéro de sa position parmi les observations ordonnées de la plus petite à la plus grande. Ainsi, la plus petite observation se voit attribuer le rang 1 et la plus grande le rang n s'il y a n observations. Cependant, que faire si un groupe d'observations sont toutes de même valeur ? Comment leur assigner un rang ? Par exemple, si les trois plus petites observations de X valent toutes 0 comme dans l'exemple fictif précédent, on pourrait leur attribuer au hasard les rangs 1, 2 et 3. Afin d'éviter une attribution au hasard des valeurs et pour être cohérent avec le fait que les trois observations prennent toutes la même valeur, on préfère souvent que le rang de ces observations prenne aussi une valeur commune. Il y a plusieurs possibilités de valeur commune pour ce rang. Par définition, une corrélation de Spearman se calcule avec des valeurs moyennes des rangs en cas d'égalité dans les observations. On parle donc de rangs moyens.

Exemple de calcul de rangs moyens :

Calculons maintenant les rangs moyens des observations du jeu de données fictif de l'exemple précédent.

Format individus :

Observation	X	rang de X	rm^X	Y	rang de Y	rm^Y
1	0	1	2	0	1	3
2	0	2	2	0	2	3
3	0	3	2	1	6	8
4	1	4	5.5	0	3	3
5	1	5	5.5	0	4	3
6	1	6	5.5	1	7	8
7	1	7	5.5	1	8	8
8	2	8	9	0	5	3
9	2	9	9	1	9	8
10	2	10	9	1	10	8

Voici deux exemples de calcul de rangs moyens pour ces données.

- La variable X prend la valeur 1 pour les observations numérotées 4, 5, 6 et 7. Les rangs ordinaires de ces observations sont 4, 5, 6 et 7, car trois individus ont une valeur de X inférieure à 1 et les trois autres individus ont une valeur de X supérieure à 1. Le rang moyen de X pour les observations numérotées 4, 5, 6 et 7 est donc la moyenne de leurs rangs ordinaires en X, soit $\frac{4+5+6+7}{4} = 5.5$.
- D'autre part, les observations numérotées 3, 6, 7, 9 et 10 ont toutes la valeur 1 en Y. Le rang moyen de Y pour ces observations est donc $\frac{6+7+8+9+10}{5} = 8$. Le plus petit rang ordinaire de cette moyenne est 6 parce que les 5 autres observations ont une valeur de Y inférieure à 1.

On peut déduire une formule pour calculer le rang moyen associé à une modalité d'une variable catégorique. Par exemple, intéressons-nous à la modalité m_i^X de X et notons :

$a =$ (nombre d'observations pour lesquelles $X < m_i^X$) $+ 1$;

$b =$ (nombre d'observations pour lesquelles $X \leq m_i^X$).

Le rang moyen à calculer est donc la somme des entiers entre a et b inclusivement, divisé par la fréquence de la modalité m_i^X, qui vaut en fait $b - a + 1$. Rappelons que la somme des entiers de 1 à b vaut $\frac{b(b+1)}{2}$. Il suffit de soustraire à cette somme la somme des entiers de 1 à $a - 1$ afin de calculer la somme

des entiers entre a et b inclusivement :

$$\frac{b(b+1)}{2} - \frac{(a-1)a}{2} = \frac{b^2+b-a^2+a}{2}.$$

On peut factoriser par $(a+b)(b-a+1)$ le numérateur de cette fraction. En divisant cette somme par $(b-a+1)$, on obtient le rang moyen en X des observations pour lesquelles X prend la modalité m_i^X, soit :

$$\frac{(a+b)(b-a+1)}{2(b-a+1)} = \frac{a+b}{2} = rm_i^X.$$

Le même raisonnement s'applique aux autres modalités de X ou à toute autre variable catégorique.

Exemple de calcul d'un coefficient de corrélation de Spearman :

Calculons maintenant le coefficient de corrélation de Spearman sur le jeu de données fictif de l'exemple précédent. Il nous faut d'abord calculer les rangs moyens des observations. Les voici :

Format individus :

Individu	X	Y	rm^X	rm^Y
1	0	0	2	3
2	0	0	2	3
3	0	1	2	8
4	1	0	5.5	3
5	1	0	5.5	3
6	1	1	5.5	8
7	1	1	5.5	8
8	2	0	9	3
9	2	1	9	8
10	2	1	9	8

Format fréquences :

$rm^X \setminus rm^Y$	3	8	Total
2	2	1	3
5.5	2	2	4
9	1	2	3
Total	5	5	10

où le rang moyen de la valeur 0 de X est $2 = \frac{1+3}{2}$, pour la valeur 1 de X il s'agit de $5.5 = \frac{4+7}{2}$ et pour la valeur 2 de X c'est $9 = \frac{8+10}{2}$. Les rangs moyens des modalités de Y sont $3 = \frac{1+5}{2}$ pour $m_1^Y = 0$ et $8 = \frac{6+10}{2}$ pour $m_2^Y = 1$. On calcule le coefficient de corrélation de Spearman pour ces observations de la

même façon que l'on a calculé le coefficient de corrélation de Pearson, mais en utilisant les valeurs de rm^X et rm^Y au lieu de prendre directement les valeurs de X et Y. On obtient $r_{s,obs} = 0.2581989$, soit exactement la même valeur que pour la corrélation de Pearson ! Ce résultat s'explique par le fait que les modalités de X, ainsi que celles de Y, sont toutes équidistantes.

Si on changeait la dernière modalité de X par 3 au lieu de 2, le coefficient de Pearson changeraient (0.2526456 au lieu de 0.2581989). Cependant, ce changement n'aurait aucun impact sur la corrélation de Spearman puisqu'il ne change pas les rangs des données.

Différences entre les coefficients de corrélation de Pearson et de Spearman

Voici un tableau énumérant des différences entre les deux coefficients de corrélation vus dans ce cours :

Pearson	vs	**Spearman**
- mesure le degré d'association linéaire entre X et Y		- mesure le degré d'association monotone entre X et Y
- est influencé par les valeurs extrêmes		- est robuste (pas influencé par les valeurs extrêmes)
- a besoin d'un score numérique		- a seulement besoin d'observations ordonnables

Interprétation des coefficients de corrélation

Les valeurs possibles des deux coefficients de corrélation se situent entre -1 et 1. On peut interpréter comme suit les valeurs des coefficients :

- $|r_p| \approx 1 \Rightarrow$ relation linéaire forte entre les variables ;
- $|r_s| \approx 1 \Rightarrow$ relation monotone forte entre les variables (strictement croissante ou décroissante) ;
- $r_p \approx 0$ ou $r_s \approx 0 \Rightarrow$ absence de relation (linéaire ou monotone) entre les variables ;
- $r_p > 0$ ou $r_s > 0 \Rightarrow$ association positive (plus X est grand, plus Y tend à être grand) ;
- $r_p < 0$ ou $r_s < 0 \Rightarrow$ association négative (plus X est grand, plus Y tend à être petit).

Test d'association entre deux variables ordinales : test de Mantel-Haenszel

En utilisant un coefficient de corrélation, on ne peut pas tout à fait tester l'indépendance entre les variables X et Y, mais on peut tester la présence ou non d'une association entre elles :

$$H_0 : X \text{ et } Y \text{ ne sont pas associées} \Leftrightarrow \rho = 0$$

Pour confronter cette hypothèse nulle à une hypothèse alternative bilatérale ou unilatérale, on peut se baser sur la statistique de test :

$$M = \sqrt{(n-1)}r$$

où r est soit le coefficient de corrélation de Pearson (r_p), soit celui de Spearman (r_s). Peu importe le coefficient choisi, sous H_0, la loi asymptotique de M est une $\mathcal{N}(0, 1)$.

Souvent, on utilise plutôt la statistique de test $M^2 = (n-1)r^2$, appelée khi-carré de Mantel-Haenszel. Sous H_0, la loi asymptotique de M^2 est une χ^2_1. Cette statistique ne permet cependant pas d'effectuer un test unilatéral.

Exemple d'association pour des variables ordinales :
malformations chez les nouveau-nés et alcool pendant la grossesse

Effectuons maintenant un test d'association entre les deux variables ordinales X et Y. Nous tenterons de faire ce test de deux façons : avec le coefficient de corrélation de Pearson et avec celui de Spearman. Il faut donc définir un score numérique pour représenter les modalités des variables, ainsi que calculer les rangs moyens des modalités.

Ici, on peut définir facilement des scores numériques très représentatifs des modalités. Pour la variable Y, le score 0 représentera l'absence de malformations et le score 1 la présence de malformations. Pour la variable X, le score numérique sera le centre des intervalles qui définissent les modalités. Pour la dernière modalité, soit ≥ 6, choisissons arbitrairement le score 7.

Pour calculer les rangs moyens, on utilise la formule donnée précédemment pour des données représentées sous le format fréquences. Par exemple, pour la modalité m_3^X, on a $a = 17114+14502+1 = 31617$ et $b = 17114+14502+793 = 32409$. Donc le rang moyen pour cette modalité est $(31617 + 32409)/2 = 32013$.

Le tableau suivant comprend les données sous le format fréquences ainsi que les scores numériques et les rangs moyens.

Rang moyen			16241	32528	
	Score		$s_1^Y = 0$	$s_2^Y = 1$	
		$X \setminus Y$	absence	présence	Total
8557.5	$s_1^X = 0$	0	$n_{11} = 17066$	$n_{12} = 48$	17114
24365.5	$s_2^X = 0.5$	< 1	$n_{21} = 14464$	$n_{22} = 38$	14502
32013	$s_3^X = 1.5$	1-2	$n_{31} = 788$	$n_{32} = 5$	793
32473	$s_4^X = 4$	3-5	$n_{41} = 126$	$n_{42} = 1$	127
32555.5	$s_5^X = 7$	≥ 6	$n_{51} = 37$	$n_{52} = 1$	38
		Total	32481	93	32574

Pour répondre à la question de recherche « Les mères qui consomment de l'alcool durant la grossesse augmentent-elles le risque de malformations chez leurs enfants ? », on va formuler les hypothèses ainsi :

$$H_0 : \rho = 0 : \text{il n'y a pas d'association entre } X \text{ et } Y,$$
$$H_1 : \rho > 0 : \text{il y a une association positive entre } X \text{ et } Y.$$

Résultats :

$$\begin{aligned} r_{p,obs} = 0.0142 \quad &\longrightarrow \quad M_{obs} = \sqrt{(32574-1)}0.0142 = 2.5632 \\ &\quad\; \textit{seuil observé} = P(N(0,1) > 2.5632) = 0.0052 \\ r_{s,obs} = 0.0033 \quad &\longrightarrow \quad M_{obs} = \sqrt{(32574-1)}0.0033 = 0.5928 \\ &\quad\; \textit{seuil observé} = P(N(0,1) > 0.5928) = 0.2767 \end{aligned}$$

Conclusion : En utilisant les scores $(0\,;\, 0.5\,;\, 1.5\,;\, 4.0\,;\, 7.0)$ et la corrélation de Pearson, on rejette H_0, il y a association positive entre la consommation d'alcool et la présence de malformations à la naissance. Par contre, en utilisant les rangs moyens et la corrélation de Spearman, on ne peut rejeter H_0.

Dans cet exemple, les modalités de la variable X ne sont pas du tout équidistantes. La méthode des rangs moyens ne fonctionne donc pas bien. Les scores numériques sont plus représentatifs de la réalité que les rangs moyens. La bonne conclusion est donc l'association positive entre la consommation d'alcool et la présence de malformations à la naissance.

Autres mesures d'association pour variables ordinales

Les coefficients de corrélation de Pearson et Spearman sont loin d'être les seules mesures à avoir été proposées pour mesurer l'association entre deux variables ordinales. Il existe d'autres corrélations, par exemple la corrélation polychorique, ainsi que plusieurs mesures basées sur les concordances et discordances (Agresti, 2002, section 2.4.3). Ces dernières mesures, tout comme la corrélation de Spearman, ne nécessitent aucun score numérique. Il suffit de pouvoir ordonner les observations pour pouvoir les calculer.

Équivalence avec d'autres méthodes

Lorsque le tableau de fréquences est de dimension $I \times 2$, la statistique M est la même que la statistique du test de tendance de Cochran-Armitage (Agresti, 2002, sections 3.4.6 et 5.3.5).

On pourrait aussi effectuer une régression linéaire simple avec les scores numériques. Le test sur le paramètre de la régression linéaire teste justement si la corrélation entre X et Y est nulle. Cependant, ce test n'est pas identique à ce qui a été présenté ici. Il se base sur une statistique de test similaire, mais de loi asymptotique Student.

2.4.2 Association entre une variable nominale et une variable ordinale

Si on croise une variable nominale et une variable ordinale, le plus simple est d'effectuer une ANOVA à un facteur pour comparer la valeur moyenne de la variable ordinale pour chacun des groupes formés par la variable nominale. On peut encore une fois travailler sur des rangs plutôt qu'avec un score numérique. On fait alors un test de Kruskal Wallis. Si la variable nominale n'a que deux modalités, ces tests reviennent au test de Mantel-Haenszel en donnant le score 1 à une des modalités de la variable et 0 à l'autre.

2.5 Cas particulier des données pairées

Nous considérons ici le cas particulier de tableaux $I \times I$ pour lesquels les variables X et Y sont en fait la même caractéristique mesurée à deux reprises sur les mêmes individus. Les deux mesures de la caractéristique peuvent se distinguer, par exemple, de la façon suivante :

- elles n'ont pas été prises au même moment ou au même endroit (mesures répétées dans le temps ou dans l'espace) ;
- deux personnes différentes ont pris les mesures ;
- les mesures ont été prises avec deux instruments de mesure différents.

Lorsqu'on est en présence de 2 mesures de la même caractéristique, on parle de « données pairées ». Dans le cas plus général d'un nombre quelconque de mesures de la même caractéristique sur les mêmes individus, on parlera plutôt de « données appariées » ou de « mesures répétées ».

On a donc ici deux variables prenant les mêmes modalités. Notons-les encore X et Y, mais posons maintenant que $I = J$ et $(m_1^X, \ldots, m_I^X) =$.

Exemple 1 de données pairées :

Deux médecins doivent diagnostiquer un groupe de 100 patients. Le diagnostic est malade/sain. Chaque patient est examiné par les 2 médecins. On obtient les résultats du tableau suivant :

		Second médecin		
		Malade	Sain	Total
Premier médecin	Malade	23	15	38
	Sain	20	42	62
	Total	43	57	100

Une question d'intérêt serait de savoir s'il y a une différence significative entre les diagnostics des 2 médecins. On veut donc comparer les probabilités marginales plutôt que les probabilités conditionnelles comme on l'a fait jusqu'ici dans ce chapitre. L'hypothèse nulle d'un test répondant à cette question de recherche serait $H_0 : \pi_{1\bullet} = \pi_{\bullet 1}$.

Exemple 2 de données pairées :

Un groupe de 1600 Canadiens en âge de voter participent à une étude. On leur demande tout d'abord leur opinion (très insatisfait, insatisfait, satisfait, très satisfait) au sujet de la performance du premier ministre canadien. On leur repose la même question 6 mois plus tard, et on obtient les résultats du tableau suivant.

		Second sondage				
		t. insat.	insat.	sat.	t. sat.	Total
Premier sondage	t insat.	188	84	18	0	290
	insat	52	338	191	6	587
	sat.	13	22	420	51	506
	t. sat.	0	1	2	214	217
	Total	253	445	631	271	1600

Une question d'intérêt est de savoir s'il y a une différence significative entre les résultats du premier et du second sondage. Si on se demandait si les deux variables sont associées, on saurait déjà comment répondre à la question. On effectuerait un test d'association entre deux variables ordinales basé sur un coefficient de corrélation. Cependant, on ne se demande pas ici si les deux variables sont associées, car on s'attend évidemment à ce qu'elles le soient. On veut plutôt étudier ici le degré d'accord entre les réponses des participants lors du premier et du second sondage. On a donc besoin de nouveaux outils.

2.5.1 Différents formats de jeux de données

On a déjà mentionné que les observations des variables représentées dans un tableau de fréquences peuvent se présenter sous deux formats : le format individus avec une ligne par individu (soit n lignes) et le format fréquences avec une ligne pour chaque combinaison des modalités des variables (soit $I \times J$ lignes si on a les deux variables X et Y telles que définies dans ce chapitre).

Pour des données pairées, un troisième format de jeux de données est fréquemment utilisé. Dans ce format de données, il y a une colonne identifiant la variable (X ou Y), une autre la modalité prise par la variable et une troisième identifiant l'individu. Le jeu de données comporte, pour chaque

individu, une ligne par variable. S'il y a deux variables, il comporte donc $2n$ lignes en tout. On l'appellera le format avec plus d'une ligne par individu.

Exemple 1 de données pairées :
différents formats de jeux de données

Sous le format le plus abrégé, soit le format fréquences, le jeu de données à l'allure suivante :

Médecin1	**Médecin2**	**Fréquence**
Malade	Malade	23
Malade	Sain	15
Sain	Malade	20
Sain	Sain	42

Il comporte donc seulement $I \times I = 2 \times 2 = 4$ lignes. Sous le format individus, il comporte $n = 100$ lignes. Ses dix premières lignes pourraient bien être les suivantes :

Individu	**Médecin1**	**Médecin2**
1	Malade	Malade
2	Sain	Sain
3	Malade	Sain
4	Sain	Malade
5	Sain	Sain
6	Malade	Sain
7	Sain	Malade
8	Malade	Malade
9	Sain	Sain
10	Sain	Sain
⋮	⋮	⋮

Le jeu de données sous le troisième format, soit celui avec plus d'une ligne par individu, aurait l'allure suivante :

Individu	**Médecin**	**Diagnostic**
1	1	Malade
1	2	Malade
2	1	Sain
2	2	Sain
3	1	Malade
3	2	Sain
4	1	Sain
4	2	Malade
5	1	Sain
5	2	Sain
⋮	⋮	⋮

Ce jeu de données comporterait $2n = 200$ lignes.

Il faut être prudent avec ce format de données. Il peut induire en erreur. En effet, on pourrait voir les deux variables comme étant X = le médecin et Y = le diagnostic plutôt que X = diagnostic du premier médecin et Y = diagnostic du deuxième médecin. On serait alors porté à construire le tableau de fréquences suivant :

		Diagnostic		
		Malade	Sain	Total
Médecin	1	38	62	100
	2	43	57	100
	Total	81	119	200

Le problème avec ce tableau est que les deux sous-populations que l'on pourrait imaginer en lignes sont en fait composées des mêmes individus. Chaque individu apporte deux comptes au tableau plutôt qu'un seul. Le n du tableau est donc 200 au lieu de 100. Ainsi, on ne respecte pas l'hypothèse de base d'indépendance entre les individus.

2.5.2 Sensibilité et spécificité

Les tableaux de fréquences 2×2 avec données pairées sont utilisés dans une application épidémiologique bien particulière : l'évaluation de la qualité d'un examen diagnostic. Pour faire cette évaluation, on utilise un échantillon de sujets dont on connaît l'état : malade ou non malade (sain). Cet état a été déterminé à l'aide d'un examen de référence dit « gold-standard ». C'est l'examen réputé être le plus fiable pour diagnostiquer la maladie en question. On fait ensuite passer aux sujets l'examen diagnostic à évaluer. On obtient des résultats représentés dans un tableau tel le suivant :

		Y = Vérité supposée	
		Malade	Sain
X = Résultat de l'examen	Positif (malade)	n_{11} = VP Vrais Positifs	n_{12} = FP Faux Positifs
	Négatif (sain)	n_{21} = FN Faux Négatifs	n_{11} = VN Vrais Négatifs

On est donc bien en présence de données pairées. Les deux variables mesurent la même caractéristique, l'état malade ou non malade de la personne, mais à l'aide de deux examens différents, l'un à évaluer et l'autre considéré comme étant très fiable.

On définit la sensibilité et la sensibilité d'un examen diagnostic comme suit (Beaucage et Viger, 1996, chapitre 6) :

sensibilité : probabilité d'obtenir un résultat positif pour un sujet malade
$= P(X = \text{Positif} \mid Y = \text{Malade})$

spécificité : probabilité d'obtenir un résultat négatif pour un sujet sain
$= P(X = \text{Négatif} \mid Y = \text{Sain})$

Ce sont les deux mesures courantes de la fiabilité d'un examen diagnostic. À partir des données recueillies, on estime ces mesures par :

- estimation de la sensibilité : $VP/(VP + FN)$;
- estimation de la spécificité : $VN/(VN + FP)$.

En pratique, ce qui importe une personne qui doit subir un examen diagnostic est plutôt :

– Si j'obtiens un résultat positif, quelles sont mes chances de réellement avoir la maladie ?
– Si j'obtiens un résultat négatif, quelles sont mes chances de réellement être non atteint de la maladie ?

Cette personne est donc intéressée par les probabilités $P(Y = \text{Malade} \mid X = \text{Positif})$ et $P(Y = \text{Sain} \mid X = \text{Négatif})$. Par le théorème de Bayes, si on connaît la probabilité pour une personne d'être atteinte de la maladie ($P(Y = \text{Malade})$), on peut trouver la valeur de ces probabilités conditionnelles à X à partir de la sensibilité et la spécificité, qui sont elles conditionnelles à Y.

Exemple de sensibilité et de spécificité :
la mammographie pour détecter le cancer du sein

Il est estimé que 1% des femmes qui subissent une mammographie à un temps donné sont atteintes du cancer du sein. On rapporte aussi que la mammographie a une sensibilité de 0.86 et une spécificité de 0.88 pour détecter un cancer du sein (Agresti, 2002, section 2.1.3). Ces valeurs sont relativement bonnes. Pourtant, en se basant sur elles, la probabilité d'être réellement atteinte du cancer du sein pour une femme qui reçoit un résultat de mammographie positif est de seulement 7%.

Vérifions ce résultat surprenant. Notons A l'événement « réellement avoir un cancer du sein » et B l'événement « recevoir un résultat de mammographie positif ». On veut calculer $P(A|B)$. On estime que $P(A) = 0.01$, sensibilité $= P(B|A) = 0.86$ et spécificité $= P(B^C|A^C) = 0.88$, où l'exposant C représente l'événement complémentaire. En vertu du théorème de Bayes (voir annexe A.1.2), on a :

$$\begin{aligned} P(A|B) &= \frac{P(B|A)P(A)}{P(B|A)P(A) + (1 - P(B^C|A^C))(1 - P(A))} \\ &= \frac{0.86 \times 0.01}{0.86 \times 0.01 + (1 - 0.88) \times (1 - 0.01)} = 0.0675 \end{aligned}$$

Cependant, la probabilité de ne pas être atteinte du cancer du sein pour une femme qui reçoit un résultat de mammographie négatif ($P(A^C|B^C)$) est 0.998. Les calculs pour obtenir cette valeur sont similaires aux calculs

précédents. Ainsi, la mammographie n'est que la première étape dans un diagnostic de cancer du sein. Il n'arrive à peu près jamais qu'une femme recevant un résultat négatif à une mammographie ait en fait un cancer du sein. La mammographie est donc bonne pour faire un premier tri parmi les femmes pour trouver celles potentiellement atteintes de ce cancer. Celles qui reçoivent un résultat de mammographie positif doivent subir d'autres examens pour poser un diagnostic final.

2.5.3 Test de la symétrie de la loi conjointe

La symétrie de la loi conjointe est un modèle mathématique utile pour juger de l'accord entre deux variables. On verra plus loin comment l'utiliser en pratique. On va ici définir ce nouveau concept et voir comment le tester. Dans un tableau de fréquences $I \times I$, on dit que la loi conjointe de X et Y est symétrique si $\pi_{ij} = \pi_{ji}$ pour tout couple (i, j). On va voir deux statistiques de test pour confronter les hypothèses suivantes :

$$\begin{array}{lcll} H_0 & : & \pi_{ij} = \pi_{ji} & \text{pour tout couple } (i,j) \\ H_1 & : & \pi_{ij} \neq \pi_{ji} & \text{pour au moins un couple } (i,j) \end{array}$$

– **Statistique du khi-deux de Pearson (test de Bowker)**

$$X^2_{sym} = \sum_{1 \leq i < j \leq I} \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}}.$$

– **Statistique du rapport de vraisemblance**

$$G^2_{sym} = \sum_{i=1}^{I} \sum_{j=1}^{I} n_{ij} \ln \frac{2n_{ij}}{n_{ij} + n_{ji}}.$$

Justification des formules pour les statistiques X^2_{sym} et G^2_{sym} :
Sous H_0, les estimateurs du maximum de vraisemblance des probabilités conjointes π_{ij} sont

$$\hat{\pi}_{ij} = \frac{n_{ij} + n_{ji}}{2n}$$

Remarquez que si $i = j$, alors $\hat{\pi}_{ii} = \frac{n_{ii}+n_{ii}}{2n} = \frac{n_{ii}}{n}$.

La statistique du khi-deux de Pearson se simplifie donc ainsi :

$$\begin{aligned} X^2_{sym} &= \sum_{i=1}^{I}\sum_{j=1}^{I} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} = \sum_{i=1}^{I}\sum_{j=1}^{I} \frac{(n_{ij} - n\hat{\pi}_{ij})^2}{n\hat{\pi}_{ij}} \\ &= \sum_{i=1}^{I}\sum_{j=1}^{I} \frac{(n_{ij} - n\frac{n_{ij}+n_{ji}}{2n})^2}{n\frac{n_{ij}+n_{ji}}{2n}} = \sum_{i=1}^{I}\sum_{j=1}^{I} \frac{(\frac{n_{ij}-n_{ji}}{2})^2}{\frac{n_{ij}+n_{ji}}{2}} = \sum_{i=1}^{I}\sum_{j=1}^{I} \frac{1}{2}\frac{(n_{ij} - n_{ji})^2}{n_{ij}+n_{ji}}. \end{aligned}$$

Les termes associés aux éléments sur la diagonale du tableau de fréquences ($i = j$) sont tous nuls puisque $n_{ii} - n_{ii} = 0$. Ainsi, la diagonale ne contribue pas à la statistique X^2_{sym}. En dehors de la diagonale, on a que les termes ij sont égaux aux termes ji puisque $\frac{(n_{ij}-n_{ji})^2}{n_{ij}+n_{ji}} = \frac{(n_{ji}-n_{ij})^2}{n_{ji}+n_{ij}}$. On peut donc changer la façon de noter la somme. Plutôt que de sommer sur tous les termes du tableau $I \times I$, on peut sommer seulement sur les termes sous la diagonale, mais en comptant chacun des termes 2 fois. On obtient ainsi

$$X^2_{sym} = \sum_{1\leq i<j\leq I} \frac{2}{2}\frac{(n_{ij} - n_{ji})^2}{n_{ij}+n_{ji}} = \sum_{1\leq i<j\leq I} \frac{(n_{ij} - n_{ji})^2}{n_{ij}+n_{ji}}.$$

Les calculs pour obtenir G^2_{sym} sont directs.

Distribution asymptotique de X^2_{sym} et G^2_{sym} sous H_0 :
Sous H_0, ces deux statistiques suivent asymptotiquement une loi $\chi^2_{\frac{I(I-1)}{2}}$. Les degrés de liberté du test se justifient comme suit :

$$\begin{aligned} \#\text{degrés de liberté} &= \text{(dim espace paramètre) - (dim espace sous } H_0) \\ &= (I^2 - 1) - \left(\frac{I(I+1)}{2} - 1\right) \\ &= \frac{I(I-1)}{2} \end{aligned}$$

2.5.4 Test d'homogénéité des marginales

Pour répondre à une question de recherche telle que formulée dans l'exemple 1 d'un diagnostic posé par deux médecins, on doit tester si les distributions

marginales de X et Y sont identiques. On teste donc

$$\begin{array}{lcll} H_0 & : & \pi_{i\bullet} = \pi_{\bullet i} \quad \forall i = 1, \ldots, I & \Leftrightarrow \qquad (\pi_{1\bullet}, \ldots, \pi_{I\bullet}) = (\pi_{\bullet 1}, \ldots, \pi_{\bullet I}) \\ H_1 & : & \pi_{i\bullet} \neq \pi_{\bullet i} \quad \text{pour au moins un } i & \end{array}$$

Cela implique donc que sous H_0, nous avons les contraintes suivantes :

$$\begin{array}{lrcl} (c_1) & \pi_{11} + \ldots + \pi_{1I} & = & \pi_{11} + \ldots + \pi_{I1}, \\ (c_2) & \pi_{21} + \ldots + \pi_{2I} & = & \pi_{12} + \ldots + \pi_{I2}, \\ & & \vdots & \\ (c_I) & \pi_{I1} + \ldots + \pi_{II} & = & \pi_{1I} + \ldots + \pi_{II}. \end{array}$$

en plus de la contrainte $\sum_{i=1}^{I} \sum_{i=1}^{J} \pi_{ij} = 1$. On a que la contrainte c_I est automatiquement satisfaite si les contraintes c_1 à c_{I-1} et la contrainte de la somme totale à 1 sont satisfaites. Il n'est pas possible de trouver explicitement le maximum de vraisemblance sous H_0, c'est-à-dire sous les contraintes ci-dessus. Cela requiert donc l'utilisation d'un logiciel, afin de trouver une solution numérique à l'étape de maximisation.

Une fois que les estimateurs du maximum de vraisemblance sous H_0 sont obtenus, le test du rapport de vraisemblance et du khi-deux de Pearson peuvent être effectués exactement de la même façon que d'habitude. La loi asymptotique des statistiques de ces tests est toujours une khi-deux de degrés de liberté :

$$\begin{array}{rcl} \#\text{degrés de liberté} & = & \text{(dim espace paramètre) - (dim espace sous } H_0) \\ & = & (I^2 - 1) - (I^2 - ((I - 1) + 1)) \\ & = & I - 1 \end{array}$$

Notons que le recours à une solution numérique complique l'utilisation de ce test en pratique. D'autres façons de tester l'homogénéité des marginales ont été proposées, notamment le test de Stuart-Maxwell (Agresti, 2007, section8.3.1) et le test de Bhapkar (voir Sun et Yang, 2008, pour plus d'information).

2.5.5 Lien entre les deux tests

Notons que

$$\text{Symétrie de la loi conjointe} \Rightarrow \text{Homogénéité des lois marginales}$$

mais que

$$\text{Homogénéité des lois marginales} \not\Rightarrow \text{Symétrie de la loi conjointe.}$$

Ainsi, la symétrie de la loi conjointe est une condition plus forte que l'homogénéité des lois marginales. Il y a donc un lien important entre le test d'homogénéité des lois marginales et le test de symétrie de la loi conjointe. Si on ne rejette pas l'hypothèse de symétrie de la loi conjointe, alors on ne rejette pas l'hypothèse d'homogénéité des lois marginales. De façon équivalente, si on rejette l'hypothèse d'homogénéité des lois marginales, alors on rejette l'hypothèse de symétrie de la loi conjointe.

Les deux tests sont parfois désignés comme étant des tests d'accord. Cependant, ils doivent être utilisés conjointement à des mesures d'accord pour permettre une interprétation correcte du degré d'accord dans les données.

2.5.6 Cas particulier du tableau 2×2 : le test de McNemar

Reconsidérons maintenant le cas des tests de symétrie de la loi conjointe et d'homogénéité des lois marginales dans le cas d'un tableau 2×2. Dans ce cas, on a :

$$\text{Symétrie de la loi conjointe} \Leftrightarrow \text{Homogénéité des lois marginales}$$

Il s'en suit que dans le cas 2×2, le test de symétrie de la loi conjointe et le test d'homogénéité des lois marginales sont exactement le même test. Notez aussi que les degrés de liberté obtenus dans les deux tests : $I(I-1)/2$ (symétrie) et $I-1$ (homogénéité) coïncident lorsque $I = 2$. La statistique du khi-deux de Pearson obtenue à la section 2.5.3 se simplifie énormément, et on obtient :

$$X^2_{sym} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} \qquad \xrightarrow[H_0]{\text{asymp.}} \qquad \chi^2_1.$$

Le test se basant sur cette statistique est appelé le test de McNemar.

Remarque : On pourrait aussi déduire de cette statistique une autre statistique se basant sur une loi normale. Cette statistique pourrait prendre des valeurs négatives et permettrait d'effectuer un test unilatéral. Ce test serait équivalent à un test apparié de comparaison de proportions.

2.5.7 Mesures d'accord entre les variables X et Y

Comme mentionné précédemment, les tests de symétrie de la loi conjointe et d'homogénéité des lois marginales sont des outils pour juger de l'accord entre deux variables mesurant la même caractéristique. Cependant, aucun de ces tests n'évalue ce qui se passe sur la diagonale du tableau $I \times I$ croisant les variables. Pourtant, s'il y a un accord parfait entre les deux variables, toutes les fréquences se retrouveront sur cette diagonale et les fréquences en dehors de la diagonale seront toutes nulles. En effet, si les deux variables sont parfaitement en accord, un individu se verra attribuer la même modalité pour les deux variables.

Lorsqu'on s'intéresse à l'accord entre deux variables, on doit donc tout d'abord mesurer cet accord. Les deux mesures les plus courantes de cet accord sont les suivantes :

Proportion d'accord observée :

Définition théorique :

$$P_0 = \sum_{i=1}^{I} \pi_{ii}$$

Estimateur :

$$\hat{P}_0 = \sum_{i=1}^{I} \hat{\pi}_{ii} = \sum_{i=1}^{I} n_{ii}/n$$

Interprétation : La valeur de cette mesure est entre 0 et 1. Évidemment, 0 représente une absence totale d'accord et 1 un accord parfait.

Kappa de Cohen :

Définition théorique :

$$\kappa = \frac{P_0 - P_e}{1 - P_e}$$

où $P_e = \sum_{i=1}^{I} \pi_{i\bullet}\pi_{\bullet i}$ est la proportion d'accord aléatoire
Estimateur :

$$\hat{\kappa} = \frac{\hat{P}_0 - \hat{P}_e}{1 - \hat{P}_e} \qquad \text{où} \qquad \hat{P}_e = \sum_{i=1}^{I} \frac{n_{i\bullet}}{n}\frac{n_{\bullet i}}{n}$$

Interprétation : La valeur maximale de cette mesure est 1. Elle représente un accord parfait. Fleiss *et al.* (2003, chapitre 18) suggèrent qu'une valeur de kappa plus grande ou égale à 0.75 représente un excellent accord, une valeur entre 0.4 et 0.75 un bon accord et une valeur inférieure à 0.4 un mauvais accord comparativement à un accord dû au hasard.

On peut facilement construire un intervalle de confiance du kappa. Aussi, il existe une version pondérée du kappa adaptée aux variables catégoriques ordinales. Fleiss *et al.* (2003, chapitre 18) est une référence très complète sur le sujet. Vous devez cependant savoir que cette mesure ne fait pas l'unanimité dans la communauté scientifique (Agresti, 2007, section 8.5.5). On lui reproche de trop dépendre des distributions marginales.

Pour des variables numériques, on utilise souvent une corrélation intra classes (ICC) pour mesurer l'accord. Cette mesure pourrait donc être utilisée pour des variables catégoriques ordinales représentées par un score numérique (voir Shoukri, 2010, pour plus d'information).

2.5.8 Interprétation des statistiques pour données pairées

Typiquement, on tente d'utiliser les statistiques présentées dans cette section pour juger de l'accord entre deux variables. Sachez que toutes ces mesures sont imparfaites. Il est important de bien observer les données et de ne pas seulement se fier aux valeurs des statistiques. Voici quelques exemples pour justifier cette mise en garde.

Exemple de différents degrés d'accord : données pairées fictives

Imaginons-nous les huit tableaux de fréquences 3×3 suivants. Nous supposons qu'il s'agit de données pairées dans tous les tableaux. On souhaite juger de l'accord entre les variables X et Y.

Cas 1 :

$X \setminus Y$	a	b	c	total
a	40	0	0	40
b	0	40	0	40
c	0	0	40	40
total	40	40	40	120

Cas 2 :

$X \setminus Y$	a	b	c	total
a	30	5	5	40
b	5	30	5	40
c	5	5	30	40
total	40	40	40	120

Cas 3 :

$X \setminus Y$	a	b	c	total
a	20	10	10	40
b	10	20	10	40
c	10	10	20	40
total	40	40	40	120

Cas 4 :

$X \setminus Y$	a	b	c	total
a	2	19	19	40
b	19	2	19	40
c	19	19	2	40
total	40	40	40	120

Cas 5 :

$X \setminus Y$	a	b	c	total
a	20	15	15	50
b	15	20	0	35
c	15	0	20	35
total	50	35	35	120

Cas 6 :

$X \setminus Y$	a	b	c	total
a	20	0	30	50
b	0	20	0	20
c	30	0	20	50
total	50	20	50	120

Cas 7 :

$X \setminus Y$	a	b	c	total
a	20	20	0	40
b	0	20	20	40
c	20	0	20	40
total	40	40	40	120

Cas 8 :

$X \setminus Y$	a	b	c	total
a	20	20	20	60
b	0	20	20	40
c	0	0	20	20
total	20	40	60	120

Pour ces 8 cas de figure, voici les résultats du test de symétrie de la loi conjointe (Bowker), du test d'homogénéité des marginales (Bhapkar), l'estimation de la proportion d'accord observée et du kappa de Cohen. Dans

la dernière colonne du tableau, on retrouve un jugement de l'accord que l'on peut établir en observant les données.

Cas	symétrie	homo. marges	proportion	kappa	jugement accord
1	oui	oui	1	1	parfait
2	non-rejet	non-rejet	0.75	0.625	bon
3	non-rejet	non-rejet	0.5	0.25	moyen
4	non-rejet	non-rejet	0.05	-0.425	mauvais
5	non-rejet	non-rejet	0.5	0.2381	comparé à 3 ?
6	non-rejet	non-rejet	0.5	0.2	comparé à 3 ?
7	rejet	non-rejet	0.5	0.25	comparé à 3 ?
8	rejet	rejet	0.5	0.3077	comparé à 3 ?

Les quatre premiers cas présentent tous une symétrie et des marges homogènes. Par contre, les fréquences sur la diagonale du tableau diminuent de plus en plus. Ces exemples illustrent clairement que la symétrie ou l'homogénéité des marginales n'implique pas l'accord entre les variables. Plus les fréquences sont faibles sur la diagonale, moins l'accord semble bon. Les mesures d'accord tiennent bien compte de ces fréquences sur la diagonale.

En outre, comment comparer les cas 3, 5, 6, 7 et 8 ? Ces cas ont tous les mêmes fréquences sur la diagonale. Est-ce donc dire qu'ils présentent des accords similaires ? Ce n'est pas tout à fait le cas puisque les fréquences en dehors de la diagonale ne sont pas réparties de la même façon. Dans le cas 3, on dira qu'elles sont réparties de façon aléatoire puisque les fréquences hors diagonales prennent des valeurs égales. Donc les désaccords ne semblent pas suivre de tendance. Pour les autres cas, les cases avec des fréquences nulles indiquent des désaccords jamais observés. Les kappa pour ces cas diffèrent légèrement. Par exemple, pour le cas 6, le kappa vaut 0.2 alors que pour le cas 8 il vaut 0.3077. Est-ce que le cas 8 présente vraiment un meilleur accord que le cas 6 ? J'en doute. On tombe maintenant dans des jugements subjectifs. Mais au fait, comment se définit exactement le concept d'accord entre deux variables ?

Remarquez que le cas 7 est un exemple d'homogénéité des marginales sans symétrie de la loi conjointe.

2.6 Résumé des formules concernant les tableaux de fréquences à deux variables

Définitions

Variable en lignes : X, indice : $i = 1, \ldots, I$, modalités : m_i^X (parfois variable explicative) ;

Variable en colonnes : Y, indice : $j = 1, \ldots, J$, modalités : m_j^Y (parfois variable réponse).

Tableau de fréquences :

		Y			
		m_1^Y	$\cdots$	m_J^Y	total
X	m_1^X $\vdots$ m_I^X	n_{ij}			$n_{i\bullet}$
	total	$n_{\bullet j}$			n

Types de fréquences :

Fréquences croisées : n_{ij} ;
Fréquences marginales : $n_{i\bullet}$ et $n_{\bullet j}$;
Fréquences conditionnelles : Une ligne ou une colonne de fréq. croisées ;
Fréquences relatives croisées : n_{ij}/n ;
Fréquences relatives marginales : $n_{i\bullet}/n$ et $n_{\bullet j}/n$;
Fréquences relatives conditionnelles :
fréquences de X conditionnelles à Y : $n_{ij}/n_{\bullet j}$;
fréquences de Y conditionnelles à X : $n_{ij}/n_{i\bullet}$.

Types d'échantillonnage :

multinomial vs Poisson :

– multinomial : la taille d'échantillon n est fixe, interprétation stat. :

$$(n_{ij}, i = 1, \ldots, I; j = 1, \ldots, J) \sim$$
$$Multinomiale(n, \pi_{ij}, i = 1, \ldots, I; j = 1, \ldots, J).$$

– Poisson : la taille d'échantillon n n'est pas fixe, interprétation stat. :

$$n_{ij} \sim Poisson(\lambda_{ij}) \qquad \text{indépendantes pour} \quad i = 1, \ldots, I \quad \text{et} \quad j = 1, \ldots, J$$

simple vs multiple :

– simple : un seul échantillon (comme ci-dessus)

– multiple : On forme des sous-populations (strates), à partir des modalités de la variable X ou de la variable Y, et on tire un échantillon dans chacune des sous-populations. Ces échantillons sont indépendants. Si l'échantillonnage est multinomial, on a l'interprétation statistique suivante :
Stratification par rapport à X : on a I sous-populations indépendantes telles que :

$$(n_{i1}, \ldots, n_{iJ}) \sim Multinomiale(n_i, \pi_{1|i}, \ldots, \pi_{J|i}) \quad \text{pour } i = 1, \ldots, I$$

où les n_i sont les $n_{i\bullet}$ vus auparavant, mais considérés fixes.
Stratification par rapport à Y : on a J sous-populations indépendantes telles que :

$$(n_{1j}, \ldots, n_{Ij}) \sim Multinomiale(n_j, \pi_{1|j}, \ldots, \pi_{I|j}) \quad \text{pour } j = 1, \ldots, J$$

où les n_j sont les $n_{\bullet j}$ vus auparavant, mais considérés fixes.

Probabilités d'intérêt et leurs estimateurs :

Type de probabilité	Probabilité	Définition	Estimateur potentiel	Bon estimateur si éch. multiple
conjointe	π_{ij}	$P(X = m_i^X, Y = m_j^Y)$	n_{ij}/n	jamais
marginale	$\pi_{i\bullet}$	$P(X = m_i^X)$	$n_{i\bullet}/n$	si var. stratif. $= Y$
	$\pi_{\bullet j}$	$P(Y = m_j^Y)$	$n_{\bullet j}/n$	si var. stratif. $= X$
conditionnelle	$\pi_{i\|j}$	$P(X = m_i^X \| Y = m_j^Y)$	$n_{ij}/n_{\bullet j}$	si var. stratif. $= Y$
	$\pi_{j\|i}$	$P(Y = m_j^Y \| X = m_i^X)$	$n_{ij}/n_{i\bullet}$	si var. stratif. $= X$

Test d'association entre deux variables nominales

Test d'indépendance :

$$H_0 \ : \ X \text{ et } Y \text{ sont indépendants} \quad \text{ou} \quad \pi_{ij} = \pi_{i\bullet}\pi_{\bullet j} \quad \forall\ i,\ j.$$

Test d'homogénéité de sous-populations (stratification selon X) :

$$\begin{aligned} H_0 \ : \ & \text{Dans les } I \text{ sous-populations déterminées par } X, \\ & Y \text{ suit la même distribution} \quad \text{ou} \\ & (\pi_{11}, \ldots, \pi_{1J}) = \cdots = (\pi_{I1}, \ldots, \pi_{IJ}) \quad \text{ou} \\ & \pi_{j|i} = \pi_{j|i'} \quad \forall\ i \neq i',\ j \quad \text{ou} \\ & \pi_{j|i} = \pi_{\bullet j} \quad \forall\ i,\ j \end{aligned}$$

Les hypothèses alternatives sont le complément des hypothèses nulles, ces tests sont toujours bilatéraux.

Ces deux sont tests équivalents car : indépendance $\Leftrightarrow$ homogénéité des sous-populations.

– **Statistique du khi-deux de Pearson** :

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - n_{i\bullet}n_{\bullet j}/n)^2}{n_{i\bullet}n_{\bullet j}/n} \quad \xrightarrow[H_0]{\text{asymp.}} \quad \chi^2_{(I-1)(J-1)}$$

– **Statistique du rapport de vraisemblance** :

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} \ln \frac{n_{ij}}{\hat{\mu}_{ij}} = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} \ln \frac{n_{ij}}{n_{i\bullet}n_{\bullet j}/n} \quad \xrightarrow[H_0]{\text{asymp.}} \quad \chi^2_{(I-1)(J-1)}$$

Cas particulier des tableaux 2×2 :

Test de comparaison de deux proportions :
Posons $\pi_1 = \pi_{1|i=1}$ et $\pi_2 = \pi_{1|i=2}$.

$$H_0 : \pi_1 = \pi_2 \qquad \text{versus} \qquad H_1 : \pi_1 \neq \text{ ou } > \text{ ou } < \pi_2$$

– **Statistique du test de Wald** :

$$Z_w = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\dfrac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_{1\bullet}} + \dfrac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_{2\bullet}}}} \quad \xrightarrow[H_0]{\text{asymp}} \quad \mathcal{N}(0,1).$$

– **Statistique du test score** :

$$Z_s = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_{1\bullet}} + \frac{1}{n_{2\bullet}}\right)}} \quad \xrightarrow[H_0]{\text{asymp.}} \quad \mathcal{N}(0,1)$$

où $\hat{\pi}_1 = n_{11}/n_{1\bullet}$, $\hat{\pi}_2 = n_{21}/n_{2\bullet}$ et $\hat{\pi} = \frac{n_{1\bullet}\hat{\pi}_1 + n_{2\bullet}\hat{\pi}_2}{n_{1\bullet}+n_{2\bullet}} = \frac{n_{11}+n_{21}}{n}$.

Forme abrégée de la statistique X^2 pour un tableau 2×2 :

$$X^2 = \frac{n\Delta^2}{n_{1\bullet}n_{2\bullet}n_{\bullet 1}n_{\bullet 2}} \qquad \text{où} \quad \Delta = n_{11}n_{22} - n_{12}n_{21}$$

Petits échantillons :

Correction pour la continuité de la statistique du khi-deux de Pearson pour un tableau 2×2 (correction de Yates) :

$$X^2_{corr} = \frac{n(|\Delta| - \frac{n}{2})^2}{n_{1\bullet}n_{2\bullet}n_{\bullet 1}n_{\bullet 2}} \qquad \xrightarrow[H_0]{\text{asymp.}} \qquad \chi^2_1$$

Test exact de Fisher pour un tableau 2×2 :
Test d'indépendance dont la statistique de test est :

$$n_{11} \quad \underset{H_0}{\longrightarrow} \quad Hypergeometrique(a = n_{\bullet 1}, b = n_{1\bullet}, c = n)$$

Cette distribution est exacte, mais elle suppose que les marges sont fixes.

Fonctions de masse de la distribution $Hypergeometrique(a, b, c)$:

$$P(X = x) = \frac{\binom{b}{x}\binom{c-b}{a-x}}{\binom{c}{a}} \qquad \text{pour} \quad max(0, a+b-c) \leq x \leq min(a, b).$$

Décrire et mesurer l'association entre deux variables nominales

- **Probabilités conditionnelles** : telles que définies précédemment.

- **Résidus** :
  - Résidus bruts : $RB_{ij} = n_{ij} - \hat{\mu}_{ij} = n_{ij} - n_{i\bullet}n_{\bullet j}/n$.
  - Résidus de Pearson :

$$RP_{ij} = \frac{n_{ij} - n_{i\bullet}n_{\bullet j}/n}{\sqrt{n_{i\bullet}n_{\bullet j}/n}} \qquad \xrightarrow[H_0]{\text{asymp.}} \qquad N(0, \sigma_{ij}^2)$$

  - Résidus de Pearson ajustés ou standardisés :

$$RAP_{ij} = \frac{n_{ij} - n_{i\bullet}n_{\bullet j}/n}{\sqrt{\frac{n_{i\bullet}n_{\bullet j}}{n}\left(1 - \frac{n_{i\bullet}}{n}\right)\left(1 - \frac{n_{\bullet j}}{n}\right)}} \qquad \xrightarrow[H_0]{\text{asymp.}} \qquad N(0,1)$$

- **Coefficient de Cramer** :

$$V = \sqrt{\frac{X^2/n}{\min\left(I-1, J-1\right)}}, \quad \text{pour un tableau } 2 \times 2 : V = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1\bullet}n_{2\bullet}n_{\bullet 1}n_{\bullet 2}}}$$

- **Différence de proportions** pour un tableau 2×2 :
  Définition théorique : $\pi_{1|i=1} - \pi_{1|i=2} = \pi_1 - \pi_2$.
  Estimateur : $\hat{\pi}_1 - \hat{\pi}_2 = n_{11}/n_{1\bullet} - n_{21}/n_{2\bullet}$.
  Intervalle de confiance de niveau $1 - \alpha$:

$$\pi_1 - \pi_2 \ \epsilon \ \hat{\pi}_1 - \hat{\pi}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_{1\bullet}} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_{2\bullet}}}.$$

- **Risque relatif** pour un tableau 2×2 :
  Définition théorique : $RR = \frac{\pi_1}{\pi_2}$.
  Estimateur : $\widehat{RR} = \frac{\hat{\pi}_1}{\hat{\pi}_2} = \frac{n_{11}/n_{1\bullet}}{n_{21}/n_{2\bullet}}$.
  Intervalle de confiance de niveau $1 - \alpha$:

$$\left[\widehat{RR}\ e^{-z_{\alpha/2}\hat{\sigma}(\ln(\widehat{RR}))},\ \widehat{RR}\ e^{z_{\alpha/2}\hat{\sigma}(\ln(\widehat{RR}))}\right] \text{ avec } \hat{\sigma}(\ln(\widehat{RR})) = \sqrt{\frac{1}{n_{11}} - \frac{1}{n_{1\bullet}} + \frac{1}{n_{21}} - \frac{1}{n_{2\bullet}}}.$$

– **Rapport de cotes** pour un tableau 2×2 :
Définition théorique : $RC = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\pi_{11}\pi_{22}}{\pi_{21}\pi_{12}}$.
Estimateur : $\widehat{RC} = \frac{\hat{\pi}_{11}\hat{\pi}_{22}}{\hat{\pi}_{21}\hat{\pi}_{12}} = \frac{n_{11}n_{22}}{n_{21}n_{12}}$.
Intervalle de confiance de niveau $1-\alpha$:

$$\left[\widehat{RC}\ e^{-z_{\alpha/2}\hat{\sigma}(\ln(\widehat{RC}))},\ \widehat{RC}\ e^{z_{\alpha/2}\hat{\sigma}(\ln(\widehat{RC}))}\right] \text{ avec } \hat{\sigma}(\ln(\widehat{RC})) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}.$$

Cas particulier des variables ordinales

Coefficient de corrélation :
mesure d'association linéaire (Pearson) ou monotone (Spearman)

$$r = \frac{\displaystyle\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} qm_i^X qm_j^Y - \frac{1}{n}\left(\sum_{i=1}^{I} n_{i\bullet} qm_i^X\right)\left(\sum_{j=1}^{J} n_{\bullet j} qm_j^Y\right)}{\sqrt{\left(\displaystyle\sum_{i=1}^{I} n_{i\bullet}(qm_i^X)^2 - \frac{1}{n}\left(\sum_{i=1}^{I} n_{i\bullet} qm_i^X\right)^2\right)\left(\sum_{j=1}^{J} n_{\bullet j}(qm_i^Y)^2 - \frac{1}{n}\left(\sum_{j=1}^{J} n_{\bullet j} qm_j^Y\right)^2\right)}}$$

où les qm_i^X et qm_j^Y sont des quantités représentant les modalités :
– pour le **coefficient de Pearson** (r_P), ces quantités sont des scores numériques subjectifs, mais choisis de façon à être les plus représentatifs possible de la réalité ;
– pour le **coefficient de Spearman**, ces quantités sont les rangs moyens des modalités de X et Y, que l'on peut calculer ainsi : pour tout $i = 1, \ldots, I$, le rang moyen de la modalité m_i^X de X est $\frac{a+b}{2}$, avec
$a =$ (nombre d'observations pour lesquelles $X < m_i^X$) $+ 1$;
$b =$ (nombre d'observations pour lesquelles $X \leq m_i^X$).
Le même raisonnement s'applique aux modalités de Y.

Test d'association entre X et Y

$$H_0 \ : X \text{ et } Y \text{ ne sont sont pas associées}$$
$$H_1 \ \begin{cases} X \text{ et } Y \text{ sont associées (test bilatéral)} \\ X \text{ et } Y \text{ sont associées positivement} \\ X \text{ et } Y \text{ sont associées négativement} \end{cases}$$

Statistique de test :

$$M = r\sqrt{(n-1)} \xrightarrow[H_0]{\text{asymp.}} N(0,1)$$

ou encore la statistique de Mantel et Haenszel (test bilatéral uniquement) :

$$M^2 = (n-1)r^2 \xrightarrow[H_0]{\text{asymp.}} \chi^2_1$$

où r est, au choix, la corrélation de Pearson (association linéaire) ou de Spearman (association monotone).

Cas particulier des données pairées

Sensibilité et spécificité d'un examen diagnostic :
X = Résultat du test diagnostique, soit m_1^X=positif (malade) ou m_2^X = négatif (sain)
Y = Vrai état d'une personne, soit m_1^Y = malade ou m_2^Y = sain

- **sensibilité** = $P(X = \text{positif} \mid Y = \text{malade})$,
- **spécificité** = $P(X = \text{négatif} \mid Y = \text{sain})$.

Test de la symétrie de la loi conjointe dans un tableau $I \times I$:

$$\begin{array}{lcll} H_0 & : & \pi_{ij} = \pi_{ji} & \text{pour tout couple } (i,j) \\ H_1 & : & \pi_{ij} \neq \pi_{ji} & \text{pour au moins un couple } (i,j) \end{array}$$

- **Statistique du khi-deux de Pearson (test de Bowker)** :

$$X^2_{sym} = \sum_{1 \leq i < j \leq I} \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}} \xrightarrow[H_0]{\text{asymp.}} \chi^2_{\frac{I(I-1)}{2}}$$

- **Statistique du rapport de vraisemblance** :

$$G^2_{sym} = \sum_{i=1}^{I} \sum_{j=1}^{I} n_{ij} \ln \frac{2n_{ij}}{n_{ij} + n_{ji}} \xrightarrow[H_0]{\text{asymp.}} \chi^2_{\frac{I(I-1)}{2}}$$

Test d'homogénéité des marginales dans un tableau $I \times I$:

$$\begin{array}{rcll} H_0 & : & \pi_{i\bullet} = \pi_{\bullet i} & \text{pour tout } i = 1, \ldots, I \\ H_1 & : & \pi_{i\bullet} \neq \pi_{\bullet i} & \text{pour au moins un } i \end{array}$$

On connaît l'existence des statistiques de Stuart-Maxwell et de Bhapkar et on sait comment les calculer en SAS, mais leurs formules n'ont pas été données en classe. Sous H_0, ces stat. suivent asymptotiquement une χ^2_{I-1}.

Test de McNemar : test de symétrie de la loi conjointe et d'homogénéité des marginales pour un tableau 2×2 :

$$X^2_{sym} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} \qquad \xrightarrow[H_0]{\text{asymp.}} \qquad \chi^2_1.$$

Mesures d'accord entre les variables X et Y :

- **Proportion d'accord observée** :
  Définition théorique : $P_0 = \sum_{i=1}^{I} \pi_{ii}$.
  Estimateur : $\hat{P}_0 = \sum_{i=1}^{I} \hat{\pi}_{ii} = \sum_{i=1}^{I} n_{ii}/n$.
- **Kappa de Cohen** :
  Définition théorique : $\kappa = \frac{P_0 - P_e}{1 - P_e}$
  où $P_e = \sum_{i=1}^{I} \pi_{i\bullet}\pi_{\bullet i}$ est la proportion d'accord aléatoire
  Estimateur :

$$\hat{\kappa} = \frac{\hat{P}_0 - \hat{P}_e}{1 - \hat{P}_e} \qquad \text{où} \qquad \hat{P}_e = \sum_{i=1}^{I} \frac{n_{i\bullet}}{n} \frac{n_{\bullet i}}{n}.$$

Chapitre 3

Tableaux de fréquences à trois variables

Dans ce chapitre, on s'intéresse encore au lien entre deux variables catégoriques X et Y. La nouveauté par rapport au chapitre 2 est que l'on va maintenant tenir compte d'une troisième variable catégorique, notée Z. Cette variable est peut-être liée à X et Y, mais on ne s'intéresse pas vraiment à ces liens. On cherche plutôt à étudier le lien entre X et Y en corrigeant pour l'effet confondant potentiel de Z sur ce lien.

Après avoir présenté quelques définitions et outils descriptifs à la section 1, la section 2 traitera de la distinction entre l'association conditionnelle et l'association marginale. Ensuite, nous présenterons quelques tests et mesures relatifs aux tableaux de fréquences $2 \times 2 \times K$. Pour finir, nous mentionnerons d'autres outils statistiques qui auraient pu être présentés dans ce chapitre.

Ce chapitre est bref, car l'objectif d'étudier la relation entre deux variables en corrigeant par rapport aux effets d'autres variables est mieux atteint en utilisant des modèles statistiques plutôt que des tests. Le dernier chapitre de ces notes présente des modèles statistiques.

3.1 Définitions et outils descriptifs

3.1.1 Tableaux conditionnels versus tableau marginal

Un tableau de fréquences à trois variables est en fait représenté par une série de tableaux de fréquences à deux variables, soit un tableau pour cha-

cune des modalités de la troisième variable, notée Z. Les modalités de cette variable seront notées m_k^Z avec $k = 1, \ldots, K$.

Sachant que $Z = m_k^Z$, on a :

		Y			
		m_1^Y	$\cdots$	m_J^Y	total
X	m_1^X $\vdots$ m_I^X		n_{ijk}		$n_{i\bullet k}$
	total		$n_{\bullet jk}$		$n_{\bullet\bullet k}$

et ce, pour $k = 1, \ldots, K$

Les K tableaux croisant X et Y en fixant la valeur de Z à une de ses modalités sont appelés « tableaux conditionnels » ou encore « tableaux partiels ». Dans ces tableaux, n_{ijk} représente le nombre d'individus dans l'échantillon pour lesquels $X = m_i^X$, $Y = m_j^Y$ et $Z = m_k^Z$. Le symbole $\bullet$ représente encore une sommation. Sa position dans l'indice indique par rapport à quelle variable on somme. La position 1 réfère à X, la deuxième position réfère à Y et la troisième à Z. Ainsi, par exemple, $n_{\bullet\bullet k} = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ijk}$.

On peut aussi produire un tableau de fréquences sommant sur toutes les modalités de Z comme suit :

		Y			
		m_1^Y	$\cdots$	m_J^Y	total
X	m_1^X $\vdots$ m_I^X		$n_{ij\bullet}$		$n_{i\bullet\bullet}$
	total		$n_{\bullet j\bullet}$		$n_{\bullet\bullet\bullet}$

Ce tableau est ici nommé « tableau marginal ». Il revient en fait à un simple tableau à deux variables tel que vu au chapitre 2, ne tenant pas compte d'une troisième variable.

Exemple de tableaux conditionnels et de tableau marginal : vols d'avions

Aux États-Unis, le Département de Transport demande aux compagnies aériennes de recueillir et de leur transmettre des données concernant les vols

d'avions qu'ils offrent. Parmi les informations ainsi recueillies, on retrouve, notamment, les variables suivantes :

- X : le nom de la compagnie aérienne ;
- Y : une indicatrice de départ en retard pour le vol ;
- Z : la ville de l'aéroport où à lieu le départ du vol.

On s'intéressera ici à ces trois variables, pour 11 000 vols d'avions en juin 1991 (Moore, 2003). Deux compagnies aériennes seront à l'étude : Alaska Airlines et America West Airlines. Les vols d'avions considérés partaient de 5 villes : Los Angeles, Phoenix, San Diego, San Francisco ou Seattle.

Question : Est-ce qu'il y a une différence entre les compagnies aériennes en ce qui concerne le respect des heures de départ prévues de leurs vols ?

On veut donc étudier le lien entre les variables X et Y. Cependant, on choisit de considérer une troisième variable, Z, potentiellement confondante puisqu'on la soupçonne d'être reliée à X et Y.

Voici les tableaux conditionnels de fréquences observées pour cet exemple :

Z	X	Y	
ville de départ	compagnie aérienne	retard oui	retard non
Los Angeles	Alaska	62	497
	AmWest	117	694
Phoenix	Alaska	12	221
	AmWest	415	4840
San Diego	Alaska	20	212
	AmWest	65	383
San Francisco	Alaska	102	503
	AmWest	129	320
Seattle	Alaska	305	1841
	AmWest	61	201

Il s'agit de 5 tableaux de fréquences croisées entre X et Y, conditionnels à la valeur de Z.

Souvent, lorsque les variables X et Y sont dichotomiques, ces tableaux partiels sont représentés plus succinctement de la façon suivante :

	Alaska Airlines		AmWest Airlines	
	en retard	à l'heure	en retard	à l'heure
Los Angeles	62	497	117	694
Phoenix	12	221	415	4840
San Diego	20	212	65	383
San Francisco	102	503	129	320
Seattle	305	1841	61	201

Si on ne tient pas compte de la ville de départ des vols, on obtient le tableau marginal de fréquences observées suivant :

X	Y	
compagnie	retard	
aérienne	oui	non
Alaska	501	3274
AmWest	787	6438

On observe bien ici que $n_{ij\bullet} = \sum_{k=1}^{5} n_{ijk}$ pour i=1,2 et j=1,2.

3.1.2 Graphiques

Il est difficile de représenter trois variables sur un même graphique. Il est souvent plus judicieux de produire des graphiques conditionnels, au même titre que les tableaux conditionnels. Il s'agit de graphiques présentant deux variables, X et Y. Un tableau est produit pour chacune des modalités de la troisième variable Z. La section 2.1.4 traitait de tels tableaux bivariés.

Le diagramme en mosaïque offre cependant une façon de représenter simultanément les trois variables sur un même graphique. On peut subdiviser les rectangles du graphique en fonction des fréquences relatives d'une troisième variable (et même d'une quatrième). La figure 3.1 présente un exemple d'un tel graphique. Cependant, ce genre de graphique présente tellement d'information qu'il est parfois difficile à interpréter.

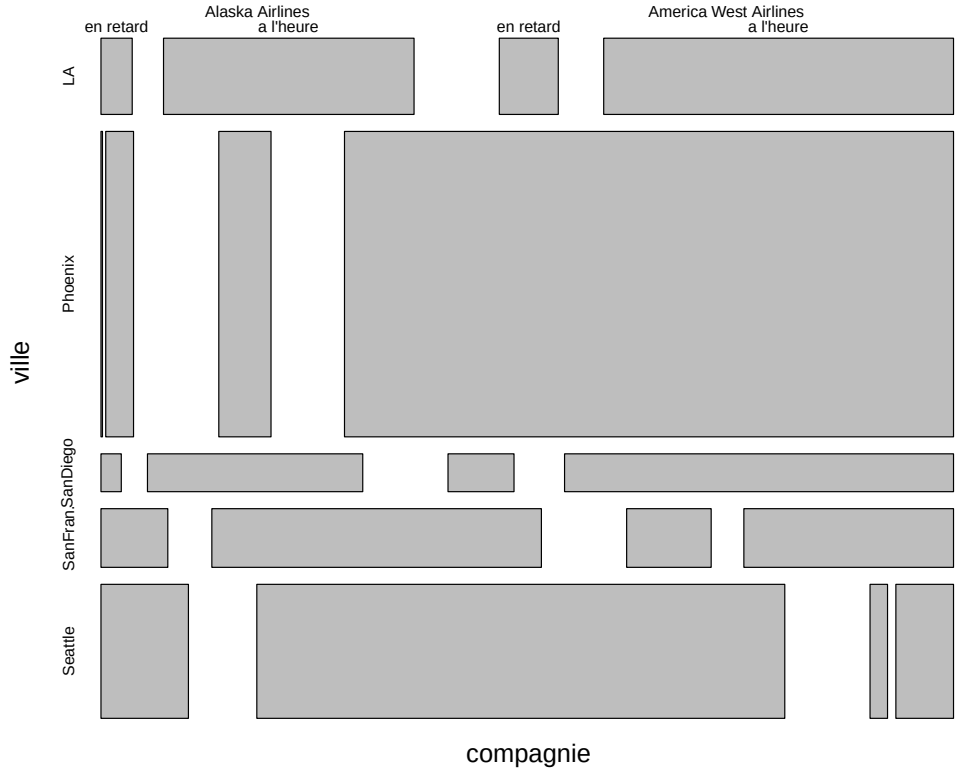


FIGURE 3.1 – Exemple de diagramme en mosaïque représentant trois variables catégoriques pour les données sur les vols d'avions, en tenant compte de la ville de départ.

3.2 Association conditionnelle versus association marginale

Lorsque l'on cherche à étudier la relation entre deux variables X et Y alors que l'on possède aussi des observations pour d'autres variables, on fait la distinction entre l'association entre X et Y conditionnelle à la valeur des autres variables, et l'association marginale entre X et Y, soit l'association ne tenant pas compte des autres variables.

Association conditionnelle : L'association conditionnelle entre les variables X et Y est étudiée par une analyse dans laquelle on fixe la valeur d'une autre ou même de plusieurs autres variables. On effectue plusieurs analyses conditionnelles du lien entre les variables X et Y puisque les variables de conditionnement peuvent prendre plusieurs valeurs. Dans le cas d'une seule variable de conditionnement catégorique Z, on est en présence de K associations conditionnelles, soit une pour chaque modalité de la variable. Ces associations peuvent être étudiées avec les tests et mesures présentés au chapitre 2.

Exemple d'association conditionnelle : vols d'avions

Reprenons l'exemple des vols d'avions. Pour chacune des villes de départ, nous allons étudier la relation entre la compagnie aérienne et les retards dans les vols. Pour ce faire, nous allons estimer la probabilité qu'un vol soit en retard conditionnellement à la ville et à la compagnie par la proportion de vols en retard dans l'échantillon pour chaque compagnie, dans chaque ville. Nous allons ensuite estimer, pour chacun des tableaux conditionnels, un rapport de cotes, lui aussi dit conditionnel et noté RC_k. Les estimations de ces rapports de cotes permettent de voir rapidement la direction de l'association entre X et Y. Si $\widehat{RC}_k < 1$, alors pour $Z = m_k^Z$ la proportion de vols en retard est plus grande pour $X = m_1^X =$ Alaska Airlines que pour $X = m_2^X =$ America West Airlines. C'est l'inverse si $\widehat{RC}_k > 1$. Puis, avec une statistique X^2 de Pearson, nous testerons si la probabilité qu'un vol soit en retard diffère entre les deux compagnies aériennes, et ce, pour chaque ville de départ.

Z	X	Y		Analyses conditionnelles				
ville de départ	cie aérienne	retard oui	retard non	prop. vols en retard	$\widehat{RC}_k$	X_k^2	seuil obs.	$n_{\bullet\bullet k}$
LA	Alaska	62	497	0.1109	0.7400	3.24	0.0718	1370
	AmWest	117	694	0.1443				
Phoenix	Alaska	12	221	0.0515	0.6333	2.35	0.1256	5488
	AmWest	415	4840	0.0790				
San Diego	Alaska	20	212	0.0862	0.5559	4.85	0.0277	680
	AmWest	65	383	0.1451				
San Fran.	Alaska	102	503	0.1686	0.5030	21.22	$< .0001$	1054
	AmWest	129	320	0.2873				
Seattle	Alaska	305	1841	0.1421	0.5459	14.90	0.0001	2408
	AmWest	61	201	0.2328				

Tous les rapports de cotes conditionnels sont inférieurs à 1, donc dans toutes les villes, la proportion de vols en retard est plus petite pour Alaska Airlines que pour America West Airlines. La différence entre ces deux proportions est significative pour trois villes : San Francisco, Seattle et San Diego. Elle est aussi presque significative à Los Angeles. La différence de proportion la moins marquée est observée pour les vols partant de Phoenix.

Ces analyses conditionnelles nous indiquent que le risque qu'un vol soit en retard est plus faible avec Alaska Airlines qu'avec America West Airlines.

Association marginale : L'association marginale entre les variables X et Y ne tient compte d'aucune autre variable. Pour des variables catégoriques, on étudie cette association à partir du tableau de fréquences bivariées entre X et Y (ici appelé tableau marginal), comme on a appris à le faire au chapitre 2.

Exemple d'association marginale : vols d'avions

Maintenant, nous ne tiendrons pas compte de la ville de départ des vols. Nous allons donc étudier marginalement la relation entre la compagnie aérienne et les retards dans les vols. Pour ce faire, nous allons calculer les mêmes statistiques que lors des analyses conditionnelles.

X	Y		Analyse marginale				
cie aérienne	retard oui	retard non	prop. vols en retard	$\widehat{RC}$	X^2	seuil obs.	n
Alaska	501	3274	0.1327	1.2518	13.57	0.0002	1100
AmWest	787	6438	0.1089				

Nous obtenons maintenant des résultats contraires à ce que nous obtenions en tenant compte de la ville de départ ! Ici, la proportion de vols en retard est plus petite pour America West Airlines que pour Alaska Airlines, et la différence entre les deux proportions est significativement différente de zéro.

3.2.1 Paradoxe de Simpson

Un paradoxe de Simpson est rencontré lorsque la direction de l'association entre deux variables X et Y est renversée lorsque l'on tient compte d'une troisième variable Z, alors dite « confondante ». En d'autres mots, on dit être en présence d'un paradoxe de Simpson lorsque les conclusions de chaque analyse conditionnelle sont opposées à la conclusion de l'analyse marginale. C'est le cas dans l'exemple des vols d'avions. Nous traitons ici de paradoxe de Simpson avec des variables catégoriques, mais sachez que ce paradoxe peut aussi être observé avec des variables numériques.

Comment expliquer un tel phénomène ? Ce paradoxe peut survenir lorsqu'il y a une forte association entre X et la troisième variable Z, ainsi qu'entre Y et la troisième variable Z.

Bonne pratique statistique à en tirer : Dans l'étude d'une relation entre deux variables X et Y, si on possède aussi des observations pour d'autres variables et que ces variables sont associées à la fois à X et à Y, il est important de faire des analyses conditionnelles à ces variables pour tirer des conclusions. Il ne faut pas se fier aux conclusions de l'analyse marginale, elles sont potentiellement trompeuses.

Exemple d'explication de paradoxe de Simpson : vols d'avions

Comment expliquer le paradoxe de Simpson observé dans l'exemple des vols d'avions ? Pour répondre à cette question, on doit étudier l'association entre la variable confondante Z et les variables X et Y.

Relation entre X et Z : Au total, parmi les 11 000 vols, 3775 sont offerts par Alaska Airlines et 7225 par America West Airlines. La majorité des vols $((1841 + 305) \times 100\%/3775 = 56.8\%)$ d'Alaska Airlines partent de Seattle, et la majorité des vols d'America West Airlines partent de Phoenix $((4840 + 415) \times 100\%/7225 = 72.7\%)$.

Relation entre Y et Z : Parmi les 11 000 vols étudiés, sans considérer la compagnie aérienne, voici les proportions de retard selon la ville de départ :

ville de départ	proportion de vols en retard
Los Angeles	0.1307
Phoenix	0.0778
San Diego	0.1250
San Francisco	0.2192
Seattle	0.1520

On voit que Phoenix se distingue des autres villes par une proportion plus faible de vols en retard.

Ainsi, le paradoxe de Simpson observé ici vient du fait que America West Airlines offre beaucoup de vols en partance de Phoenix et que dans cette ville la proportion de vols en retard est plus petite qu'ailleurs. Si on ne tient pas compte de la ville, les nombreux vols à l'heure d'America West Airlines en partance de Phoenix viennent gonfler leur nombre total de vols à l'heure.

Si on fait une analyse marginale, mais en omettant les vols en partance de Phoenix, on n'observe plus de paradoxe de Simpson. L'analyse marginale et les analyses conditionnelles donnent alors des résultats similaires.

3.2.2 Indépendance conditionnelle versus marginale

L'indépendance conditionnelle et l'indépendance marginale sont deux notions distinctes. L'une n'implique pas l'autre. Pour s'en convaincre, voici deux contre-exemples fictifs.

Contre-exemple prouvant que
indépendance conditionnelle $\not\Rightarrow$ indépendance marginale :

Voici des données fictives présentant de l'indépendance conditionnelle, mais pas de l'indépendance marginale.

Z	X	Y m_1^Y	m_2^Y	$\widehat{RC}$
m_1^Z	m_1^X	18	12	$\frac{18 \times 8}{12 \times 12} = 1$
	m_2^X	12	8	
m_2^Z	m_1^X	2	8	$\frac{2 \times 32}{8 \times 8} = 1$
	m_2^X	8	32	
Total	m_1^X	20	20	$\frac{20 \times 40}{20 \times 20} = 2$
	m_2^X	20	40	

Pour juger de l'indépendance ou non entre les variables X et Y, nous avons calculé des rapports de cotes. Comme on l'a appris au chapitre 2, un rapport de cotes prenant la valeur 1 est un indicateur d'indépendance.

Contre-exemple prouvant que
indépendance marginale $\not\Rightarrow$ indépendance conditionnelle :

Voici encore des données fictives, présentant maintenant de l'indépendance marginale, mais pas de l'indépendance conditionnelle.

Z	X	Y m_1^Y	m_2^Y	$\widehat{RC}$
m_1^Z	m_1^X	8	12	$\frac{8 \times 8}{12 \times 12} = 0.\bar{4}$
	m_2^X	12	8	
m_2^Z	m_1^X	12	8	$\frac{12 \times 12}{8 \times 8} = 2.25$
	m_2^X	8	12	
Total	m_1^X	20	20	$\frac{20 \times 20}{20 \times 20} = 1$
	m_2^X	20	20	

3.3 Cas particulier des tableaux $2 \times 2 \times K$: homogénéité de l'association conditionnelle

Lorsque la variable de conditionnement Z a plusieurs modalités, il peut être fastidieux de présenter les résultats de toutes les analyses conditionnelles. D'ailleurs, si toutes ces analyses conditionnelles donnent des résultats similaires, on aimerait pouvoir les regrouper en une seule analyse. Il existe des tests d'homogénéité d'association pour vérifier si, statistiquement, des associations conditionnelles sont équivalentes (donc homogènes). Si c'est le cas, on peut alors produire des tests et des mesures d'association conditionnelle communs.

Dans cette section, nous traitons de tels outils statistiques dans le cas particulier de tableaux $2 \times 2 \times K$, c'est-à-dire lorsque les variables X et Y sont dichotomiques.

3.3.1 Test d'homogénéité de l'association conditionnelle

Pour deux variables dichotomiques, une mesure usuelle d'association est le rapport de cotes. On pourrait donc tester l'homogénéité de l'association entre X et Y conditionnellement à Z en testant l'égalité des K rapports de cotes, notés RC_k pour $k = 1, \ldots, K$, provenant des K analyses conditionnelles. Les hypothèses de ce test seraient :

$$\begin{aligned} H_0 &: RC_1 = \ldots = RC_K, \\ H_1 &: RC_k \neq RC_{k'} \text{ pour au moins un couple } (k, k'). \end{aligned}$$

Nous ne présenterons pas ici les détails mathématiques d'un tel test, mais nous mentionnerons qu'il en existe, notamment le test asymptotique de Breslow et Day (1980) et le test exact de Zelen (1971). La statistique du test de Breslow-Day suit asymptotiquement, sous H_0, une loi khi-deux à $K - 1$ degrés de liberté. Elle peut être calculée à l'aide de certains logiciels statistiques.

Nous verrons dans le prochain chapitre comment faire un test équivalent en régression logistique.

Exemple de test d'homogénéité des rapports de cotes : vols d'avions

Dans l'exemple des vols d'avions, nous avions calculé des rapports de cotes conditionnels, soit un pour chacun des tableaux conditionnels. Ces rapports de cotes prenaient des valeurs relativement égales. L'association entre X et Y semble donc homogène par rapport à la variable de conditionnement Z. À l'aide du logiciel SAS, nous avons effectué le test de Breslow-Day. Nous avons obtenu une valeur observée pour la statistique de test de 3.2701. Sous l'hypothèse nulle d'homogénéité des rapports de cotes, la statistique de ce test suit asymptotiquement une loi khi-deux à $K - 1 = 4$ degrés de liberté. Ici, le seuil observé du test, soit $P(\chi^2_4 \geq 3.2701) = 0.5137$, est plus grand que le niveau de signification de 5%. Ainsi, nous ne pouvons pas rejeter l'hypothèse nulle d'homogénéité des rapports de cotes. Nous concluons donc que l'association conditionnelle entre X et Y est homogène.

3.3.2 Mesure commune d'association conditionnelle

Si les rapports de cotes sont homogènes, c'est qu'ils ont une valeur commune. Mantel et Haenszel ont proposé l'estimateur suivant de cette valeur commune :

$$\widehat{RC}_{MH} = \frac{\sum_{k=1}^{K}(n_{11k}n_{22k}/n_{\bullet\bullet k})}{\sum_{k=1}^{K}(n_{12k}n_{21k}/n_{\bullet\bullet k})}.$$

Exemple de rapport de cotes commun : vols d'avions

La valeur observée de l'estimateur de Mantel et Haenszel du rapport de cotes commun est ici :

$$\widehat{RC}_{MH} = \frac{62 \times 694/1370 + \cdots + 305 \times 201/2408}{497 \times 117/1370 + \cdots + 1841 \times 61/2408} = 0.5846$$

Remarque : Mantel et Haenszel ont aussi proposé un estimateur du risque relatif commun. Cependant, l'homogénéité des rapports de cotes n'implique pas l'homogénéité des risques relatifs. Ainsi, même si l'homogénéité des rapports de cotes a été prouvée par un test, il est préférable de calculer les risques

relatifs conditionnels et de s'assurer qu'ils sont similaires avant de calculer un risque relatif commun.

3.3.3 Test d'indépendance conditionnelle

Si les rapports de cotes sont homogènes, on peut aussi tester l'indépendance conditionnelle. Cette indépendance est définie par l'indépendance entre X et Y peu importe la valeur de Z. Pour des variables dichotomiques, on peut traduire l'indépendance conditionnelle en terme de rapport de cotes. En effet, indépendance conditionnelle $\Leftrightarrow RC_1 = \ldots = RC_K = 1$. Le test de Cochran-Mantel-Haenszel est un test bilatéral d'hypothèse nulle H_0 : X et Y sont conditionnellement indépendantes. La statistique de ce test suit asymptotiquement, sous H_0 et en supposant que les marges des tableaux conditionnels sont fixes, une loi khi-deux à 1 degrés de liberté. Comme pour le test de Breslow-Day d'homogénéité des rapports de cotes, nous ne donnerons pas ici la formule pour calculer cette statistique de test (Agresti, 2007, voir section 4.4.3). Elle peut être calculée avec certains logiciels statistiques.

Exemple de test d'homogénéité des rapports de cotes : vols d'avions

À l'aide du logiciel SAS, nous avons effectué le test de Cochran-Mantel-Haenszel. Nous avons obtenu une valeur observée pour la statistique de test de 42.0019. Sous l'hypothèse nulle d'indépendance conditionnelle et celle que les marges des tableaux conditionnels sons fixes, la statistique de ce test suit asymptotiquement une loi khi-deux à 1 degrés de liberté. Ici, le seuil observé du test, soit $P(\chi^2_1 \geq 42.0019) < 0.0001$. Ainsi, nous rejetons avec conviction l'hypothèse nulle d'indépendance conditionnelle. Nous concluons donc qu'il existe bel et bien une association conditionnelle entre X et Y.

Remarque : Le test de Cochran-Mantel-Haenszel se généralise au cas où les variables X et Y possèdent plus de 2 modalités, et même au cas où une ou les deux variables sont ordinales (Landis *et al.*, 1978). Cependant, le test de Breslow-Day, lui, n'a pas été généralisé à ces cas. On est donc à court de ressources pour tester l'homogénéité de l'association dans un cas général. Il est toujours préférable d'observer nos données afin d'évaluer si les associations conditionnelles semblent homogènes avant d'effectuer un test de Cochran-Mantel-Haenszel.

3.4 Résumé des formules concernant les tableaux de fréquences à trois variables

– Notation pour un tableau de fréquences $I \times J \times K$:

Un tableau à 3 variables permet d'étudier la relation entre X et Y en tenant compte d'une troisième variable Z. On note les modalités des variables comme suit :
– modalités de X : m_i^X pour $i = 1, \ldots, I$;
– modalités de Y : m_j^Y pour $j = 1, \ldots, J$;
– modalités de Z : m_k^Z pour $k = 1, \ldots, K$;

Un tableau à 3 variables se compose de K **tableaux conditionnels** (ou partiels) $I \times J$:

Sachant que $Z = m_k^Z$, on a :

		Y			
		m_1^Y	$\cdots$	m_J^Y	total
X	m_1^X $\vdots$ m_I^X	n_{ijk}			$n_{i\bullet k}$
	total	$n_{\bullet jk}$			$n_{\bullet\bullet k}$

et ce, pour $k = 1, \ldots, K$

Le **tableau marginal** $I \times J$ est obtenu en sommant sur Z :

		Y			
		m_1^Y	$\cdots$	m_J^Y	total
X	m_1^X $\vdots$ m_I^X	$n_{ij\bullet}$			$n_{i\bullet\bullet}$
	total	$n_{\bullet j\bullet}$			$n_{\bullet\bullet\bullet}$

Dans ces tableaux, n_{ijk} représente le nombre d'individus dans l'échantillon aléatoire pour lesquels $X = m_i^X$, $Y = m_j^Y$ et $Z = m_k^Z$. Le symbole $\bullet$ représente une sommation.

– **Association conditionnelle** = association entre les variables X et Y en fixant la valeur de Z.
– **Association marginale** = association entre les variables X et Y sans tenir compte de Z.

– **Paradoxe de Simpson** = la direction de l'association entre deux variables X et Y est renversée lorsque l'on tient compte d'une troisième variable Z. En d'autres mots, les conclusions de chaque analyse conditionnelle sont opposées à la conclusion de l'analyse marginale.

– Statistique du **test d'homogénéité de l'association** entre X et Y dans un tableau $2 \times 2 \times K$:
H_0 : les rapports de cotes des K tableaux conditionnels sont égaux

On connaît l'existence de la statistique de Breslow-Day et on sait comment la calculer en SAS, mais sa formule n'a pas été donnée en classe. Sous H_0, cette statistique suit asymptotiquement une χ^2_{K-1}.

– Estimation de Mantel et Haenszel du **rapport de cotes commun** dans un tableau $2 \times 2 \times K$:

$$\widehat{RC}_{MH} = \frac{\sum_{k=1}^{K}(n_{11k}n_{22k}/n_{\bullet\bullet k})}{\sum_{k=1}^{K}(n_{12k}n_{21k}/n_{\bullet\bullet k})}.$$

– Statistique du **test d'indépendance conditionnelle** (indépendance entre X et Y peu importe la valeur de Z) dans un tableau $2 \times 2 \times K$: si les rapports de cotes sont homogènes, on peut tester H_0 : X et Y sont conditionnellement indépendant.

On connaît l'existence de la statistique de Cochran, Mantel et Haenszel et on sait comment la calculer en SAS, mais sa formule n'a pas été donnée en classe. Sous H_0, et sous l'hypothèse que les marges sont fixes, cette statistique suit asymptotiquement une χ^2_1.

Chapitre 4

Modèles linéaires généralisés (GLM)

Les modèles linéaires généralisés (McCullagh et Nelder, 1989) sont souvent désignés par leur acronyme : GLM (Generalized Linear Models). Cet acronyme sert aussi à désigner le modèle linéaire général (General Linear Model). Par exemple, la procédure `GLM` de SAS permet d'ajuster un modèle linéaire général. Le mot général signifie ici qu'il s'agit d'un modèle linéaire pouvant contenir plus d'une variable réponse. La fonction `glm()` de R ajuste quant à elle des modèles linéaires généralisés. Ce n'est pas la même chose. Le présent document traite de ces derniers. Ils généralisent les modèles linéaires de deux façons :

- La variable réponse du modèle peut suivre une distribution autre que normale.
- L'espérance de la variable aléatoire peut être reliée aux variables explicatives par un lien autre que l'identité.

Ainsi, les GLM sont une solution de rechange aux modèles linéaires ajustés sur une variable réponse transformée en raison de résidus non normaux ou présentant une variance hétérogène.

Dans ce chapitre, on verra comment utiliser les GLM pour étudier le lien entre une variable réponse catégorique ou représentant un dénombrement et des variables explicatives de type quelconque (Agresti, 2007). Quelques avantages de ces modèles sur l'utilisation de tests et mesures pour les tableaux de fréquences vus dans la première partie du cours sont que :

- les variables explicatives peuvent être numériques continues ;

– ils intègrent facilement plusieurs variables explicatives, ce qui permet d'étudier des liens entre les variables explicatives et la variable réponse en corrigeant pour les effets de toutes les autres variables explicatives ;
– ils offrent la possibilité de faire des prédictions.

Les tests et mesures pour les tableaux de fréquences peuvent faire partie d'une étude exploratoire préliminaire à l'ajustement d'un GLM.

Changements dans la notation employée

À partir d'ici, la notation change légèrement afin d'être conforme à la littérature sur les GLM. La lettre i était jusqu'à maintenant utilisée pour indicer les modalités de la variable catégorique X. Cette lettre va maintenant plutôt servir à représenter les individus de l'échantillon, avec i allant de 1 à n. Ce n'est donc plus l'indice u qui représente les individus à partir d'ici, mais bien i comme on a l'habitude de voir dans la présentation de modèles statistiques.

On utilise la lettre Y pour représenter la variable réponse du modèle. On postule qu'il s'agit d'une variable aléatoire, ce qui explique l'utilisation d'une lettre majuscule pour représenter cette variable.

La lettre x désignera quant à elle une variable explicative du modèle. On utilise une lettre minuscule, car on considère que cette quantité est fixe et non aléatoire. Un GLM peut contenir une seule variable explicative (modèle simple) ou plusieurs (modèle multiple). On indicera par j les différentes variables explicatives s'il y en a plus qu'une, avec j allant de 1 à p. Ainsi, p représente le nombre total de variables explicatives dans le modèle.

4.1 Composantes d'un GLM et notation

Tous les modèles linéaires généralisés ont trois composantes :

1. composante aléatoire ;
2. composante systématique ;
3. fonction de lien.

1. Composante aléatoire

Il s'agit de la variable réponse, notée Y. On suppose que les autres variables, dites explicatives, influencent potentiellement la valeur de cette variable, considérée aléatoire.

Hypothèses du modèle :
- on possède des observations indépendantes de la variable Y,
- Y suit une distribution de la famille exponentielle, à identifier (exemple de distribution de la famille exponentielle : normale, binomiale (cas particulier Bernoulli), multinomiale, Poisson, binomiale négative, exponentielle, gamma, etc.).

Un modèle linéaire généralisé ne possède qu'une seule variable réponse. On le qualifie donc de modèle univarié.

2. Composante systématique

Il s'agit d'une combinaison linéaire des variables explicatives, notées x_1 à x_p :

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \boldsymbol{x}^t \boldsymbol{\beta}$$

où $\boldsymbol{x}^t = (1, x_1, \ldots, x_p)$ et $\boldsymbol{\beta}^t = (\beta_0, \beta_1, \ldots, \beta_p)$. Cette quantité est parfois appelée prédicteur linéaire et notée η.

Coefficients à estimer :
- ordonnée à l'origine β_0 (pas obligatoire) et
- coefficients β_j, $j = 1, \ldots, p$, pour les variables explicatives.

Caractéristiques de la composante systématique :
- linéaire ;
- peut contenir une seule (modèle simple) ou plusieurs (modèle multiple) variables explicatives ;
- peut contenir des interactions entre les variables explicatives ;
- les variables explicatives peuvent être numériques ou catégoriques (voir section 4.2.1) ;
- les variables explicatives sont considérées fixes, donc le modèle ne comporte pas d'effets aléatoires (par contre les GLMM, modèles linéaires généralisés mixtes, peuvent en contenir).

3. Fonction de lien

Il s'agit du lien entre les composantes aléatoire et systématique.

Soit $E(Y) = \mu$ la valeur espérée ou moyenne de Y, la fonction de lien $g()$ relie μ à la composante systématique ainsi :

$$g(\mu) = \boldsymbol{x}^t \boldsymbol{\beta}.$$

La fonction $g()$ doit être monotone et différenciable.

Souvent, la fonction de lien est choisie de façon à s'assurer que les valeurs prédites de μ appartiennent à l'ensemble de ses valeurs possibles.

Énoncé du modèle

Nous considérons ici un échantillon aléatoire de n observations $Y_1, \ldots, Y_n$ et un ensemble de p variables explicatives mesurées pour chacune des observations. Pour l'observation i, on note $\boldsymbol{x}_i^t = (1, x_{i1}, \ldots, x_{ip})$ le vecteur des valeurs observées des variables explicatives. Le chiffre 1 en première position de ce vecteur sert à intégrer une ordonnée à l'origine au modèle.

Le modèle linéaire généralisé peut s'énoncer ainsi :

$$Y_i \sim \mathcal{L}(\mu_i, \phi) \text{ indépendantes, avec } g(\mu_i) = \boldsymbol{x}_i^t \boldsymbol{\beta}, \text{ pour } i = 1, \ldots, n$$

où

- Y_i est la variable aléatoire Y sachant que $\boldsymbol{x} = \boldsymbol{x}_i$;
- $\mathcal{L}$ est une distribution à spécifier de la famille exponentielle ;
- ϕ est un paramètre de dispersion de la distribution choisie (pour certaines distributions, telles que Bernoulli et Poisson, $\phi = 1$) ;
- $\mu_i = E(Y_i)$ est la valeur espérée ou moyenne de Y_i.

Exemple : régression linéaire

Un modèle de régression linéaire est souvent énoncé comme suit :

$$Y_i = \boldsymbol{x}_i^t \boldsymbol{\beta} + \epsilon_i \quad \text{où} \quad \epsilon_i \; iid \; N(0, \sigma^2), \quad \text{pour } i = 1, \ldots, n.$$

Ce modèle est équivalent au GLM suivant :

$Y_i \sim N(\mu_i, \sigma^2)$ indépendantes, avec $g(\mu_i) = \mu_i = \boldsymbol{x}_i^t \boldsymbol{\beta}$, pour $i = 1, \ldots, n$.

Dans ce GLM, on a que
- la fonction de lien $g()$ est l'identité ;
- $\mathcal{L}$ est la distribution normale ;
- σ^2 est le paramètre de dispersion de la distribution normale ($\phi = \sigma^2$).

4.1.1 Régression logistique : composantes du modèle

La régression logistique (Hosmer et Lemeshow, 2000) de base sert à modéliser une variable réponse binaire en fonction de variables explicatives. Cette méthode peut aussi être généralisée aux variables réponses à plus de deux modalités. Nous discuterons de cette généralisation plus loin. Pour l'instant, considérons le cas d'une variable réponse à deux modalités.

Soit Y la variable réponse binaire. Attribuons la valeur 1 à une de ses modalités et 0 à l'autre. On suppose donc que Y suit une distribution Bernoulli de paramètre π avec $\pi = P(Y = 1) = E(Y) = \mu$. Le modèle s'énonce ainsi :

$$\begin{aligned} Y_i \sim Bernoulli(\mu_i = \pi_i) \quad &\text{indépendantes} \\ \text{avec} \quad g(\pi_i) = \boldsymbol{x}_i^t \boldsymbol{\beta}&, \text{ pour } i = 1, \ldots, n. \end{aligned}$$

Plusieurs fonctions de lien $g()$ sont intéressantes pour ce modèle :

$$\begin{aligned} g(\pi_i) &= \ln\left(\frac{\pi_i}{1-\pi_i}\right) \text{ lien logit;} \\ g(\pi_i) &= \Phi^{-1}(\pi_i) \text{ lien probit, où } \Phi(x) = P(\mathcal{N}(0,1) \leq x); \\ g(\pi_i) &= \ln(-\ln(1-\pi_i)) \text{ lien log-log ou Gumbel;} \\ g(\pi_i) &= \pi_i \text{ lien identité.} \end{aligned}$$

Les trois premières fonctions de lien énumérées ci-dessus permettent de s'assurer que les valeurs prédites sont dans l'intervalle $(0, 1)$.

Dans la très vaste majorité des cas, on utilise le lien logit, car contrairement aux autres liens, il permet aux paramètres du modèle d'être interprétable. Notez que le choix de la fonction de lien ne change pas grand-chose aux conclusions finales.

Les 3 hypothèses principales du modèle de régression logistique sont donc :
1) les observations Y_i sont indépendantes ;
2) les observations Y_i suivent une loi de Bernoulli ;
3) la probabilité de succès $\pi = P(Y = 1)$ est monotone en x, pour toutes les variables x du modèle.

4.1.2 Régression Poisson : composantes du modèle

La régression Poisson (Cameron et Trivedi, 1998) sert à modéliser une variable réponse représentant un dénombrement. On suppose que Y suit une loi Poisson de paramètre $\mu = E(Y)$. Le modèle s'énonce ainsi :

$$\begin{aligned} Y_i &\sim Poisson(\mu_i) \quad \text{indépendantes} \\ \text{avec} \quad \ln(\pi_i) &= \boldsymbol{x}_i^t \boldsymbol{\beta}, \text{ pour } i = 1, \ldots, n. \end{aligned}$$

Le logarithme naturel (parfois noté seulement log), est la fonction de lien la plus utilisée en régression Poisson. Certains utilisent aussi parfois une fonction de lien identité ou encore racine carrée.

Les 3 hypothèses principales du modèle de régression Poisson sont similaires à celles pour la régression logistique :
1) les observations Y_i sont indépendantes ;
2) les observations Y_i suivent une loi de Poisson ;
3) l'espérance μ est monotone en x, pour toutes les variables x du modèle.

4.1.3 Comparaison de différents GLM

Voici un tableau comparatif de différents modèles usuels faisant partie de la grande famille des modèles linéaires généralisés :

Composante aléatoire	**Composante systématique**	**Fonction de lien**	**Type de modèle**
normale	var. numériques	identité	régression linéaire
normale	var. catégoriques	identité	ANOVA
normale	var quelconque	identité	ANCOVA
binomiale	var quelconque	logit*	régression logistique
Poisson	var quelconque	ln*	régression Poisson

*fonction de lien la plus usuelle, mais elle peut être autre

4.2 Interprétation des paramètres

De façon générale, si un paramètre β_j prend une valeur nulle, c'est que la variable explicative dont il est le coefficient dans le modèle n'a pas de lien avec la variable réponse. Cependant, si la valeur d'un paramètre est non nulle, la signification que l'on peut lui attribuer dépend du type de la variable explicative dont il est le coefficient ainsi que de la fonction de lien du modèle.

4.2.1 Variables explicatives catégoriques

Pour que le modèle $g(\mu) = \boldsymbol{x}^t\boldsymbol{\beta}$ ait du sens, il faut que les variables explicatives x_1 à x_p prennent des valeurs numériques. Alors, comment intégrer à un GLM une variable explicative catégorique nominale ? Il faut utiliser une « paramétrisation » pour ramener la variable sur une échelle numérique. Si la variable est catégorique ordinale, c'est facile. On va représenter la variable par un score numérique et utiliser ce score comme variable explicative dans la composante systématique du modèle. Cependant, si les modalités de la variable ne sont pas ordonnables, on utilise typiquement des variables indicatrices.

Le cas le plus simple est celui d'une variable catégorique à 2 modalités. On attribue alors la valeur 1 à une modalité et 0 à l'autre. Si la variable catégorique nominale a k modalités possibles, on va typiquement sélectionner une des modalités comme étant la modalité de référence. Pour toutes les autres modalités, on crée une indicatrice prenant la valeur 1 si la variable catégorique est égale à la modalité en question, 0 sinon. On ajoute ainsi $k-1$ variables indicatrices dans le GLM pour représenter une seule variable explicative à k modalités.

Exemple de paramétrisation d'une variable catégorique :
fautes de frappe

Voici un petit jeu de données fictif qui sera utilisé dans ce chapitre pour illustrer quelques notions théoriques. Les variables présentent dans ces données sont Y le nombre de fautes de frappe contenu dans un texte, X_1 le nombre de mots dans le texte et X_2 l'auteur du texte.

Y	$x_1 =$ nbre mots	$x_2 =$ auteur	I_A	I_B	I_C
1	102	A	1	0	0
4	242	B	0	1	0
1	97	A	1	0	0
7	299	C	0	0	1
3	210	B	0	1	0
1	107	A	1	0	0
3	198	B	0	1	0
5	261	C	0	0	1
1	100	A	1	0	0
8	310	C	0	0	1

L'auteur est une variable nominale. On ne peut donc pas définir la composante systématique du modèle par $\beta_0 + \beta_1 x_1 + \beta_2 x_2$. En effet, x_2 n'est pas un nombre. Multiplier l'auteur A par le coefficient numérique β_2 n'a pas de sens. On va plutôt utiliser des variables indicatrices pour incorporer l'auteur dans le modèle. Une indicatrice pour chacun des auteurs $(I_\text{A}, I_\text{B}, I_\text{C})$ a été juxtaposée au jeu de données ci-dessus. On voit que, par définition, la somme des trois indicatrices vaut toujours 1 ($I_{\text{A}i} + I_{\text{B}i} + I_{\text{C}i} = 1 \; \forall \; i$). Ainsi, inclure ces trois indicatrices dans le modèle causerait une multicolinéarité parfaite : on peut toujours définir une des trois indicatrices par une combinaison linéaire des deux autres. On doit donc choisir seulement deux de ces indicatrices pour les inclure dans le modèle. L'indicatrice omise est celle de la modalité de référence. Par exemple, on pourrait choisir l'auteur C comme modalité de référence et définir la composante systématique du modèle comme suit : $\beta_0 + \beta_1 x_1 + \beta_2 I_\text{A} + \beta_3 I_\text{B}$.

On appelle parfois « variables de design » les variables incluses dans la composante systématique d'un modèle pour représenter une variable explicative catégorique. Pour affirmer qu'une variable explicative catégorique n'a pas de lien avec la variable réponse, il faut que les paramètres multipliant toutes les variables de design représentant cette variable dans le modèle soient nuls.

D'autres paramétrisations que celles utilisant des indicatrices sont possibles pour inclure une variable catégorique dans la composante systématique d'un modèle. Cependant, elles ne seront pas approfondies dans ce cours. L'aide du logiciel SAS en présente quelques-unes.

4.2.2 Lien identité : effet additif

Avec un lien linéaire, on a $\mu = \boldsymbol{x}^t\boldsymbol{\beta}$. Notons

$$\mu(x_j) = \beta_0 + \beta_1 x_1 + \cdots \beta_j x_j + \cdots + \beta_p x_p = \boldsymbol{x}^t\boldsymbol{\beta}$$

et

$$\mu(x_j + 1) = \beta_0 + \beta_1 x_1 + \cdots \beta_j (x_j + 1) + \cdots + \beta_p x_p = \boldsymbol{x}^t\boldsymbol{\beta} + \beta_j.$$

On a donc que

$$\beta_j = \mu(x_j + 1) - \mu(x_j).$$

Ainsi, lorsque l'on passe de x_j à $x_j + 1$ alors que toutes les autres variables explicatives restent inchangées, la quantité β_j est ajoutée (effet additif) à la moyenne de la variable réponse Y. Il s'agit d'une augmentation si la valeur de β_j est positive et d'une diminution si la valeur de β_j est négative.

Pour une variable x_j indicatrice, β_j représente la différence de moyennes entre le groupe d'individus pour lesquels $x_j = 1$ comparativement à ceux pour lesquels $x_j = 0$:

$$\beta_j = \mu(x_j = 1) - \mu(x_j = 0).$$

4.2.3 Lien logarithmique : effet multiplicatif

Avec un lien logarithmique, on a $\mu = \exp(\boldsymbol{x}^t\boldsymbol{\beta})$. Notons

$$\mu(x_j) = \exp(\beta_0 + \beta_1 x_1 + \cdots \beta_j x_j + \cdots + \beta_p x_p) = e^{\boldsymbol{x}^t\boldsymbol{\beta}}$$

et

$$\begin{aligned} \mu(x_j + 1) &= \exp(\beta_0 + \beta_1 x_1 + \cdots \beta_j (x_j + 1) + \cdots + \beta_p x_p) \\ &= e^{\boldsymbol{x}^t\boldsymbol{\beta} + \beta_j} = e^{\boldsymbol{x}^t\boldsymbol{\beta}} e^{\beta_j} = e^{\beta_j}\mu(x_j). \end{aligned}$$

Ainsi, lorsque l'on passe de x_j à $x_j + 1$ alors que toutes les autres variables explicatives restent inchangées, la moyenne de la variable réponse Y est multipliée par e^{β_j} (effet multiplicatif). Si la valeur de β_j est positive, la moyenne

sera augmentée, si la valeur de β_j est négative, la moyenne diminuera.

Si x_j est une indicatrice,

$$e^{\beta_j} = \frac{\mu(x_j = 1)}{\mu(x_j = 0)}.$$

Alors e^{β_j} est le ratio entre la moyenne pour les individus avec $x_j = 1$ et celle pour les individus avec $x_j = 0$.

Remarque : un effet multiplicatif, contrairement à un effet additif, ne dépend pas de l'échelle des données.

4.2.4 Lien logit : rapport de cotes

Le lien logit est utilisé en régression logistique. Ce lien est défini uniquement pour des valeurs de μ entre 0 et 1. Étant donné que nous somme dans le contexte de la régression logistique, μ est en fait $\pi = P(Y = 1)$. Plutôt que de parler d'effet sur la moyenne de Y nous parlerons donc d'effet sur le risque ou la probabilité que $Y = 1$.

Avec un lien logit, on a :

$$\begin{aligned} logit(\pi) = \ln\left(\frac{\pi}{1-\pi}\right) &= \boldsymbol{x}^t\boldsymbol{\beta} \\ \frac{\pi}{1-\pi} &= e^{\boldsymbol{x}^t\boldsymbol{\beta}} \\ \pi &= (1-\pi)e^{\boldsymbol{x}^t\boldsymbol{\beta}} \\ \pi(1+e^{\boldsymbol{x}_i^t\boldsymbol{\beta}}) &= e^{\boldsymbol{x}^t\boldsymbol{\beta}} \\ \pi &= \frac{e^{\boldsymbol{x}^t\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}^t\boldsymbol{\beta}}} \end{aligned}$$

Notons $\pi(x_j) = \frac{e^{\boldsymbol{x}^t\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}^t\boldsymbol{\beta}}}$, donc $\frac{\pi(x_j)}{1-\pi(x_j)} = e^{\boldsymbol{x}^t\boldsymbol{\beta}}$. On a que

$$\begin{aligned} \frac{\pi(x_j+1)}{1-\pi(x_j+1)} &= e^{\boldsymbol{x}^t\boldsymbol{\beta}}e^{\beta_j} \qquad \text{(développement omis car idem au lien log)} \\ &= e^{\beta_j}\frac{\pi(x_j)}{1-\pi(x_j)}. \end{aligned}$$

Ainsi, lorsque l'on passe de x_j à $x_j + 1$ alors que toutes les autres variables explicatives restent inchangées, la cote de la probabilité que $Y = 1$ est multipliée par e^{β_j}.

Si la variable x_j est une indicatrice, e^{β_j} est le rapport de cote de la probabilité que $Y = 1$ pour les individus pour lesquels $x_j = 1$ par rapport aux individus pour lesquels $x_j = 0$, en conservant les autres variables explicatives inchangées :

$$e^{\beta_j} = \frac{\pi(x_j = 1)/(1 - \pi(x_j = 1))}{\pi(x_j = 0)/(1 - \pi(x_j = 0))}.$$

On peut facilement illustrer ce résultat dans le cas d'un tableau de fréquences 2×2. On a alors une seule variable explicative x.

	$y = 1$	$y = 0$
$x = 1$	$\pi(x = 1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$1 - \pi(x = 1) = \frac{1}{1+e^{\beta_0+\beta_1}}$
$x = 0$	$\pi(x = 0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$	$1 - \pi(x = 0) = \frac{1}{1+e^{\beta_0}}$

$$\begin{aligned} RC &= \frac{\text{cote chez les sujets pour qui } x = 1}{\text{cote chez les sujets pour qui } x = 0} \\ &= \frac{\pi(x = 1)/(1 - \pi(x = 1))}{\pi(x = 0)/(1 - \pi(x = 0))} \\ &= \frac{\pi(x = 1)(1 - \pi(x = 0))}{(1 - \pi(x = 1))\pi(x = 0)} \\ &= \frac{\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\frac{1}{1+e^{\beta_0}}}{\frac{1}{1+e^{\beta_0+\beta_1}}\frac{e^{\beta_0}}{1+e^{\beta_0}}} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_1}} = e^{\beta_1} \end{aligned}$$

Comme on l'a appris précédemment, si la probabilité de $Y = 1$ est faible, on peut interpréter ce rapport de cotes comme un risque relatif.

4.2.5 Autres liens

Pour les autres fonctions de liens mentionnées jusqu'à maintenant, soit les liens probit et log-log proposés en régression logistique, il n'y a pas d'interprétation aussi intéressante que le rapport de cotes pour le lien logit. Cependant, ces fonctions de lien étant monotones croissantes, on peut affirmer qu'un paramètre positif dénote une association positive entre une variable explicative

et la variable réponse et un paramètre négatif représente une association négative.

4.2.6 Interactions

Supposons le modèle linéaire généralisé suivant :

$$g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2.$$

Le paramètre β_{12} devant l'interaction double $x_1 x_2$ représente l'effet de x_2 sur la relation entre Y et x_1. Si ce paramètre prend une valeur nulle, c'est que l'association entre Y et x_1 n'est pas influencée par x_2. Elle est donc « homogène » par rapport à x_2 (elle est toujours la même peu importe la valeur de x_2). On pourrait aussi voir plutôt β_{12} comme l'effet de x_1 sur la relation entre Y et x_2.

En général, si tous les paramètres devant les interactions comprenant une certaine variable explicative x_j sont nuls, c'est que l'association entre Y et x_j est homogène par rapport à toutes les autres variables du modèle.

Cependant, si le paramètre devant une interaction n'est pas nul, comment l'interpréter ? Encore une fois, ça dépend de la fonction de lien. Voici, par exemple, l'interprétation qu'on pourrait en faire pour une fonction de lien linéaire.

Interaction double avec un lien linéaire :

Notons

$$\mu(x_1) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

et

$$\begin{aligned} \mu(x_1 + 1) &= \beta_0 + \beta_1(x_1 + 1) + \beta_2 x_2 + \beta_{12}(x_1 + 1)x_2 \\ &= \mu(x_1) + \beta_1 + \beta_{12} x_2. \end{aligned}$$

Ainsi, l'ajout à la moyenne de Y lorsque l'on passe de x_1 à $x_1 + 1$ alors que x_2 reste inchangée n'est pas seulement β_1, il s'agit plutôt de $\beta_1 + \beta_{12} x_2$. On peut donc voir β_1 comme l'effet additif de base et $\beta_{12} x_2$ comme l'effet additif supplémentaire, fonction de la valeur de x_2.

4.3 Ajustement du modèle

Jusqu'à maintenant, nous avons présenté les modèles linéaires généralisés de façon théorique. En pratique, on veut estimer les paramètres $\boldsymbol{\beta}$ du modèle à partir des données d'un échantillon. Pour un GLM, cette estimation est faite par maximum de vraisemblance, tel que présenté dans l'article de Nelder et Wedderburn (1972) qui a introduit les GLM.

4.3.1 Exemple de maximum de vraisemblance : régression Poisson simple avec lien logarithmique

Prenons l'exemple de la régression Poisson simple avec lien logarithmique pour illustrer la méthode d'estimation par maximum de vraisemblance. Le modèle s'écrit comme suit :

$Y_i \sim Poisson(\mu_i)$ indépendantes, avec $\ln(\mu_i) = \beta_0 + \beta_1 x_i$, pour $i = 1, \ldots, n$.

La vraisemblance du modèle s'écrit donc :

$$\begin{aligned} L(\boldsymbol{\beta}) = L((\beta_0, \beta_1)) &= P(Y_1 = y_1, \ldots, Y_n = y_n) \\ &= \prod_{i=1}^{n} P(Y_i = y_i) \quad \text{par indépendance entre les } Y_i \\ &= \prod_{i=1}^{n} \frac{\mu_i^{y_i}}{y_i!} e^{-\mu_i} \quad \text{car } Y_i \sim Poisson(\mu_i) \end{aligned}$$

où $\mu_i = \exp(\beta_0 + \beta_1 x_i)$.

On veut maximiser cette vraisemblance $L(\boldsymbol{\beta})$. Maximiser $L(\boldsymbol{\beta})$ revient à maximiser $\ln(L(\boldsymbol{\beta}))$ puisque le logarithme naturel est une fonction monotone croissante. Nous allons travailler plutôt avec cette quantité puisqu'elle a une forme algébrique plus facilement différenciable.

$$\begin{aligned} \ln(L(\boldsymbol{\beta})) &= \sum_{i=1}^{n} \ln\left(\frac{\mu_i^{y_i}}{y_i!} e^{-\mu_i}\right) \\ &= \sum_{i=1}^{n} (y_i \ln(\mu_i) - \mu_i - \ln(y_i!)) \\ &= \sum_{i=1}^{n} (y_i(\beta_0 + \beta_1 x_i) - \exp(\beta_0 + \beta_1 x_i)) - \sum_{i=1}^{n} \ln(y_i!) \end{aligned}$$

Dans cette somme, le dernier élément $\sum_{i=1}^{n} \ln(y_i!)$ est une constante, car il ne dépend pas des paramètres. On va donc l'omettre dans les calculs subséquents.

Nous allons maintenant dériver la quantité $\ln(L(\boldsymbol{\beta}))$ par rapport à chacun des paramètres. Pour maximiser la vraisemblance, il faut d'abord trouver les valeurs de β_0 et β_1 qui annulent les dérivées partielles de $\ln(L(\boldsymbol{\beta}))$.

$$\begin{aligned} S(\boldsymbol{\beta}) = \begin{pmatrix} \frac{\partial \ln(L(\boldsymbol{\beta}))}{\partial \beta_0} \\ \frac{\partial \ln(L(\boldsymbol{\beta}))}{\partial \beta_1} \end{pmatrix} &= \begin{pmatrix} \sum_{i=1}^{n} (y_i - \exp(\beta_0 + \beta_1 x_i)) \\ \sum_{i=1}^{n} (y_i x_i - \exp(\beta_0 + \beta_1 x_i) x_i) \end{pmatrix} \\ &= \sum_{i=1}^{n} (y_i - \mu_i) \begin{pmatrix} 1 \\ x_i \end{pmatrix} \end{aligned}$$

Tentons de trouver une solution algébrique au système à deux équations et deux inconnus $S(\boldsymbol{\beta}) = (0, 0)^t$. La première équation se développe ainsi :

$$\begin{aligned} \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \exp(\beta_0 + \beta_1 x_i)) &= 0 \\ -e^{\beta_0} \sum_{i=1}^{n} e^{\beta_1 x_i} &= -\sum_{i=1}^{n} y_i \\ e^{\beta_0} &= \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} e^{\beta_1 x_i}} \end{aligned}$$

Remplaçons maintenant cette quantité dans la deuxième équation :

$$\begin{aligned} \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \exp(\beta_0 + \beta_1 x_i) x_i &= 0 \\ -e^{\beta_0} \sum_{i=1}^{n} x_i e^{\beta_1 x_i} &= -\sum_{i=1}^{n} y_i x_i \\ \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} e^{\beta_1 x_i}} \sum_{i=1}^{n} x_i e^{\beta_1 x_i} &= \sum_{i=1}^{n} y_i x_i \\ \frac{\sum_{i=1}^{n} x_i e^{\beta_1 x_i}}{\sum_{i=1}^{n} e^{\beta_1 x_i}} &= \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} y_i} \end{aligned}$$

Isoler β_1 dans cette équation est impossible. Il n'y a donc pas de solution algébrique à la maximisation de cette vraisemblance. Pour une distribution

normale, la solution algébrique existe et elle est la même que la solution trouvée par la méthode des moindres carrés. Cependant, pour les autres distributions de la famille exponentielle, il n'y a en général pas de solution algébrique. La méthode d'ajustement du modèle doit fonctionner, peu importe la distribution de la famille exponentielle choisie. Ainsi, la vraisemblance d'un GLM est maximisée en utilisant un algorithme numérique.

4.3.2 Algorithme numérique de maximisation

Nous présenterons ici brièvement deux algorithmes numériques usuels de maximisation de la log-vraisemblance pour un GLM : l'algorithme de Newton-Raphson et le Fisher scoring.

Le problème est le suivant : on cherche à trouver numériquement la solution de

$$S(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \ln\left(L(\boldsymbol{\beta})\right) = \mathbf{0}$$

où L est la vraisemblance du modèle linéaire généralisé à ajuster, $\boldsymbol{\beta}$ et $\mathbf{0}$ sont des vecteurs. La fonction S est nommée en statistique fonction score. On la nomme parfois aussi gradient puisque $S(\boldsymbol{\beta})$ est le vecteur des dérivées partielles de la log-vraisemblance par rapport aux paramètres $\boldsymbol{\beta}$. Une propriété des distributions de la famille exponentielle est que leur fonction de vraisemblance ou de log-vraisemblance est strictement concave. En conséquence, il existe une seule solution à $S(\boldsymbol{\beta}) = \mathbf{0}$ et le point en lequel $S(\boldsymbol{\beta}) = \mathbf{0}$ est nécessairement un maximum. Nous noterons $\hat{\boldsymbol{\beta}}$ ce point, il s'agit de l'estimateur du maximum de vraisemblance de $\boldsymbol{\beta}$, le vecteur des paramètres du modèle linéaire généralisé.

Algorithme de Newton-Raphson

Par expansion en série Taylor autour d'un point quelconque $\boldsymbol{\beta}_0$, on a que :

$$S(\hat{\boldsymbol{\beta}}) = 0 \approx S(\boldsymbol{\beta}_0) + \left.\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0).$$

Notons

$$I_o(\boldsymbol{\beta}) = -H(\boldsymbol{\beta}) = -\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\frac{\partial^2}{\partial \boldsymbol{\beta_j} \boldsymbol{\beta_{j'}}} \ln\left(L(\boldsymbol{\beta})\right),$$

soit moins le hessien ou matrice des dérivées secondes de la fonction de log-vraisemblance par rapport aux paramètres $\boldsymbol{\beta}$. La matrice $I_o(\boldsymbol{\beta})$ se nomme matrice d'information observée. On veut donc résoudre :

$$0 \approx S(\boldsymbol{\beta}_0) - I_o(\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0).$$

En isolant $\hat{\boldsymbol{\beta}}$ dans cette expression, on obtient :

$$\hat{\boldsymbol{\beta}} \approx \boldsymbol{\beta}_0 + I_o^{-1}(\boldsymbol{\beta}_0)S(\boldsymbol{\beta}_0).$$

Cette approximation est le principe de base de l'algorithme de Newton-Raphson. Cet algorithme fonctionne peu importe la distribution de la famille exponentielle choisie pour la composante aléatoire du modèle, car $S(\boldsymbol{\beta})$ et $I_o(\boldsymbol{\beta})$ sont définis de façon générale.

Étapes de l'algorithme de Newton-Raphson :

1. Fixer des valeurs initiales pour le vecteur de paramètres : $\boldsymbol{\beta}^{iter=0}$.

2. Calculer $S(\boldsymbol{\beta}^{iter})$ et $I_o(\boldsymbol{\beta}^{iter})$, soit la fonction score et la matrice d'information observée au point $\boldsymbol{\beta} = \boldsymbol{\beta}^{iter}$.

3. Mettre à jour la valeur des paramètres à l'aide de la formule :

$$\boldsymbol{\beta}^{iter+1} = \boldsymbol{\beta}^{iter} + I_o^{-1}(\boldsymbol{\beta}^{iter})S(\boldsymbol{\beta}^{iter}).$$

4. Calculer la valeur d'un indice de convergence, par exemple la distance euclidienne entre $\boldsymbol{\beta}^{iter}$ et $\boldsymbol{\beta}^{iter+1}$.

5. Si la valeur de cet indice est supérieure au critère de convergence que l'on a choisi préalablement (par exemple 0.0001) et que le nombre d'itérations effectuées est inférieur au nombre maximum d'itérations que l'on a aussi choisi préalablement (par exemple 100) :
   – on incrémente le nombre d'itérations ($iter$ devient $iter + 1$) et
   – on recommence les étapes 2, 3, 4 et 5 ;
   sinon l'algorithme s'arrête et
   – on dit qu'il a convergé si le critère de convergence a été atteint ;
   – on dit qu'il n'a pas convergé si le nombre maximum d'itérations a été atteint avant que le critère de convergence ne soit atteint, ou si des problèmes numériques (par exemple des divisions par zéro) ont forcé l'algorithme à arrêter avant même que le critère de convergence ou le nombre maximum d'itérations soit atteint.

Solutions potentielles aux problèmes de non-convergence :
Les logiciels statistiques dans lesquels cet algorithme d'ajustement des GLM est implanté proposent habituellement des valeurs par défaut judicieuses pour :

- les valeurs initiales $\boldsymbol{\beta}^{iter=0}$;
- le critère de convergence ;
- le nombre maximum d'itérations.

Cependant, si on rencontre en pratique des problèmes de non-convergence, on peut essayer d'autres valeurs pour ces paramètres de l'algorithme numérique. Si on fournit des valeurs initiales proches de la solution, l'algorithme devrait converger facilement et en peu d'itérations. Par exemple, on pourrait obtenir de bonnes valeurs initiales en ajustant un modèle linéaire classique (avec des résidus normaux) avec la même composante systématique que le GLM, sur la variable réponse transformée par la fonction de lien du GLM.

Souvent, les valeurs par défaut des paramètres de l'algorithme sont déjà très appropriées et modifier leurs valeurs ne nous aide pas à faire converger l'algorithme. Dans ce cas, c'est probablement que le modèle linéaire généralisé choisi s'ajuste mal aux données. Il faut alors envisager de modifier le modèle en :

- choisissant une autre distribution de la famille exponentielle pour la composante aléatoire du modèle, ou en
- modifiant la composante systématique du modèle, par exemple en ajoutant ou en enlevant des variables explicatives ou des interactions entre des variables explicatives, ou en
- changeant la fonction de lien.

Estimateur du maximum de vraisemblance de $\boldsymbol{\beta}$:
Si l'algorithme à convergé, disons à l'itération $iter = niter$, on obtient que l'estimateur du maximum de vraisemblance des paramètres du modèle est :

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{niter}.$$

Une estimation de la matrice de variance-covariance de $\hat{\boldsymbol{\beta}}$ est l'inverse de la matrice d'information observée calculée au point $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$:

$$\hat{\sigma}^2(\hat{\boldsymbol{\beta}}) = I_o^{-1}(\hat{\boldsymbol{\beta}}).$$

Fisher scoring

Un autre algorithme fréquemment utilisé pour ajuster un GLM est le Fisher scoring. La seule différence entre le Fisher scoring et l'algorithme de Newton-Raphson est que l'on utilise l'espérance de la matrice d'information observée plutôt que la matrice d'information observée $I_o(\boldsymbol{\beta})$ elle-même. Cette matrice est nommée matrice d'information espérée ou matrice d'information de Fisher :

$$I_e(\boldsymbol{\beta}) = E[I_o(\boldsymbol{\beta})] = -E\left[\frac{\partial^2}{\partial \boldsymbol{\beta_j} \boldsymbol{\beta_{j'}}} \ln\left(L(\boldsymbol{\beta})\right)\right].$$

Donc en remplaçant $I_0(\boldsymbol{\beta})$ par $I_e(\boldsymbol{\beta})$ dans les étapes de l'algorithme numérique énoncées précédemment, on effectue du Fisher scoring. Cet algorithme est parfois présenté sous la forme d'une méthode de « moindres carrés itérativement repondérés ».

Aussi, lorsque l'on utilise le Fisher scoring , l'estimation de la matrice de variance-covariance de $\hat{\boldsymbol{\beta}}$ est l'inverse de la matrice d'information espérée calculée au point $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$:

$$\hat{\sigma}^2(\hat{\boldsymbol{\beta}}) = I_e^{-1}(\hat{\boldsymbol{\beta}}).$$

Cet estimateur est tiré de l'inégalité de Cramer-Rao (Casella et Berger, 2002, section 7.3.2) qui dit que :

$$Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\hat{\boldsymbol{\beta}}) \geq I_e^{-1}(\boldsymbol{\beta}).$$

Équivalence entre les deux algorithmes : Si la fonction de lien choisie dans le modèle linéaire généralisé est la fonction dite canonique pour la distribution de la famille exponentielle choisie, alors les matrices d'informations espérées et observées sont les mêmes. Sans entrer dans les détails des distributions de la famille exponentielle, mentionnons seulement que les fonctions de liens canoniques pour les distributions abordées dans ce cours sont les suivantes :

Composante aléatoire	Fonction de lien canonique
normale	identité
binomiale	logit
Poisson	ln

Donc avec la fonction de lien canonique, $I_0(\boldsymbol{\beta}) = I_e(\boldsymbol{\beta})$ et les algorithmes de Newton-Raphson et Fisher scoring sont équivalents.

Exemple de matrices d'information :
régression Poisson simple avec lien logarithmique

On a déjà trouvé la formule explicite pour $S(\boldsymbol{\beta})$ pour la régression Poisson simple avec lien logarithmique. Il s'agit de

$$S(\boldsymbol{\beta}) = \begin{pmatrix} \sum_{i=1}^n (y_i - \exp(\beta_0 + \beta_1 x_i)) \\ \sum_{i=1}^n (y_i x_i - \exp(\beta_0 + \beta_1 x_i) x_i) \end{pmatrix}.$$

La matrice d'information observée a la forme suivante :

$$\begin{aligned} I_o(\boldsymbol{\beta}) &= -\begin{pmatrix} \frac{\partial^2 \ln(L(\boldsymbol{\beta}))}{\partial \beta_0^2} & \frac{\partial^2 \ln(L(\boldsymbol{\beta}))}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 \ln(L(\boldsymbol{\beta}))}{\partial \beta_0 \partial \beta_1} & \frac{\partial^2 \ln(L(\boldsymbol{\beta}))}{\partial \beta_1^2} \end{pmatrix} \\ &= -\begin{pmatrix} -\sum_{i=1}^n \exp(\beta_0 + \beta_1 x_i) & -\sum_{i=1}^n \exp(\beta_0 + \beta_1 x_i) x_i \\ -\sum_{i=1}^n \exp(\beta_0 + \beta_1 x_i) x_i & -\sum_{i=1}^n \exp(\beta_0 + \beta_1 x_i) x_i^2 \end{pmatrix} \\ &= \sum_{i=1}^n \mu_i \begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \end{pmatrix} \end{aligned}$$

La fonction de lien ln est la fonction de lien canonique pour la distribution Poisson. Ainsi, les matrices d'information observée et espérée sont égales. On voit d'ailleurs ici que $I_o(\boldsymbol{\beta})$ ne contient plus de y_i, elle ne contient donc plus rien d'aléatoire. C'est une constante, donc son espérance est égale à elle-même :

$$I_e(\boldsymbol{\beta}) = E[I_o(\boldsymbol{\beta})] = I_o(\boldsymbol{\beta}).$$

Exemple de matrices d'information :
régression Poisson simple avec lien identité

Tentons maintenant de trouver la forme explicite des matrices d'information observée et espérée pour un GLM n'utilisant pas une fonction de lien

canonique. Nous utiliserons l'exemple de la régression Poisson simple avec lien identité. Le modèle s'écrit comme suit :

$Y_i \sim Poisson(\mu_i)$ indépendantes, avec $\mu_i = \beta_0 + \beta_1 x_i$, pour $i = 1, \ldots, n$.

La vraisemblance du modèle s'écrit donc encore :

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{\mu_i^{y_i}}{y_i!} e^{-\mu_i}$$

mais cette fois $\mu_i = \beta_0 + \beta_1 x_i$. La log-vraisemblance s'écrit :

$$\begin{aligned} \ln(L(\boldsymbol{\beta})) &= \sum_{i=1}^{n} (y_i \ln(\mu_i) - \mu_i - \ln(y_i!)) \\ &= \sum_{i=1}^{n} (y_i \ln(\beta_0 + \beta_1 x_i) - (\beta_0 + \beta_1 x_i)) - \sum_{i=1}^{n} \ln(y_i!). \end{aligned}$$

Laissons tomber le dernier terme qui est une constante.

Les dérivées partielles de la log-vraisemblance par rapport à chacun des paramètres sont :

$$\begin{aligned} S(\boldsymbol{\beta}) = \begin{pmatrix} \frac{\partial \ln(L(\boldsymbol{\beta}))}{\partial \beta_0} \\ \frac{\partial \ln(L(\boldsymbol{\beta}))}{\partial \beta_1} \end{pmatrix} &= \begin{pmatrix} \sum_{i=1}^{n} \left(\frac{y_i}{\beta_0 + \beta_1 x_i} - 1 \right) \\ \sum_{i=1}^{n} \left(\frac{y_i x_i}{\beta_0 + \beta_1 x_i} - x_i \right) \end{pmatrix} \\ &= \sum_{i=1}^{n} \left(\frac{y_i}{\mu_i} - 1 \right) \begin{pmatrix} 1 \\ x_i \end{pmatrix} \end{aligned}$$

La matrice d'information observée a donc la forme suivante :

$$\begin{aligned} I_o(\boldsymbol{\beta}) &= - \begin{pmatrix} \frac{\partial^2 \ln(L(\boldsymbol{\beta}))}{\partial \beta_0^2} & \frac{\partial^2 \ln(L(\boldsymbol{\beta}))}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 \ln(L(\boldsymbol{\beta}))}{\partial \beta_0 \partial \beta_1} & \frac{\partial^2 \ln(L(\boldsymbol{\beta}))}{\partial \beta_1^2} \end{pmatrix} \\ &= - \begin{pmatrix} -\sum_{i=1}^{n} \frac{y_i}{(\beta_0 + \beta_1 x_i)^2} & -\sum_{i=1}^{n} \frac{y_i x_i}{(\beta_0 + \beta_1 x_i)^2} \\ -\sum_{i=1}^{n} \frac{y_i x_i}{(\beta_0 + \beta_1 x_i)^2} & -\sum_{i=1}^{n} \frac{y_i x_i^2}{(\beta_0 + \beta_1 x_i)^2} \end{pmatrix} \\ &= \sum_{i=1}^{n} \frac{y_i}{\mu_i^2} \begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \end{pmatrix} \end{aligned}$$

Trouvons maintenant la forme explicite de la matrice d'information espérée. Il s'agit de l'espérance de la matrice d'information observée. Pour calculer cette espérance, il faut maintenant voir les y_i comme des quantités aléatoires. Utilisons donc maintenant une lettre majuscule pour les représenter :

$$\begin{aligned} I_e(\boldsymbol{\beta}) = E[I_o(\boldsymbol{\beta})] &= E\left[\sum_{i=1}^{n} \frac{Y_i}{\mu_i^2} \begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \end{pmatrix}\right] \\ &= \sum_{i=1}^{n} \frac{E[Y_i]}{\mu_i^2} \begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \end{pmatrix} \\ &= \sum_{i=1}^{n} \frac{\mu_i}{\mu_i^2} \begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \end{pmatrix} = \sum_{i=1}^{n} \frac{1}{\mu_i} \begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \end{pmatrix}. \end{aligned}$$

4.3.3 Estimation du paramètre de dispersion ϕ

Tout ce qui précède traitait de l'estimation des paramètres de la composante systématique du modèle : $\boldsymbol{\beta}$. Cependant, le modèle comporte parfois un paramètre supplémentaire associé à la composante aléatoire du modèle : le paramètre de dispersion ϕ. En régression logistique et Poisson, ce paramètre n'a pas à être estimé. Il prend la valeur fixe de 1. Cependant, si la distribution de la famille exponentielle choisie pour la composante aléatoire est autre que Bernoulli ou Poisson, il doit être estimé. Cette estimation peut se faire par maximum de vraisemblance avec les algorithmes Newton-Raphson et Fisher scoring en ajoutant un élément à la fonction score, soit la dérivée de la log-vraisemblance par rapport à ϕ. La matrice d'information contiendra alors une ligne et une colonne de plus. Cette estimation peut aussi se faire à partir de statistiques utilisées pour valider le modèle, soit la déviance et la statistique X^2 de Pearson. Ces statistiques seront présentées plus loin.

4.4 Format de données : une ligne par individu versus données groupées

Lorsque l'on travaille avec des variables explicatives comprenant un petit nombre de valeurs observées (par exemple des variables catégoriques), il arrive souvent que plusieurs individus aient la même combinaison de valeurs pour les variables explicatives. Dans ce cas, on n'observe pas n vecteurs $\boldsymbol{x}_i$ différents, mais plutôt une certaine quantité notée C, avec $C < n$.

Exemple de données impossible à grouper : fautes de frappe

Dans cet exemple, les données comportent une variable explicative numérique. Toutes les observations de cette variable sont distinctes. On ne peut donc pas former de groupes d'individus ayant les mêmes valeurs des variables explicatives.

Exemple de données groupées avec une variable explicative :
test de thé de Fisher

On reprend ici le plus petit jeu de données traité dans la première partie du cours, celui du test de thé de Fisher. Les notes de cours présentaient les données sous le format fréquences à l'intérieur d'un tableau de fréquences à deux variables. Ce jeu de données aurait l'allure suivante dans un fichier de données :

Format fréquences :

Type réel du thé	Évaluation	Fréquence
anglais	bon	3
anglais	moins bon	1
ordinaire	bon	1
ordinaire	moins bon	3

Les fréquences représentent un nombre de tasses de thé. Ici, un individu est une tasse de thé. On peut donc en déduire que le jeu de données sous le format une ligne par individu est le suivant :

Format individus :

Type réel du thé	Évaluation
anglais	bon
anglais	bon
anglais	bon
anglais	moins bon
ordinaire	bon
ordinaire	moins bon
ordinaire	moins bon
ordinaire	moins bon

Ce jeu de données possède $n = 8$ lignes. Ici, on considère que la variable réponse Y est l'évaluation du thé et la variable explicative X est le type réel du thé. Il n'y a que deux valeurs possibles pour cette variable explicative ($C = 2$). Ainsi, les quatre premiers individus peuvent être regroupés. Ce sont ceux pour lesquels le lait a été versé en premier lors de la préparation du thé. Les quatre derniers individus forment un autre groupe : ceux pour lesquels le thé a été versé en premier lors de la préparation du thé. On crée donc à partir du jeu de données complet le jeu de données groupées suivant :

Format données groupées :

X = Type réel	Fréquence de Y = bon	Nombre de tasses
anglais	3	4
ordinaire	1	4

Ce jeu de données résume toute l'information comprise dans le jeu de données complet.

Exemple de données groupées avec plus d'une variable explicative : vols d'avions

Reprenons maintenant l'exemple des vols d'avions vus dans la section du cours sur les tableaux de fréquences à trois variables. Ce jeu de données comprend $n = 11\ 000$ individus. Si l'on traite ces données avec un GLM, il faut identifier une variable réponse. Il s'agit ici d'une variable réponse binaire indiquant si le vol est en retard. Les variables explicatives sont la

compagnie aérienne et la ville de départ du vol. Le jeu de données en format individus serait très long, il comporterait 11 000 lignes. Le jeu de données groupées pour sa part est très succinct. Les deux variables explicatives sont catégoriques, la première a deux modalités et la deuxième en a cinq. Ainsi, il y a $C = 2 \times 5 = 10$ combinaisons possibles de ces deux variables. Les données groupées sont les suivantes :

Compagnie	Ville	Nombre de retards	Nombre de vols
Alaska	LA	62	559
AmWest	LA	117	811
Alaska	Phoenix	12	233
AmWest	Phoenix	415	5255
Alaska	San Diego	20	232
AmWest	San Diego	65	448
Alaska	San Fran.	102	605
AmWest	San Fran.	129	449
Alaska	Seattle	305	2146
AmWest	Seattle	61	262

La façon proposée précédemment pour noter un GLM en général, soit

$$Y_i \sim \mathcal{L}(\mu_i, \phi) \text{ indépendantes, avec } g(\mu_i) = \boldsymbol{x}_i^t \boldsymbol{\beta}, \text{ pour } i = 1, \ldots, n,$$

est appropriée lorsque les observations ont toutes un vecteur $\boldsymbol{x}_i$ unique. En effet, Y_i est la variable aléatoire Y sachant que $\boldsymbol{x} = \boldsymbol{x}_i$. Si plusieurs $\boldsymbol{x}_i$ sont identiques, il serait plus clair de noter le modèle ainsi :

$$Y|\boldsymbol{x}_c \sim \mathcal{L}(\mu_c, \phi) \text{ indépendantes, avec } g(\mu_c) = \boldsymbol{x}_c^t \boldsymbol{\beta}, \text{ pour } c = 1, \ldots, C.$$

Le modèle postule donc que l'espérance de Y dépend la forme de la distribution $\mathcal{L}$ ainsi que des valeurs des paramètres de la distribution. Le paramètre μ_c dépend pour sa part de la valeur prise par les variables explicatives. Tous les individus pour lesquels $\boldsymbol{x} = \boldsymbol{x}_c$ font partie du même groupe, nommé groupe c.

Pour ajuster ce modèle, on peut utiliser le jeu de données complet, soit sous le format une ligne par individu (n lignes), ou des *données groupées*. Ce jeu de données abrégé contient C lignes, soit une ligne par combinaison

observée de valeurs des variables explicatives. Dans ce jeu de données, on somme les valeurs observées de la variable réponse pour tous les individus d'un même groupe. On doit cependant ajouter une colonne au jeu de données pour contenir une information dont on a besoin pour ajuster le modèle sur ces sommes au lieu des données brutes : le nombre d'individus dans chacun des groupes.

On possède donc une nouvelle variable réponse, notée Y^*, qui est la somme de la variable réponse initiale. Les observations de cette variable dans l'échantillon sont :

$$Y_c^* = \sum_{i \in \text{groupe } c} Y_i \quad \text{pour } c = 1, \ldots, C.$$

La distribution de cette nouvelle variable aléatoire dépend de la distribution de Y. On est certain cependant, par indépendance entre les individus, que l'espérance de cette nouvelle variable réponse est :

$$E[Y_c^*] = E\left[\sum_{i \in \text{groupe } c} Y_i\right] = n_c^* \mu_c$$

où n_c^* est le nombre d'individus dans le groupe c. Le modèle postule que $g(\mu_c) = \boldsymbol{x}_c^t \boldsymbol{\beta}$. Si on veut estimer les paramètres $\boldsymbol{\beta}$ de ce modèle sur les données groupées, il faut trouver le bon modèle à ajuster sur la variable réponse groupée Y^*. Voyons ce qui en est pour les distributions Bernoulli et Poisson.

4.4.1 Régression logistique avec des données groupées

Lorsque la variable réponse est binaire (avec une modalité supposée être un succès et l'autre un échec), plutôt que d'avoir la valeur de la variable pour chaque individu, on peut résumer le jeu de données en ayant le nombre de succès et le nombre d'essais pour chaque combinaison observée de valeurs pour les variables explicatives, comme dans l'exemple du test de thé de Fisher ou dans celui des vols d'avions.

Pour les données sous le format individu,le modèle de régression logistique s'énonce ainsi :

$$\begin{aligned} &Y|\boldsymbol{x}_c \sim Bernoulli(\pi_c) \quad \text{indépendantes} \\ &\text{avec} \quad g(\pi_c) = \boldsymbol{x}_c^t\boldsymbol{\beta}, \text{ pour } c = 1, \ldots, C. \end{aligned}$$

Pour les données groupées, la variable réponse change. Elle n'est plus binaire, elle représente maintenant un nombre de succès, noté Y^*, parmi n^* observations. On a alors :

$$Y^*|\boldsymbol{x}_c \sim Binomiale(n_c^*, \pi_c) \quad \text{indépendantes, pour } c = 1, \ldots, C.$$

L'approche adoptée en régression logistique pour estimer $\boldsymbol{\beta}$ à partir des données groupées est de pondérer la variable réponse par l'inverse des nombres d'essais n^*. On a que $E[Y_c^*/n_c^*] = n_c^*\pi_c/n_c^* = \pi_c$. Le modèle demeure donc $g(\pi_c) = \boldsymbol{x}_c^t\boldsymbol{\beta}$.

4.4.2 Régression Poisson avec des données groupées

La somme de variables aléatoires Poisson indépendantes suit aussi une loi Poisson, de paramètre égal à la somme des paramètres des variables aléatoires sommées. Ainsi, si

$$Y|\boldsymbol{x}_c \sim Poisson(\mu_c) \quad \text{indépendantes,}$$

alors

$$Y^*|\boldsymbol{x}_c \sim Poisson(\mu_c^* = n_c^*\mu_c) \quad \text{indépendantes.}$$

En régression Poisson, on n'adopte pas l'approche de pondérer la nouvelle variable réponse. On ajuste plutôt un modèle sur Y^* non pondéré, mais on écrit la composante systématique du nouveau modèle de façon à respecter le modèle sur la variable d'origine qui dit que $g(\mu_c) = \boldsymbol{x}_c^t\boldsymbol{\beta}$. Cette écriture dépend de la fonction de lien choisi. Voyons comment faire avec les fonctions de lien identité et logarithmique.

Fonction de lien identité : Si la fonction de lien est l'identité, le modèle sur la variable réponse initiale Y postule que :

$$g(\mu_c) = \mu_c = \boldsymbol{x}_c^t \boldsymbol{\beta}.$$

Pour la variable réponse groupée Y^* on a :

$$g(\mu_c^*) = \mu_c^* = n_c^* \mu_c = n_c^* \boldsymbol{x}_c^t \boldsymbol{\beta}.$$

Ainsi, dans ce modèle les variables explicatives ne sont plus $\boldsymbol{x}^t = (1, x_1, \ldots, x_p)$. Elles deviennent $(n^*, n^* \times x_1, \ldots, n^* \times x_p)$. Le modèle ne comporte plus d'ordonnée à l'origine.

Fonction de lien identité : Si la fonction de lien est le logarithme naturel, le modèle sur la variable réponse initiale Y postule que :

$$g(\mu_c) = \ln(\mu_c) = \boldsymbol{x}_c^t \boldsymbol{\beta}.$$

Pour la variable réponse groupée Y^* on a :

$$g(\mu_c^*) = \ln(\mu_c^*) = \ln(n_c^* \mu_c) = \ln(n_c^*) + \ln(\mu_c) = \ln(n_c^*) + \boldsymbol{x}_c^t \boldsymbol{\beta}.$$

Ainsi, dans ce modèle les variables explicatives ne changent pas, mais une variable explicative supplémentaire dont le coefficient est 1 doit être ajoutée. Ce type de variable est appelé « **variable offset** ». Il n'y a pas de paramètre ajouté au modèle.

4.5 Inférence sur les paramètres

Faire de l'inférence statistique signifie émettre des conclusions concernant une population à partir de résultats obtenus sur un échantillon de cette population. On a vu comment ajuster un GLM à des données provenant d'un échantillon. Nous allons maintenant chercher à établir des conclusions sur la population cible à l'étude à partir des estimations calculées pour les coefficients $\boldsymbol{\beta}$.

4.5.1 Estimation ponctuelle et par intervalle de confiance

Lors de l'ajustement d'un modèle linéaire généralisé, on estime par maximum de vraisemblance les paramètres du modèle. Ces paramètres sont les $\boldsymbol{\beta}$ de la composante systématique et parfois aussi le paramètre de dispersion ϕ de la composante aléatoire. Étant donné que l'on s'intéresse ici aux régressions logistique et Poisson pour lesquelles le paramètre de dispersion ϕ est fixé à 1, on considérera que le vecteur des paramètres à estimer est uniquement $\boldsymbol{\beta}$. En plus d'estimer ce vecteur de paramètres, on calcule une matrice de variance-covariance asymptotique pour le vecteur des estimations obtenues. Cette matrice est $\hat{\sigma}^2(\hat{\boldsymbol{\beta}}) = I^{-1}(\hat{\boldsymbol{\beta}})$ où I est la matrice d'information observée si l'algorithme de Newton-Raphson a été employé ou la matrice d'information espérée si le Fisher scoring a été utilisé. Les erreurs-types des paramètres sont la racine carrée des éléments sur la diagonale de cette matrice. Ainsi, $\hat{\sigma}(\hat{\beta}_j)$ est la racine carrée de l'élément correspondant à β_j dans la matrice $\hat{\sigma}^2(\hat{\boldsymbol{\beta}})$. On peut se servir de ces erreurs-types pour construire des intervalles de confiance de Wald de niveau $(1-\alpha)\%$ pour tous les paramètres du vecteur $\boldsymbol{\beta}$ comme suit :

$$\beta_j \ \in \ \hat{\beta}_j \ \pm \ z_{\alpha/2} \, \hat{\sigma}(\hat{\beta}_j).$$

4.5.2 Test de Wald sur un paramètre

Comme mentionné précédemment, un paramètre β_j nul signifie que la variable explicative dont il est le coefficient, x_j, n'a pas de lien avec la variable réponse Y. Il est donc d'intérêt de tester si la valeur d'un paramètre est nulle dans la population à l'étude. On peut faire un test de Wald pour confronter

les hypothèses :

$$\begin{aligned} H_0 &: \beta_j = 0 \quad \text{(il n'y a pas de lien entre } x_j \text{ et } Y) \\ H_1 &: \beta_j \neq 0 \quad \text{(il y a un lien entre } x_j \text{ et } Y) \end{aligned}$$

La statistique de ce test est la suivante :

$$Z_w = \frac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)} \qquad \xrightarrow[H_0]{\text{asymp.}} \qquad N(0,1).$$

Sous H_0, la loi asymptotique de la statistique de test est normale standard. On peut effectuer un test bilatéral ou unilatéral avec cette statistique de test.

La statistique de Wald peut aussi être définie ainsi :

$$Z_w^2 = \frac{\hat{\beta}_j^2}{\hat{\sigma}^2(\hat{\beta}_j)} \qquad \xrightarrow[H_0]{\text{asymp.}} \qquad \chi_1^2.$$

Avec cette statistique, qui suit asymptotiquement une loi du khi-deux à un degré de liberté, seul un test bilatéral peut être effectué.

Remarque : Il existe aussi une version multivariée (test sur plus d'un paramètre) du test de Wald, mais on ne le verra pas dans ce cours. On utilisera plutôt le test de rapport de vraisemblance décrit dans la sous-section suivante pour effecteur des tests multivariés.

4.5.3 Test de rapport de vraisemblance sur plusieurs paramètres

En utilisant la vraisemblance maximale des modèles linéaires généralisés, on peut facilement construire un test de rapport de vraisemblance pour vérifier si un groupe de paramètres β_j avec $j \in \mathcal{B}$ sont nuls. Les hypothèses du test sont les suivantes :

$$\begin{aligned} H_0 &: \beta_j = 0 \text{ pour tout } j \in \mathcal{B} \\ &: \text{le modèle réduit est équivalent au modèle complet} \\ H_1 &: \beta_j \neq 0 \text{ pour au moins un } j \in \mathcal{B} \\ &: \text{le modèle réduit n'est pas équivalent au modèle complet} \end{aligned}$$

où le modèle complet est celui comprenant les paramètres β_j avec $j \in \mathcal{B}$ et le modèle réduit est celui dans lequel ces paramètres sont fixés à zéro. Ainsi, le modèle réduit ne contient pas les paramètres β_j avec $j \in \mathcal{B}$, ni les variables explicatives x_j avec $j \in \mathcal{B}$ dont ces paramètres sont les coefficients .

La statistique de test est :

$$LR = 2\{max\ln(L_C) - max\ln(L_R)\} \qquad \xrightarrow[H_0]{\text{asymp.}} \qquad \chi^2_b$$

où $max\ln(L_C)$ est la log-vraisemblance maximale du modèle complet,
$max\ln(L_R)$ est la log-vraisemblance maximale du modèle réduit et
b est le nombre de paramètres β_j testés (soit le nombre d'éléments dans l'ensemble $\mathcal{B}$).

Notez que les degrés de liberté b peuvent aussi se calculer par le nombre de paramètres du modèle complet moins le nombre de paramètres du modèle réduit. Les programmes informatiques ajustant des modèles linéaires généralisés fournissent habituellement les quantités $max\ln(L_R)$ et $max\ln(L_R)$.

Utilités potentielles du test de rapport de vraisemblance :
La statistique de test LR peut permettre d'effectuer les tests suivants :

- Test d'un lien entre la variable réponse Y et une variable explicative catégorique, représentée dans la composante systématique du modèle par une ou plusieurs variables (la plupart du temps des indicatrices).
  Si tous les paramètres devant les variables du modèle représentant la variable explicative catégorique sont nuls, alors le lien n'est pas significatif. Cependant, si au moins un paramètre prend une valeur significativement différente de zéro, on dit qu'il y a un lien entre la variable explicative et la variable réponse. On peut le décrire en étudiant les paramètres devant les variables du modèle représentant la variable explicative catégorique ou par des comparaisons multiples, comme en ANOVA.

- Test pour savoir s'il est préférable d'inclure une variable explicative numérique, ou pouvant être représentée par un score numérique, ayant peu de modalités sous la forme numérique ou catégorique.
  Le modèle avec la variable numérique est le modèle réduit alors que

celui avec la variable catégorique est le modèle complet. Ces modèles sont bien emboîtés puisque, en considérant dans le modèle complet la paramétrisation de la variable explicative numérique sous la forme de polynômes, le modèle réduit est celui conservant uniquement le terme linéaire du modèle complet.

– Test de l'homogénéité de l'association entre une variable explicative et la variable réponse Y.
Si tous les paramètres devant des termes d'interaction comprenant la variable explicative sont nuls, alors l'association est homogène.

Remarque :

Ce test peut en fait être vu comme un test de comparaison de modèles. Attention, pour que le test soit valide, il faut que les modèles comparés soient emboîtés, c'est-à-dire que le modèle réduit soit un cas particulier du modèle complet. Un cas particulier est formé en fixant à zéro certains paramètres du modèle complet.

Pour comparer des modèles non emboîtés, on ne peut pas faire un test de rapport de vraisemblance. On peut cependant comparer la valeur d'un indice d'ajustement du modèle tel le critère d'information d'Akaike (AIC). Ce critère est défini par :

$$AIC = 2(p+1) - 2max\ln(L)$$

où $p+1$ représente le nombre de paramètres dans le modèle et
$max\ln(L)$ est la log-vraisemblance maximale du modèle.
On recherche une valeur la plus petite possible de cet indice. Ainsi, les modèles comprenant un grand nombre de paramètres sont pénalisés. Pour un seul modèle, cet indice n'a pas d'interprétation intéressante. Il a vraiment été développé pour comparer des modèles ajustés sur les mêmes données. C'est le modèle avec le plus petit AIC qui est préféré selon ce critère.

4.6 Prédiction de Y

Une valeur prédite par le modèle est définie par

$$\hat{\mu} = g^{-1}(\boldsymbol{x}^t \hat{\boldsymbol{\beta}})$$

où g^{-1} est l'inverse de la fonction de lien et $\hat{\boldsymbol{\beta}}$ est l'estimateur du maximum de vraisemblance du vecteur de paramètres $\boldsymbol{\beta}$. Donc, pour une observation y_i associée au vecteur de variables explicatives $\boldsymbol{x}_i$, sa valeur prédite par le modèle est $\hat{\mu}_i = g^{-1}(\boldsymbol{x}_i^t \hat{\boldsymbol{\beta}})$. Ainsi, on prédit la variable réponse Y par son espérance, qui est elle fonction des valeurs prises par les variables explicatives. On peut ainsi prédire la valeur moyenne de Y pour des valeurs de $\boldsymbol{x}$ non observées dans l'échantillon.

4.6.1 Prédiction par intervalle de confiance

En plus de fournir une estimation ponctuelle de μ (soit une prédiction), on peut vouloir construire un intervalle de confiance pour μ. Pour ce faire, on construit d'abord un intervalle de confiance pour l'estimation du prédicteur linéaire

$$\hat{\eta} = \boldsymbol{x}^t \hat{\boldsymbol{\beta}}.$$

On tire de l'ajustement du modèle une estimation de la matrice de variance-covariance de $\hat{\boldsymbol{\beta}}$, notée $\hat{\sigma}^2(\hat{\boldsymbol{\beta}})$. L'estimation de l'erreur-type de $\hat{\eta}$ est :

$$\hat{\sigma}(\hat{\eta}) = \sqrt{\boldsymbol{x}^t \hat{\sigma}^2(\hat{\boldsymbol{\beta}}) \boldsymbol{x}}.$$

On peut donc construire un intervalle de confiance de Wald de niveau $(1-\alpha)\%$ pour le prédicteur linéaire η comme suit :

$$\eta \ \in \ \hat{\eta} \ \pm \ z_{\alpha/2} \ \hat{\sigma}(\hat{\eta}).$$

On calcule ensuite un intervalle de confiance de Wald de niveau $(1-\alpha)\%$ pour $\mu = g^{-1}(\eta)$ comme suit :

$$\mu \ \in \ [g^{-1}(\hat{\eta} - z_{\alpha/2} \ \hat{\sigma}(\hat{\eta})) \ , \ g^{-1}(\hat{\eta} + z_{\alpha/2} \ \hat{\sigma}(\hat{\eta}))].$$

On applique donc l'inverse de la fonction de lien aux deux bornes de l'intervalle de confiance de η pour calculer un intervalle de confiance pour μ.

4.7 Validation du modèle

Comme avec n'importe quel modèle statistique, avant de l'utiliser pour faire de l'inférence, il faut s'assurer qu'il s'ajuste bien aux données. C'est l'étape de la validation du modèle. Une statistique de validation propre aux modèles linéaires généralisés est la déviance. Avant de voir comment juger de l'ajustement d'un modèle avec cette statistique, nous allons la définir.

4.7.1 Définition de la déviance

Il existe deux statistiques de déviance pour un GLM : la déviance standardisée et la déviance non standardisée, appelée simplement déviance. La déviance standardisée est définie par :

$$D^* = 2\{max\ln(L_S) - max\ln(L_P)\}.$$

où $max\ln(L_S)$ est la log-vraisemblance maximale d'un modèle saturé,
$max\ln(L_P)$ est la log-vraisemblance maximale du modèle proposé.
Le modèle saturé est un modèle général ne comprenant aucune contrainte particulière sur $E(Y) = \mu$. Un tel modèle comporte un paramètre pour chaque observation et ses valeurs prédites sont, par définition, égales aux valeurs observées : $\hat{\mu}_i = y_i$ pour $i = 1, \ldots, n$. Ainsi, nul besoin de paramétrer et d'ajuster le modèle saturé pour calculer sa log-vraisemblance, on utilise plutôt le fait que les valeurs prédites par ce modèle sont égales aux valeurs observées.

Pour sa part, la déviance (non standardisée) est définie à partir de la déviance standardisée par :

$$D = \phi D^*.$$

En régression logistique et Poisson, étant donné que $\phi = 1$, les deux déviances sont égales. Dérivons l'expression algébrique de la déviance en régression Poisson (à faire en exercice pour la régression logistique).

Exemple de déviance : régression Poisson

Comme trouvée précédemment, en régression Poisson la log-vraisemblance s'écrit :

$$\ln(L(\boldsymbol{\beta})) = \sum_{i=1}^{n}(y_i \ln(\mu_i) - \mu_i - \ln(y_i!)).$$

Le maximum de log-vraisemblance pour le modèle saturé est atteint en $\mu_i = y_i$, il s'écrit donc :

$$max \ln(L_S) = \sum_{i=1}^{n} (y_i \ln(y_i) - y_i - \ln(y_i!)).$$

Le maximum de log-vraisemblance pour le modèle proposé est atteint au point $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, soit lorsque $\mu_i = \hat{\mu}_i = \boldsymbol{x}^t \hat{\boldsymbol{\beta}}$. Il s'écrit :

$$max \ln(L_P) = \sum_{i=1}^{n} (y_i \ln(\hat{\mu}_i) - \hat{\mu}_i - \ln(y_i!)).$$

La formule de la déviance est donc :

$$\begin{aligned} D = D^* &= 2\{max \ln(L_S) - max \ln(L_P)\} \\ &= 2 \left\{ \sum_{i=1}^{n} (y_i \ln(y_i) - y_i - \ln(y_i!)) - \sum_{i=1}^{n} (y_i \ln(\hat{\mu}_i) - \hat{\mu}_i - \ln(y_i!)) \right\} \\ &= 2 \sum_{i=1}^{n} \left(y_i \ln \left(\frac{y_i}{\hat{\mu}_i} \right) - (y_i - \hat{\mu}_i) \right) \end{aligned}$$

Si le modèle comporte une ordonnée à l'origine, la somme $\sum_{i=1}^{n} (y_i - \hat{\mu}_i)$ est en fait la dérivée de la log-vraisemblance par rapport à cette ordonnée à l'origine β_0 (premier élément du vecteur score). Au point $\mu_i = \hat{\mu}_i$, soit le point maximisant la log-vraisemblance, cette dérivée est nulle. Alors dans ce cas la somme $\sum_{i=1}^{n} (y_i - \hat{\mu}_i)$ est égale à zéro.

Test de rapport de vraisemblance avec la déviance en régression logistique et Poisson

Pour les distributions avec $\phi = 1$, soit les distributions Bernoulli et Poisson, la déviance peut servir à calculer la statistique du test de rapport de vraisemblance sur plusieurs paramètres. En effet, la statistique

$$LR = 2\{max \ln(L_C) - max \ln(L_R)\}$$

présentée précédemment peut aussi se calculer par

$$LR = D_R - D_C$$

où D_R est la déviance du modèle réduit et
D_C est la déviance du modèle complet.

Remarque : Lorsque le paramètre de dispersion ϕ ne prend pas la valeur 1, la statistique du test de rapport de vraisemblance en terme de déviance s'écrit plutôt $LR = (D_R - D_C)/\phi$. Le paramètre de dispersion ϕ peut prendre une valeur fixe, ou être estimé. Dans ce cas, le paramètre de déviance au dénominateur de la statistique LR est remplacé par un estimateur de celui-ci, calculé à partir du modèle complet. La loi asymptotique de la statistique LR demeure χ^2_b, avec b égale à la différence dans les nombres de paramètres des modèles complet et réduit. Une statistique F suivant une loi de Fisher peut aussi être construite.

4.7.2 Statistiques d'ajustement du modèle : la déviance et la statistique khi-deux de Pearson

Un modèle présente un bon ajustement, ou une bonne adéquation aux données (en anglais goodness-of-fit), si les valeurs prédites par le modèle sont similaires aux valeurs observées. Le modèle saturé dont on vient de parler présente donc un ajustement parfait puisque ses valeurs prédites sont toutes égales aux valeurs observées. Cependant, un tel modèle n'est pas désirable puisqu'il comporte trop de paramètres, il est trop complexe. De plus, trop coller aux données n'est pas toujours une bonne chose puisque celles-ci proviennent d'un échantillon aléatoire et non de la population complète. On désire un modèle parcimonieux qui présente des relations théoriques interprétables.

On peut voir la déviance standardisée $D^* = 2\{max\ln(L_S) - max\ln(L_P)\}$ comme une statistique de test de rapport de vraisemblance comparant le modèle proposé à un modèle saturé représentant un ajustement parfait. De ce point de vue, la déviance standardisée est une statistique d'ajustement du modèle. Elle permet de tester H_0 : le modèle saturé et le modèle proposé sont équivalents, en d'autres mots, le modèle proposé s'ajuste bien aux données. Elle suit asymptotiquement, *sous certaines conditions* vues plus loin et sous H_0, une loi du khi-deux à $n - (p + 1)$ degrés de liberté, où n représente le nombre de données et $p+1$ le nombre de paramètres dans le modèle à valider.

Remarque : Lorsque la composante aléatoire est normale, sous l'hypothèse que ce postulat soit respecté et que le modèle s'ajuste bien, D^* suit exactement et non asymptotiquement une $\chi^2_{n-(p+1)}$.

Définition générale de la statistique d'ajustement de Pearson

Une autre statistique permet de juger de l'adéquation d'un modèle, il s'agit de la statistique de Pearson ainsi définie :

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

où $V(\hat{\mu}_i) = \widehat{Var}(Y_i)/\hat{\phi}$. On appelle la quantité $V(\hat{\mu})$ la fonction de variance estimée. La forme de cette fonction dépend évidemment de la distribution de la famille exponentielle choisie pour modéliser Y. Pour les distributions étudiées dans ce cours, on a :

Distribution	Estimation de la variance de Y_i
$Y \sim Binomiale(n, \pi)$	$V(\hat{\mu}) = n\hat{\pi}(1-\hat{\pi})$
$Y \sim Poisson(\mu)$	$V(\hat{\mu}) = \hat{\mu}$

Ainsi, pour la loi Poisson, on a $X^2 = \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$. La statistique X^2 a alors la même forme que la statistique du khi-deux de Pearson pour un test d'indépendance dans un tableau de fréquences à 2 variables. Cependant, pour une distribution autre que Poisson, la statistique d'ajustement X^2 n'a pas tout à fait la même forme que la statistique X^2 du test d'indépendance.

La statistique de Pearson standardisée est quant à elle définie par :

$$X^{2\,*} = \frac{X^2}{\phi}.$$

Il n'y a pas de distinction entre ces deux statistiques en régression logistique ou Poisson puisque dans ce cas $\phi = 1$.

Sous l'hypothèse nulle que les valeurs observées sont égales aux valeurs prédites, et *sous certaines conditions* discutées ci-dessous, la distribution asymptotique de la statistique de Pearson standardisée est khi-deux à $n - (p+1)$ degrés de liberté, comme pour la déviance.

Problèmes avec la validité de la loi asymptotique des statistiques d'ajustement

Lorsque l'on calcule ces deux statistiques d'ajustement sur les données au format une ligne par individu, on rencontre souvent un problème avec les conditions à respecter pour s'assurer de la validité de la loi asymptotique que l'on vient de mentionner. Pour expliquer le problème, rappelons de la matière vue dans la première partie du cours. La loi asymptotique de la statistique X^2 du test d'indépendance dans un tableau à 2 variables était considérée valide si au moins 80% des fréquences espérées étaient supérieures à 5. De façon similaire, les valeurs prédites $\hat{\mu}_i$ doivent prendre des valeurs assez grandes pour que la loi asymptotique des statistiques d'ajustement D^* et $X^{2\,*}$ soit valide. Cette condition n'est pas du tout respectée en régression logistique (valeurs prédites entre 0 et 1 pour la plupart des fonctions de lien), ni en régression Poisson lorsque l'on modélise des petits dénombrements.

Les données groupées apportent souvent une solution à ce problème. En effet, ces données comprennent une nouvelle variable réponse, qui est la somme des observations de la variable Y d'origine pour tous les individus présentant le même vecteur de variables explicatives $\boldsymbol{x}$. Les valeurs prédites par le modèle ajusté sur les données groupées sont donc plus grandes ou égales à celles provenant du modèle ajusté sur les données par individu. Voici les formules de la déviance et de la statistique de Pearson calculées sur les données groupées, en régression logistique et Poisson :

Régression	Déviance standardisée avec les données groupées
logistique	$2\sum_{c=1}^{C}\left(y_c^* \ln\left(\frac{y_c^*}{n_c^*\hat{\pi}_c}\right) + (n_c^* - y_c^*)\ln\left(\frac{(n_c^* - y_c^*)}{n_c^*(1-\hat{\pi}_c)}\right)\right)$
Poisson	$2\sum_{c=1}^{C}\left(y_c^* \ln\left(\frac{y_c^*}{\hat{\mu}_c^*}\right) - (y_c^* - \hat{\mu}_c^*)\right)$

Régression	X^2 standardisée avec les données groupées
logistique	$\sum_{c=1}^{C}\frac{(y_c^* - n_c^*\hat{\pi}_c)^2}{n_c^*\hat{\pi}_c(1-\hat{\pi}_c)}$
Poisson	$\sum_{c=1}^{C}\frac{(y_c^* - \hat{\mu}_c^*)^2}{\hat{\mu}_c^*}$

La loi asymptotique de toutes ces statistiques d'ajustement de modèle, sous l'hypothèse nulle d'un bon ajustement et celle que les valeurs prédites ne sont pas trop petites, est $\chi^2_{C-(p+1)}$. Ce résultat nous permet d'effectuer un test formel d'ajustement du modèle.

Règle du pouce : On peut aussi utiliser le fait que l'espérance d'une χ^2 est égale à ses degrés de liberté pour formuler une règle simple pour juger de l'adéquation d'un modèle. Si le modèle s'ajuste bien, la déviance devrait prendre une valeur proche de son espérance, soit $C-(p+1)$. Ainsi, une façon approximative de juger de l'ajustement d'un GLM est de calculer $D^*/(C-(p+1))$ ou ${X^2}^*/(C-(p+1))$. Si ces ratios prennent des valeurs proches de 1, c'est le signe que le modèle s'ajuste bien aux données. Lorsque l'on doute de la validité de la loi asymptotique de D^* et ${X^2}^*$, il est plus logique d'utiliser une telle règle du pouce que de faire un test formel. Cependant, lorsque le nombre de degrés de liberté $C-(p+1)$ est très petit, cette règle ne fonctionne pas bien.

Remarque 1 : En se basant sur la même idée, la déviance non standardisée ou la statistique X^2 non standardisée peuvent être utilisées pour estimer ϕ si ce paramètre n'est pas connu. C'est une méthode autre que celle du maximum de vraisemblance pour estimer ϕ. Cette estimation est très simple. On vient de dire que si le modèle s'ajuste bien et que les valeurs prédites ne sont pas trop petites, on devrait avoir :

$$\frac{D^*}{(C-(p+1))} \approx \frac{{X^2}^*}{(C-(p+1))} \approx 1.$$

De la relation entre les statistiques standardisées et leurs versions non standardisées, on obtient :

$$\begin{aligned} \frac{D}{\phi(C-(p+1))} \approx \frac{X^2}{\phi(C-(p+1))} &\approx 1 \\ \frac{D}{(C-(p+1))} \approx \frac{X^2}{(C-(p+1))} &\approx \phi. \end{aligned}$$

Ainsi, des estimateurs potentiels de ϕ sont :

$$\hat{\phi} = D/(C-(p+1)) \qquad \text{et} \qquad \hat{\phi} = X^2/(C-(p+1)).$$

Pour les raisons déjà mentionnées, on aura plus confiance en la validité de ces estimateurs s'ils sont calculés sur les données groupées que sur les données par individu.

Remarque 2 : Il arrive souvent que le nombre de combinaisons observées des variables explicatives soit pratiquement égal au nombre d'individus dans l'échantillon ($C \approx n$). C'est souvent le cas notamment lorsque le modèle comprend une variable explicative continue. Dans ce cas, les données groupées sont très similaires aux données par individu et on peut douter de la validité de la loi asymptotique des statistiques d'ajustement même avec ces données. C'est particulièrement problématique en régression logistique étant donné que l'on modélise une probabilité, donc les valeurs prédites sont sur une petite échelle. D'autres outils de validation des modèles de régression logistique ont donc été développés, notamment le test de Hosmer et Lemeshow et les courbes ROC. Nous ne présenterons pas ces méthodes ici, elles sont décrites dans Hosmer et Lemeshow (2000).

4.7.3 Étude des valeurs prédites et des résidus pour valider les postulats du modèle

Rappelons les postulats de base d'un modèle linéaire généralisé :

- les observations de la variable réponse Y sont indépendantes ;
- la relation entre la variable réponse Y et les variables explicatives est bien modélisée par la fonction de lien $g()$ choisie ;
- la variable réponse suit bien la distribution $\mathcal{L}$ choisie.

On s'assure d'abord que les observations ont été recueillies de façon à ce que le premier postulat, celui d'indépendance, soit respecté. Pour ce qui est des deux autres postulats, on peut étudier leur validité à partir de graphiques faisant intervenir les valeurs prédites par le modèle ainsi que des résidus du modèle, qui seront définis plus loin.

Étude de la justesse de la fonction de lien choisie

Nous présentons ici deux types de graphiques utiles pour juger si la fonction de lien est adéquate ou non :

Graphique des valeurs observées en fonction des valeurs prédites : Si la fonction de lien est adéquate, le graphique des valeurs observées de Y en fonction des valeurs prédites de Y devrait être linéaire.

Graphiques des valeurs observées et prédites en fonction d'un x_j : Si des variables explicatives numériques sont incluses dans le modèle, on peut tracer le graphique des valeurs observées de Y en fonction d'une de ces variables, disons x_j. On ajoute ensuite à ce graphique les valeurs prédites de Y en fonction de x_j. Les valeurs prédites devraient être proches des valeurs observées. On peut faire ce graphique pour toutes les variables x_j numériques de la composante systématique du modèle.

Ces graphiques sont plus informatifs s'ils sont produits avec les données groupées. On réduit ainsi la variabilité dans les valeurs observées. En fait, en régression logistique, ces graphiques n'ont de sens qu'avec les données groupées. En effet, avec les données au format une ligne par individu, les valeurs observées sont des 0 et des 1.

Étude de la justesse de la distribution choisie

Pour évaluer si la distribution choisie est adéquate, on utilise des résidus du modèle. Pourquoi ne testons-nous pas si la variable réponse suit bien la distribution choisie ? Parce que le modèle postule que l'espérance de Y dépend de la valeur du vecteur de variables explicatives $\boldsymbol{x}$:

$$Y|\boldsymbol{x}_c \sim \mathcal{L}(\mu_c, \phi).$$

Ainsi, dans notre échantillon, le paramètre de centralité de la distribution ne prend pas toujours la même valeur. Les tests d'adéquation de données à une loi supposent que les paramètres de la loi prennent une valeur constante dans l'échantillon, ce qui n'est pas le cas ici.

On utilise donc plutôt des résidus. Un **résidu brut** est simplement défini par la différence entre une valeur observée et une valeur prédite :

$$y_i - \hat{\mu}_i.$$

Cependant, il est très difficile d'examiner la validité de la relation postulée entre l'espérance et la variance à partir de ces résidus. Par exemple, pour un modèle Poisson, on sait que la variance est égale à l'espérance. Ainsi, un graphique des résidus bruts en fonction des valeurs prédites devrait présenter une forme d'entonnoir avec une variabilité dans les résidus qui est de plus en plus grande pour des valeurs prédites de plus en plus grandes. Il serait alors difficile de juger si cette forme d'entonnoir présente un accroissement de la variance proportionnel à l'espérance, comme attendu avec la loi Poisson, où plutôt un accroissement de la variance proportionnel à, par exemple, le carré de l'espérance, ce qui ne respecterait pas le modèle postulé. En régression linéaire ordinaire, avec la distribution normale, on n'a pas ce problème puisque la variance est postulée constante. Cependant, en régression Poisson, de même qu'en régression logistique, on ne postule pas l'homogénéité de la variance de la variable réponse à cause des propriétés des lois Poisson et binomiale.

Ainsi, avec les modèles linéaires généralisés, il est plus judicieux de valider les postulats du modèle en utilisant des résidus définis de façon à ce que leur variance soit à peu près constante sous l'hypothèse que la distribution choisie soit adéquate. Il existe deux principaux types de résidus :

Résidus de Pearson : Il s'agit des résidus bruts divisés par un facteur stabilisateur de variance :

$$r_{P_i} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}.$$

Le dénominateur $\sqrt{V(\hat{\mu}_i)}$ est la racine carrée de la fonction de variance estimée (voir la définition générale de la statistique d'ajustement de Pearson pour plus de détails concernant $V(\hat{\mu}_i)$). En fait, on constate qu'il s'agit des composantes, non élevées au carré, de la statistique d'ajustement de Pearson :

$$X^2 = \sum_{i=1}^{n} r_{P_i}^2.$$

Si le modèle s'ajuste bien aux données, et si les valeurs prédites $\hat{\mu}_i$ sont suffisamment grosses, ces résidus suivent asymptotiquement une

loi normale, d'espérance nulle et de variance constante, mais pas unitaire. Afin de s'assurer que les valeurs prédites $\hat{\mu}_i$ sont suffisamment grosses pour que cette loi asymptotique soit valide, il est préférable d'utiliser les résidus du modèle ajusté sur les données groupées : r_{P_c}. En régression logistique, c'est même essentiel de travailler avec les données groupées lors de l'étude des résidus.

Il existe une version standardisée de ces résidus, que nous noterons sr_{P_c}, qui sont de variance unitaire. Les détails de cette standardisation peuvent être trouvés dans McCullagh et Nelder (1989, section 12.5), ils ne seront pas présentés ici. Un résidu de Pearson standardisé supérieur à 3 en valeur absolue est donc un signe que la loi postulée pour la composante systématique du modèle n'est pas adéquate.

Résidus de déviance : À la manière des résidus de Pearson, les résidus de déviance sont la racine carrée des composantes de la statistique de déviance :

$$r_{D_i} = signe(y_i - \mu_i)\sqrt{d_i}$$

- avec d_i défini en fonction de la distribution choisie, tel que $\sum_{i=1}^{n} d_i$ soit égale à la déviance D,
- et $signe(y_i - \mu_i)$ qui représente le signe de la différence entre la valeur observée et la valeur prédite ($signe(y_i - \mu_i)$ vaut -1 si $y_i < \mu_i$ et $+1$ si $y_i \geq \mu_i$).

Encore une fois, il est préférable de calculer ces résidus sur les données groupées, afin d'augmenter nos chances que les conditions nécessaires à la validité de la loi asymptotique des résidus soient respectées. Cette loi asymptotique, sous l'hypothèse que le modèle s'ajuste bien, est, comme pour les résidus de Pearson, normale d'espérance nulle et de variance constante, mais pas unitaire.

Il existe aussi une version standardisée des résidus de déviance, notés sr_{D_c}. Ces résidus ont une loi asymptotique $N(0, 1)$ sous l'hypothèse que le modèle s'ajuste bien.

Comment utiliser ces résidus afin de juger de la validité des postulats du modèle ? Plusieurs graphiques et statistiques ont été proposés pour ce faire.

Nous présentons ici seulement les outils classiques.

Graphique des résidus de Pearson ou de déviance en fonction des valeurs prédites : On ne doit pas y voir de forme d'entonnoir, car si le modèle s'ajuste bien ces résidus devraient être de variance homogène.

Étude de la normalité des résidus de Pearson ou de déviance :
Comme en régression classique, si le modèle est adéquat, les résidus devraient être normaux, et ce, même si la loi postulée pour Y n'est pas la loi normale. On peut donc utiliser les outils que l'on connaît déjà pour tester la normalité des résidus de Pearson ou de déviance. Par exemple, on peut utiliser des graphiques (histogramme, boxplot, droite de Henry, etc.) ou des statistiques (coefficients d'asymétrie et d'aplatissement, test de Shapiro-Wilk, etc.).

Nous verrons à la section 4.9 quoi faire si on doute que les postulats du modèle soient respectés.

4.8 Correction pour sur ou sous dispersion

En régression Poisson et logistique, le paramètre de dispersion ϕ vaut en théorie 1. Il arrive cependant que les données présentent plus ou moins de dispersion, en d'autres mots de variabilité, que ce que la distribution suppose. On peut détecter de la sur ou de la sous dispersion en utilisant le ratio de la déviance ou de la statistique X^2 de Pearson sur ses degrés de liberté ($D/(C-(p+1))$ ou $X^2/(C-(p+1))$). Si celui-ci est nettement plus grand que 1, c'est un signe de sur dispersion. S'il est nettement inférieur à 1, c'est un signe de sous dispersion.

Un cas fréquent de variabilité non conforme au modèle en régression Poisson est ce qu'on appelle de la « dispersion extra poissonnienne ». Il s'agit de données pour lesquelles la variance, qui devrait être environ égale à l'espérance selon la loi Poisson, tend en fait à être plus grande que l'espérance.

En présence de sur ou de sous dispersion, les inférences sur les coefficients ne sont plus fiables, car la matrice de variance-covariance du vecteur des estimations des paramètres est mal estimée. Il existe cependant une façon simple de corriger les inférences. On va utiliser une estimation du paramètre de dispersion ϕ pour faire des corrections. On peut utiliser :

$$\hat{\phi} = D/(C-(p+1)) \qquad \text{ou} \qquad \hat{\phi} = X^2/(C-(p+1)).$$

Mieux ce paramètre est estimé, plus la correction sera juste. Il est donc préférable de l'estimer à partir des données groupées.

La matrice corrigée de variance-covariance du vecteur des estimations des paramètres est :

$$\hat{\sigma}^2(\hat{\boldsymbol{\beta}})_{corr} = \hat{\sigma}^2(\hat{\boldsymbol{\beta}})\hat{\phi}.$$

Il en découle que la statistique corrigée du test de Wald sur un paramètre est :

$$Z_{w,corr} = \frac{Z_w}{\sqrt{\hat{\phi}}} = \frac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)\sqrt{\hat{\phi}}}.$$

On peut aussi corriger la statistique du test de rapport de vraisemblance sur plusieurs paramètres. La statistique corrigée est :

$$LR_{corr} = \frac{LR}{\hat{\phi}}.$$

4.9 Étapes d'une analyse de données avec un GLM

Voici une suggestion d'étapes à suivre dans une analyse de données avec un modèle linéaire généralisé.

1. **Analyse exploratoire** :

   Comme dans toute analyse statistique, il est bon de produire quelques statistiques descriptives sur les données à analyser afin de mieux connaître la nature des variables.

2. **Détermination de la composante aléatoire du modèle** :

   Il faut d'abord sélectionner une distribution pour la variable réponse en fonction de la nature de celle-ci. L'étape précédente est utile pour faire un choix judicieux.

3. **Détermination de la fonction de lien** :

   Une fonction de lien doit être choisie. Certains conseillent de prendre initialement la fonction de lien canonique[1] de la distribution choisie et de la modifier uniquement si ce changement permet de régler un problème décelé lors de la validation du modèle. Rappelons quelques avantages de la fonction de lien canonique :
   - les valeurs prédites sont dans l'intervalle des valeurs possibles de la variable réponse ;
   - les paramètres ont une interprétation intéressante ;
   - l'algorithme d'estimation des paramètres est plus simple, donc on risque moins de rencontrer des problèmes numériques.

   On pourrait aussi choisir la même fonction de lien que celle utilisée dans une autre étude à laquelle on souhaite comparer nos résultats.

   Si on souhaite baser le choix de la fonction de lien sur un critère statistique, on pourrait prendre un modèle incluant toutes les variables

1. identité pour la loi normale, ln pour la loi Poisson, logit pour la loi binomiale

explicatives (sans interactions), ajuster le modèle avec toutes les fonctions de lien concurrentes et ensuite baser son choix sur des graphiques et des statistiques en recherchant le meilleur ajustement possible aux données. Des graphiques utiles ici sont les mêmes que ceux permettant de valider la justesse d'une fonction de lien. Il s'agit donc des graphiques suivants :

– les graphiques des valeurs observées en fonction des valeurs prédites de la variable réponse pour toutes les fonctions de lien : il devrait être le plus linéaire possible ;
– le graphique des valeurs observées de la variable réponse en fonction d'une variable explicative numérique auquel on ajoute les *courbes* de régression pour les différentes fonctions de lien : la *courbe* de régression doit se rapprocher le plus possible des valeurs observées.

Pour une variable explicative catégorique, on voudrait que les moyennes des valeurs observées par modalité soient les plus similaires possible aux moyennes des valeurs prédites.

4. **Détermination de la composante systématique du modèle** :

   On doit ensuite écrire la composante systématique du modèle. Si on a des observations pour plusieurs variables explicatives, il est bien de suivre une procédure pour guider la sélection des termes à inclure dans le modèle. Voici une suggestion de procédure inspirée de Hosmer et Lemeshow (2000) :

   (a) Tester si chacune des p variables est significative (à l'aide de tests de rapport de vraisemblance) à un seuil très élevé (ex. : 25% ou 30%) en ajustant p modèles ne contenant qu'une seule variable.

   (b) Mettre toutes les variables significatives de l'étape précédente dans un même modèle et les éliminer une à une (à l'aide de tests de rapport de vraisemblances) jusqu'à ce que toutes les variables restantes soient significatives au seuil choisi (par exemple 5%). Il s'agit d'une procédure de sélection « backward » (élimination descendante).

   (c) Essayer d'ajouter des termes plus complexes (ex. : interactions, variables élevées au carré, etc.) pour voir s'ils seraient significatifs. À cette étape, certains choisissent de travailler avec un seuil encore

plus petit qu'à l'étape précédente (par exemple 1%), car on ne désire pas avoir de termes complexes dans le modèle. On le veut le plus simple possible.

On cherche parfois ensuite à raffiner la façon d'inclure une variable dans le modèle. Pour une variable catégorique, on peut parfois regrouper des modalités. On se demande aussi parfois si on va inclure une variable numérique à peu de modalités sous sa forme numérique ou encore sous une forme catégorique.

Notez que si le nombre de variables explicatives au départ est assez petit, vous pouvez sauter l'étape (a) et commencer directement par un modèle très complexe. L'étape (a) comporte des avantages et des inconvénients. On pourrait appeler les modèles simples des « modèles marginaux ». Les autres variables explicatives sont peut-être des variables confondantes pour la relation entre une variable explicative x_j et la variable réponse Y. Le modèle simple ne corrige pas pour ces effets confondants. Ainsi, une variable pourrait ne pas avoir d'effet marginalement, mais avoir en réalité un effet si on l'étudie conditionnellement aux valeurs des autres variables. On avait beaucoup discuté de ce sujet dans la section du cours sur les tableaux de fréquences à trois variables. C'est ce qui explique l'utilisation d'un seuil élevé à cette étape. On souhaite faire un premier tri très grossier sans laisser tomber des variables importantes.

Un avantage cependant des modèles simples est qu'ils ne peuvent pas présenter de multicolinéarité. La multicolinéarité est un phénomène qui rend instables les estimations des paramètres en raison d'une trop forte relation entre des variables explicatives. Par exemple, la multicolinéarité peut avoir pour conséquence que deux variables sont non significatives si on les met simultanément dans le modèle, alors que prises séparément elles sont toutes deux très significatives. C'est la raison pour laquelle il est préférable de retirer les variables du modèle une à une.

Notez aussi que puisque la formation de groupes dans les données dépend des variables explicatives incluses dans le modèle, il est plus simple à cette étape de travailler sur les données une ligne par individus, et non les données groupées.

5. **Validation du modèle** :

   Avant d'utiliser un modèle pour faire de l'inférence, il faut le valider pour s'assurer de la qualité de nos résultats. Rappelons qu'il est préférable de faire cette étape sur les données groupées. Si on décèle un problème lors de cette étape, on peut soit :

   (a) changer la distribution postulée pour la variable aléatoire (et refaire les étapes subséquentes) ;

   (b) changer la fonction de lien (et refaire les étapes subséquentes) ;

   (c) changer la composante systématique du modèle, par exemple en enlevant un terme, en ajoutant une variable, en changeant le format d'une variable, etc. (et revalider le modèle) ;

   (d) corriger les résultats pour tenir compte d'une sur ou sous dispersion (s'applique uniquement aux modèles avec variable réponse binomiale ou Poisson).

   Il n'est pas rare de se retrouver à essayer toutes sortes de solutions potentielles, et donc de refaire plusieurs fois les étapes de l'analyse, jusqu'à l'obtention d'un modèle satisfaisant.

6. **Inférence et/ou prédiction** :

   Lorsqu'on a en main un modèle que l'on considère valide, on peut faire de l'inférence et/ou de la prédiction. L'inférence permet de décrire des liens entre les variables explicatives et la variable réponse.

   – Si une variable explicative a été complètement retirée du modèle lors du choix de la composante systématique, c'est qu'il n'y a pas de lien significatif entre elle et la variable réponse. Par contre, il est aussi possible qu'une variable ait été retirée du modèle parce qu'elle causait un problème de multicolinéarité. Dans ce cas, il est possible

qu'elle soit reliée à la variable réponse, mais qu'une autre variable avec laquelle elle est très corrélée le soit encore plus.

– Si un lien est significatif, il faut évaluer sa direction et, si possible, le quantifier. C'est l'interprétation des valeurs des paramètres, et des comparaisons multiples pour les variables explicatives catégoriques, qui permettent de décrire les liens.

4.10 Régression logistique pour une variable réponse polytomique

Jusqu'à maintenant, on a parlé du traitement d'une variable réponse binaire par la régression logistique et d'une variable réponse représentant un dénombrement par la régression Poisson. Que faire d'une variable réponse catégorique ayant plus de deux modalités possibles ? Quelques modèles de régression logistique ont été développés pour traiter ce type de variable réponse. On parle parfois de modèles de régression logistique polytomique ou encore multinomiale. Cependant, il n'existe pas de consensus sur la terminologie dans ce domaine. Ainsi, pour certains, la régression logistique multinomiale réfère à toute forme de régression logistique avec une variable réponse multicatégorielle, c'est-à-dire polytomique. Pour d'autres, la régression logistique multinomiale est le modèle logit généralisé présenté ci-dessous.

Dans cette section, deux modèles seront présentés sans être approfondis. Pour plus de détails, une bonne référence est Agresti (2002, chapitre 7).

4.10.1 Réponse nominale : modèle logit généralisé

Pour modéliser une variable réponse polytomique nominale, c'est-à-dire avec des modalités ne pouvant pas être ordonnées, un modèle simple et courant est le modèle logit généralisé. Ce modèle est nommé en anglais par certains « baseline-category logit model ».

Notons Y la variable réponse ayant J modalités. Nous voulons modéliser les probabilités d'occurrence des différentes modalités. Soit

$$\pi_k(\boldsymbol{x}) = P(Y = k \mid \boldsymbol{x}), \qquad \text{pour } k = 1, \ldots, J,$$

la probabilité que la variable réponse Y soit égale à sa modalité k sachant que le vecteur des variables explicatives prend la valeur $\boldsymbol{x}$. On a bien sûr que $\sum_{k=1}^{J} \pi_k(\boldsymbol{x}) = 1$. On doit choisir une des modalités de la variable comme étant la modalité de référence. Disons que l'on prend la dernière modalité, J, comme modalité de référence (baseline). Chacune des autres modalités est associée à la modalité de référence pour former le ratio $\pi_k(\boldsymbol{x})/\pi_J(\boldsymbol{x})$, et ce, pour $k = 1, \ldots, J-1$. En régression logistique pour variable réponse binaire,

il n'y a qu'un seul ratio, soit $\frac{P(Y=1 \mid \boldsymbol{x})}{P(Y=0 \mid \boldsymbol{x})} = \frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}$. On a maintenant $J-1$ ratio, à partir desquelles on définit $J-1$ modèles :

$$\ln\left(\frac{\pi_k(\boldsymbol{x})}{\pi_J(\boldsymbol{x})}\right) = \boldsymbol{x}^t\boldsymbol{\beta}_k \qquad \text{pour } k = 1, \ldots, J-1. \tag{4.1}$$

On voit apparaître ici une définition généralisée de la fonction de lien logit, d'où le nom du modèle. Ce n'est plus simplement le logarithme naturel du ratio entre la probabilité de succès par rapport à son complément, c'est maintenant le logarithme naturel du ratio entre la probabilité d'occurrence d'une modalité par rapport à la probabilité d'occurrence de la modalité de référence. Notez que les $J-1$ équations déterminent en fait le logit de toute paire de deux probabilités $\pi_k(\boldsymbol{x})$ et $\pi_{k'}(\boldsymbol{x})$ puisque

$$\ln\left(\frac{\pi_k(\boldsymbol{x})}{\pi_{k'}(\boldsymbol{x})}\right) = \ln\left(\frac{\pi_k(\boldsymbol{x})}{\pi_J(\boldsymbol{x})}\right) - \ln\left(\frac{\pi_{k'}(\boldsymbol{x})}{\pi_J(\boldsymbol{x})}\right).$$

L'équation 4.1 représente en fait $J-1$ modèles de régression logistique binaire. On va les ajuster simultanément et non individuellement. Les estimations du maximum de vraisemblance des paramètres et leurs erreurs-types devraient être similaires que l'on ajuste simultanément ou non les modèles. Cependant, un ajustement simultané des équations permet d'effectuer facilement un test global du lien entre une variable explicative x_j et la variable réponse Y. L'hypothèse nulle $H_0 : \beta_{j,1} = \ldots = \beta_{j,J-1} = 0$ représente une absence d'association. Cette hypothèse fait intervenir au moins un β de chacune des $J-1$ équations.

Mentionnons finalement que les valeurs prédites par ce modèle sont les suivantes :

$$\hat{\pi}_k(\boldsymbol{x}) = \frac{exp(\boldsymbol{x}^t\hat{\boldsymbol{\beta}}_k)}{1 + \sum_{h=1}^{J-1} exp(\boldsymbol{x}^t\hat{\boldsymbol{\beta}}_h)}.$$

4.10.2 Réponse ordinale : modèle à rapports de cotes proportionnels

On a vu par le passé que lorsque les modalités d'une variable polytomique sont ordonnables, utiliser ce caractère ordinal permet de construire des tests plus puissants sur cette variable. Pour cette raison, des modèles de régression logistique ont été élaborés pour modéliser des variables réponses catégoriques ordinales. Le plus connu de ces modèles est probablement le modèle à rapports de cotes proportionnels, en anglais « proportional odds model ».

La fonction de lien de ce modèle est appelée « logit cumulatif ». Elle est ainsi définie :

$$\begin{aligned} logit(P(Y \leq k \mid \boldsymbol{x})) &= \ln\left(\frac{P(Y \leq k \mid \boldsymbol{x})}{1 - P(Y \leq k \mid \boldsymbol{x})}\right) \\ &= \ln\left(\frac{\pi_1(\boldsymbol{x}) + \cdots + \pi_k(\boldsymbol{x})}{\pi_{k+1}(\boldsymbol{x}) + \cdots + \pi_J(\boldsymbol{x})}\right), \end{aligned}$$

en supposant que les modalités 1 à J sont ordonnées de la plus petite à la plus grande. Un logit cumulatif est défini pour toute modalité $k = 1, \ldots, J-1$. Remarquez qu'un modèle pour $logit(P(Y \leq k \mid \boldsymbol{x}))$ est en fait un modèle de régression logistique avec une variable réponse binaire pour laquelle une modalité est formée par les modalités 1 à k de Y et l'autre modalité est formée des modalités $k+1$ à J de Y.

Le modèle à rapports de cotes proportionnels est le suivant :

$$logit(P(Y \leq k \mid \boldsymbol{x})) = \beta_{0_k} + \boldsymbol{x}_{-0}^t \boldsymbol{\beta}_{-0} \qquad \text{pour } k = 1, \ldots, J-1$$

où $\boldsymbol{x}_{-0}$ est le vecteur $\boldsymbol{x}$ auquel on a enlevé le premier élément unitaire et $\boldsymbol{\beta}_{-0}$ est le vecteur des paramètres $\boldsymbol{\beta}$ excluant l'ordonnée à l'origine.

Il est important de remarquer que ce modèle comporte une ordonnée à l'origine distincte pour chaque modalité $k = 1, \ldots, J-1$, mais que les paramètres multipliant les variables explicatives (ceux composant le vecteur $\boldsymbol{\beta}_{-0}$) sont communs à toutes les modalités. C'est l'hypothèse que ces paramètres ne varient pas selon la modalité qui rend les rapports de cotes proportionnels. En pratique, avant d'utiliser ce modèle pour analyser des données, il faut s'assurer que l'hypothèse de proportionnalité des rapports de cotes est

rencontrée. Si cette hypothèse n'est pas rencontrée, on peut construire un modèle à rapports de cotes partiellement proportionnels, en anglais « partial proportional odds model ». Dans ce modèle, on permet à certains paramètres devant des variables explicatives de varier selon la modalité de Y modélisée.

Les valeurs prédites par ce modèle sont des probabilités cumulatives :

$$\hat{P}(Y \leq k \mid \boldsymbol{x}) = \frac{exp(\hat{\beta}_{0_k} + \boldsymbol{x}_{-0}^t \hat{\boldsymbol{\beta}}_{-0})}{1 + exp(\hat{\beta}_{0_k} + \boldsymbol{x}_{-0}^t \hat{\boldsymbol{\beta}}_{-0})}.$$

Pour estimer la probabilité que Y soit égale à une modalité, il faut utiliser le fait que :

$$\pi_k(\boldsymbol{x}) = P(Y = k \mid \boldsymbol{x}) = P(Y \leq k \mid \boldsymbol{x}) - P(Y \leq k-1 \mid \boldsymbol{x}).$$

4.11 Régression logistique conditionnelle

Si le temps le permettait, il serait intéressant de présenter dans ce cours la régression logistique conditionnelle. Il s'agit d'une régression logistique dans laquelle la vraisemblance est conditionnelle à l'appartenance des individus à des strates. Ces strates sont formées à partir de certaines variables de nuisance. L'écriture conditionnelle de la vraisemblance permet d'omettre l'estimation de paramètres pour les variables de nuisance tout en corrigeant les effets des variables explicatives d'intérêt pour les effets des variables de nuisance. Cette méthode est particulièrement utilisée pour analyser des données provenant d'étude cas-témoin avec appariement. Pour en savoir plus, une bonne référence est Hosmer et Lemeshow (2000, chapitre 7).

4.12 Notes complémentaires

4.12.1 Modélisation de taux : régression Poisson avec variable offset

La régression Poisson avec variable offset permet de modéliser des taux d'occurrence de certains événements.

Exemple de modélisation de taux : les accidents d'auto

Pour étudier la relation entre le sexe et le risque d'accident d'auto, nous avons des données concernant 16262 personnes âgées de 65 à 84 ans pour une période de 4 ans.

Question : Les hommes sont-ils plus sujets aux accidents que les femmes ?

Les données sont les suivantes :

Sexe	nombre d'accidents	nombre d'années-personnes
Femme	175	17.30
Homme	320	21.40

où nombre d'années-personnes = somme des temps d'observation (temps de conduite de son véhicule) pour toutes les femmes et les hommes ayant participé à l'étude (unité de mesure = mille années).

Le taux d'occurrence d'un événement se calcule en divisant le nombre d'occurrences de l'événement, représenté par variable la Y, par rapport à une certaine mesure de taille t. Si on cherche à étudier le lien entre le taux Y/t et des variables explicatives potentielles, on peut utiliser le modèle linéaire généralisé avec lien logarithmique suivant :

$$\ln\left(E\left(\frac{Y}{t}|\boldsymbol{x}\right)\right) = \ln\left(\frac{E(Y|\boldsymbol{x})}{t}\right) = \ln\left(\frac{\mu(\boldsymbol{x})}{t}\right) = \boldsymbol{x}^t\boldsymbol{\beta}.$$

La mesure de taille est considérée fixe, comme les variables explicatives. Grâce au lien logarithmique, on peut transférer t de la composante aléatoire du

modèle vers la composante systématique, où elle devient, sur l'échelle logarithmique, une variable offset :

$$\begin{aligned} \ln\left(\frac{\mu(\boldsymbol{x})}{t}\right) &= \boldsymbol{x}^t\boldsymbol{\beta} \\ \ln(\mu(\boldsymbol{x})) - \ln(t) &= \boldsymbol{x}^t\boldsymbol{\beta} \\ \ln(\mu(\boldsymbol{x})) &= \ln(t) + \boldsymbol{x}^t\boldsymbol{\beta}. \end{aligned}$$

Étant donné que $\mu(\boldsymbol{x}) = E(Y|\boldsymbol{x})$ et que Y est un dénombrement, on peut supposer que Y suit une loi Poisson et faire de ce modèle une régression Poisson avec variable offset $\ln(t)$.

Selon ce modèle, le nombre espéré d'occurrences de l'événement étudié est :

$$\mu(\boldsymbol{x}) = t\exp(\boldsymbol{x}^t\boldsymbol{\beta}).$$

Ainsi, l'espérance $\mu(\boldsymbol{x})$ est proportionnelle à la mesure de taille t.

4.12.2 Comparaison d'un GLM avec un modèle linéaire classique sur une variable réponse transformée

Avec un modèle linéaire classique (ex. : régression linéaire, ANOVA), lorsque les postulats du modèle sont violés, il arrive que l'on transforme la variable réponse. Est-ce qu'ajuster un modèle linéaire classique sur $g(Y)$ est équivalent à ajuster un modèle linéaire généralisé avec variable réponse Y et fonction de lien $g()$?

Soit $Z = g(Y)$ la variable réponse transformée. Le modèle classique suppose que cette variable transformée suit une loi normale $N(\mu_Z(\boldsymbol{x}), \sigma^2)$. On modélise donc l'espérance de Z en fonction des variables explicatives par $\mu_Z(\boldsymbol{x}) = \boldsymbol{x}^t\boldsymbol{\beta}$. Pour sa part, le modèle linéaire généralisé postule une loi pour Y et non pour $g(Y)$. Il modélise l'espérance de Y en fonction des variables explicatives ($g(\mu_Y(\boldsymbol{x})) = \boldsymbol{x}^t\boldsymbol{\beta}$), et non l'espérance de Z. À la base, les deux approches proposent donc des modèles différents. Les fonctions de vraisemblance des modèles ne sont pas tout à fait les mêmes, et, en conséquence, les estimations par maximum de vraisemblance des paramètres ne sont pas identiques, même si les composantes systématiques des modèles sont les mêmes.

Si on se penche sur les valeurs prédites par les modèles, on voit aussi une différence importante. Avec le modèle classique ajusté sur la variable transformée, on doit faire une retransformation sur les valeurs prédites par le modèle pour se ramener sur l'échelle de la variable réponse d'origine. En effet, on veut en réalité estimer $\mu_Y(\boldsymbol{x})$ alors que le modèle fournit des estimations de $\mu_Z(\boldsymbol{x})$. Cependant, on a que :

$$\mu_Y(\boldsymbol{x}) = E(Y|\boldsymbol{x}) = E(g^{-1}(Z|\boldsymbol{x}))$$

ce qui n'est pas égal à $g^{-1}(E(Z|\boldsymbol{x})) = g^{-1}(\mu_Z(\boldsymbol{x})) = g^{-1}(\boldsymbol{x}^t\boldsymbol{\beta})$ pour une fonction de lien g autre que l'identité. Ainsi, l'estimation de $\mu_Y(\boldsymbol{x})$ par $g^{-1}(\boldsymbol{x}^t\hat{\boldsymbol{\beta}})$ est biaisée.

Par exemple, avec la fonction de lien logarithmique on a :

$$\begin{aligned}
\mu_Y(\boldsymbol{x}) &= E(\exp(Z|\boldsymbol{x})) \qquad \text{où } Z \sim N(\mu_Z(\boldsymbol{x}), \sigma^2) \\
&= \exp(\mu_Z(\boldsymbol{x}) + \sigma^2/2) \qquad \text{de la f.g.m. d'une normale} \\
&= \exp(\mu_Z(\boldsymbol{x}))\exp(\sigma^2/2)
\end{aligned}$$

Ainsi, si on veut estimer sans biais $\mu_Y(\boldsymbol{x})$ avec cette fonction de lien, il faut corriger la transformation inverse de $\mu_Z(\boldsymbol{x})$ par le facteur correctif $\exp(\sigma^2/2)$. D'autres facteurs correctifs ont été proposés (Duan, 1983).

Cependant, avec un GLM, les valeurs prédites sont non biaisées, car :

$$\mu_Y(\boldsymbol{x}) = E(Y|\boldsymbol{x}) = g^{-1}(\boldsymbol{x}^t\boldsymbol{\beta}).$$

En plus de cet avantage important, le travail avec un GLM plutôt qu'un modèle linéaire classique sur une variable réponse transformée comporte les points positifs suivants :

- Avec un GLM, on est plus attentif à correctement interpréter les paramètres en tenant compte de la fonction de lien. Par exemple, avec un modèle classique sur le logarithme de Y, est-ce que l'on pense à interpréter les paramètres en terme d'effet multiplicatif ?

- Il arrive que des valeurs de $g(y_i)$ ne soient pas définies (par exemple $\ln(0)$) et compliquent la transformation de la variable réponse. Pour ajuster un GLM, on n'a pas besoin de calculer $g(y_i)$, on n'a donc pas besoin de se soucier de ces valeurs problématiques.

4.13 Résumé des formules concernant les modèles linéaires généralisés

Écriture générale d'un modèle linéaire généralisé selon les trois notations utilisées dans le cours :

Par individu :
$Y_i|\boldsymbol{x}_i \overset{\text{ind.}}{\longrightarrow} \mathcal{L}(\mu_i, \phi)$ avec $g(\mu_i) = \boldsymbol{x}_i^t\boldsymbol{\beta}$ pour $i = 1, \ldots, n$
Par groupe (vecteurs $\boldsymbol{x}$ distincts) :
$Y|\boldsymbol{x}_c \overset{\text{ind.}}{\longrightarrow} \mathcal{L}(\mu_c, \phi)$ avec $g(\mu_c) = \boldsymbol{x}_c^t\boldsymbol{\beta}$ pour $c = 1, \ldots, C$
Sans référence aux observations :
$Y|\boldsymbol{x} \overset{\text{ind.}}{\longrightarrow} \mathcal{L}(\mu(\boldsymbol{x}), \phi)$ avec $g(\mu(\boldsymbol{x})) = \boldsymbol{x}^t\boldsymbol{\beta}$
où

- $\mathcal{L}$ est n'importe quelle distribution de la famille exponentielle ;
- $\mu(\boldsymbol{x}) = E(Y|\boldsymbol{x})$ est l'espérance de Y sachant $\boldsymbol{x}$;
- ϕ est un paramètre de dispersion pour la distribution $\mathcal{L}$ ($\phi = 1$ pour les distributions Poisson et binomiale) ;
- $\boldsymbol{x}^t = (1, x_1, \ldots, x_p)$ est le vecteur des variables explicatives ;
- $\boldsymbol{\beta}^t = (\beta_0, \beta_1, \ldots, \beta_p)$ est un vecteur de paramètres, comprenant une ordonnée à l'origine β_0.

Ce modèle comporte 3 composantes :

composante aléatoire : La variable réponse Y pour laquelle on postule une distribution sachant $\boldsymbol{x}$.

composante systématique : $\boldsymbol{x}^t\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ où

- $\boldsymbol{\beta}$ est le vecteur des paramètres à estimer ;
- les x_j sont des termes considérés fixes représentant
  soit des variables explicatives numériques,
  soit des variables de design (souvent des indicatrices) qui permettent d'inclure des variables explicatives catégoriques dans le modèle,
  soit des termes plus complexes tels que des interactions.

  Note : une variable x_j dont le coefficient β_j est fixé à 1 est appelée **variable offset**.

fonction de lien : Une fonction mathématique $g()$ qui relie l'espérance de Y à la composante systématique. Exemples :

- identité : $\mu(\boldsymbol{x}) = \boldsymbol{x}^t\boldsymbol{\beta}$
- logarithme : $\ln(\mu(\boldsymbol{x})) = \boldsymbol{x}^t\boldsymbol{\beta}$
- logit : $logit(\mu(\boldsymbol{x})) = \log\left(\frac{\mu(\boldsymbol{x})}{1-\mu(\boldsymbol{x})}\right) = \boldsymbol{x}^t\boldsymbol{\beta}$
- probit : $\Phi^{-1}(\mu(\boldsymbol{x})) = \boldsymbol{x}^t\boldsymbol{\beta}$ où Φ est la fonction de répartition d'une $N(0,1)$
- log-log : $\ln(-\ln(1-\mu(\boldsymbol{x}))) = \boldsymbol{x}^t\boldsymbol{\beta}$

Régression Poisson :
Variable réponse = dénombrement (nombre entier non négatif)
On postule $Y|\boldsymbol{x} \xrightarrow{\text{ind.}} Poisson(\mu(\boldsymbol{x}))$.
La fonction de lien canonique est le logarithme naturel.

Régression logistique :
Variable réponse = variable binaire (succès ou échec)
On postule $Y|\boldsymbol{x} \xrightarrow{\text{ind.}} Bernoulli(\pi(\boldsymbol{x}))$
où $\pi(\boldsymbol{x}) = \mu(\boldsymbol{x}) = E(Y|\boldsymbol{x}) = P(Y = \text{succès} \mid \boldsymbol{x})$.
La fonction de lien canonique est le logit.

Interprétation des paramètres

Variable explicative catégorique x_{catego} à k modalités :

On doit inclure dans le modèle $k-1$ variables numériques pour la représenter. Celles-ci sont souvent des indicatrices de chacune des modalités (valeur de 1 si x_{catego} prend la modalité en question, 0 sinon), sauf une modalité, dite de référence (mais d'autres paramétrisations existent).
Pas de lien entre x_{catego} et $Y \Leftrightarrow$ tous les paramètres β devant les variables numériques du modèle représentant la variable explicative catégorique (variables de design) sont nuls.

Variable explicative numérique x_j (incluant une indicatrice) :

L'interprétation du coefficient devant x_j dans le modèle, noté β_j, dépend de la fonction de lien $g()$:
- Toute fonction de lien monotone croissante :
  β_j positif = association positive entre x_j et Y ($\uparrow$ de $x_j \Rightarrow \uparrow$ de Y)
  β_j négatif = association négative entre x_j et Y ($\uparrow$ de $x_j \Rightarrow \downarrow$ de Y)

- Lien identité : $\mu(x_j + 1) = \mu(x_j) + \beta_j$
  (augmentation d'une unité de $x_j \Rightarrow$ addition de β_j à la valeur de μ)
- Lien logarithmique : $\mu(x_j + 1) = \exp(\beta_j)\mu(x_j)$
  (augmentation d'une unité de $x_j \Rightarrow$ multiplication par e^{β_j} de la valeur de μ)
- Lien logit : $\frac{\pi(x_j+1)}{1-\pi(x_j+1)} = e^{\beta_j}\frac{\pi(x_j)}{1-\pi(x_j)}$
  (augmentation d'une unité de $x_j \Rightarrow$ multiplication par e^{β_j} de la cote de la probabilité π)
  Si x_j est une indicatrice : $e^{\beta_j} = \frac{\pi(x_j=1)/(1-\pi(x_j=1))}{\pi(x_j=0)/(1-\pi(x_j=0))}$ = rapport de cote de la probabilité π pour les individus pour lesquels $x_j = 1$ par rapport aux individus pour lesquels $x_j = 0$.

Interaction : Supposons le modèle $g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$. Le paramètre β_{12} devant l'interaction double $x_1 x_2$ représente l'effet de x_2 sur la relation entre Y et x_1. Si ce paramètre prend une valeur nulle, c'est que l'association entre Y et x_1 n'est pas influencée pas x_2.

Ajustement du modèle : estimation des paramètres par maximum de vraisemblance

Définitions :

- **Vraisemblance** pour une distribution discrète :
  $L(\boldsymbol{\beta}) = P(Y_1 = y_1, \ldots, Y_n = y_n) = \prod_{i=1}^{n} P(Y_i = y_i)$ du postulat d'indépendance entre les observations ;
- **Vecteur des dérivées premières de la log vraisemblance** (fonction score) : $S(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \ln L(\boldsymbol{\beta})$;
- **Matrice d'information observée** : $I_o(\boldsymbol{\beta}) = -Hessien(\boldsymbol{\beta}) = -\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\frac{\partial^2}{\partial \boldsymbol{\beta_j} \boldsymbol{\beta_{j'}}} \ln L(\boldsymbol{\beta})$;
- **Matrice d'information espérée** (ou de Fisher) : $I_e(\boldsymbol{\beta}) = E[I_o(\boldsymbol{\beta})] = -E\left[\frac{\partial^2}{\partial \boldsymbol{\beta_j} \boldsymbol{\beta_{j'}}} \ln L(\boldsymbol{\beta})\right]$.

Formules pour la régression simple Poisson avec lien logarithmique ou identité et la régression simple logistique avec lien logit :

	rég. Poisson lien ln	rég. Poisson lien identité	rég. logistique lien logit
$\ln L(\boldsymbol{\beta})$	$\sum_{i=1}^{n} (y_i \ln(\mu_i) - \mu_i - \ln(y_i!))$ où $\mu_i = exp(\beta_0 + \beta_1 x_i)$	où $\mu_i = \beta_0 + \beta_1 x_i$	$\sum_{i=1}^{n} (y_i \ln(\pi_i) + (1-y_i)\ln(1-\pi_i))$ où $\pi_i = \frac{exp(\beta_0+\beta_1 x_i)}{1+exp(\beta_0+\beta_1 x_i)}$
$S(\boldsymbol{\beta})$	$\sum_{i=1}^{n} \begin{pmatrix} (y_i - \mu_i) \\ x_i(y_i - \mu_i) \end{pmatrix}$	$\sum_{i=1}^{n} \begin{pmatrix} (y_i/\mu_i - 1) \\ x_i(y_i/\mu_i - 1) \end{pmatrix}$	$\sum_{i=1}^{n} \begin{pmatrix} (y_i - \pi_i) \\ x_i(y_i - \pi_i) \end{pmatrix}$
$I_o(\boldsymbol{\beta})$	$\sum_{i=1}^{n} \mu_i \begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \end{pmatrix}$	$\sum_{i=1}^{n} \frac{y_i}{\mu_i^2} \begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \end{pmatrix}$	$\sum_{i=1}^{n} \pi_i(1-\pi_i) \begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \end{pmatrix}$
$I_e(\boldsymbol{\beta})$	$\sum_{i=1}^{n} \mu_i \begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \end{pmatrix}$	$\sum_{i=1}^{n} \frac{1}{\mu_i} \begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \end{pmatrix}$	$\sum_{i=1}^{n} \pi_i(1-\pi_i) \begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \end{pmatrix}$

Si la fonction de lien du modèle est la fonction de lien canonique pour la distribution de la famille exponentielle postulée, les matrices d'information espérée et observée sont égales.

La maximisation de la vraisemblance revient à trouver la solution pour $\boldsymbol{\beta}$ de $S(\boldsymbol{\beta}) = 0$. Cette maximisation est mise en oeuvre à l'aide de l'**algorithme numérique Newton-Raphson** (utilise $I_o(\boldsymbol{\beta})$) ou **Fisher Scoring** (utilise $I_e(\boldsymbol{\beta})$). Ces algorithmes itératifs mettent à jour la valeur des paramètres à l'aide de la formule : $\boldsymbol{\beta}^{iter+1} = \boldsymbol{\beta}^{iter} + I_o^{-1}(\boldsymbol{\beta}^{iter}) S(\boldsymbol{\beta}^{iter})$.

Données groupées

- Un groupe = toutes les observations ayant les mêmes valeurs des variables explicatives dans la composante systématique du modèle (même valeur observée pour le vecteur $\boldsymbol{x}$).
- C = nombre de lignes du jeu de données groupées = nombre de groupes = nombre de configurations différentes des variables explicatives qui ont été observées.
- Nouvelle variable explicative modélisée : $Y_c^* = \sum_{\{i \in \text{ grp } c\}} Y_i$
  où Y est la variable réponse d'origine (données non groupées). La somme est faite sur toutes les observations d'un même groupe. La distribution postulée pour cette nouvelle variable est :

Régression logistique :
$Y_c^*|x_c \sim Binomiale(n_c^*, \pi_c)$ où $\pi_c = P(Y = \text{succès} \mid \boldsymbol{x}_c)$;

Régression Poisson :
$Y_c^*|x_c \sim Poisson(\mu_c^*)$ où $\mu_c^* = n_c^* \mu_c$ et $\mu_c = E(Y|\boldsymbol{x}_c)$;

où n_c^* est le nombre d'individus dans le groupe c, pour $c = 1, \ldots, C$.

– On doit tenir compte des n_c^* dans les modèles ajustés sur ces données groupées. Cependant, on ne tient pas compte des n_c^* de la même façon pour les deux types de régression vues dans le cours.

Régression logistique : On pondère les observations Y_c^* par $1/n_c^*$. On a que $E[Y_c^*/n_c^*] = n_c^* \pi_c / n_c^* = \pi_c$. Le modèle demeure donc le même que sur les données non groupées : $g(\pi_c) = \boldsymbol{x}_c^t \boldsymbol{\beta}$.

Régression Poisson : On n'utilise pas l'approche de la pondération. Selon la fonction de lien, on ajuste sur la nouvelle variable réponse groupée Y^* une régression Poisson avec une composante systématique ajustée par rapport à celle du modèle sur les données non groupées ($g(\mu_c) = \boldsymbol{x}_c^t \boldsymbol{\beta}$) :

Lien identité : $g(\mu_c^*) = \mu_c^* = n_c^* \mu_c = n_c^* \boldsymbol{x}_c^t \boldsymbol{\beta}$.
Dans ce modèle, les variables explicatives ne sont plus $\boldsymbol{x}^t = (1, x_1, \ldots, x_p)$. Elles deviennent $(n^*, n^* \times x_1, \ldots, n^* \times x_p)$. Le modèle ne comporte plus d'ordonnée à l'origine.

Lien logarithmique : $g(\mu_c^*) = \ln(\mu_c^*) = \ln(n_c^* \mu_c) = \ln(n_c^*) + \ln(\mu_c) = \ln(n_c^*) + \boldsymbol{x}_c^t \boldsymbol{\beta}$, où $\ln(n_c^*)$ est une variable offset.

Inférence sur les paramètres

– **Estimateur du maximum de vraisemblance de $\boldsymbol{\beta}$** :
$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{niter}$ où *niter* est la dernière itération de l'algorithme numérique de maximisation de la vraisemblance, si celui-ci a convergé.

– **Estimateur de la matrice de variance-covariance asymptotique du vecteur $\hat{\boldsymbol{\beta}}$** :
$\hat{\sigma}^2(\hat{\boldsymbol{\beta}}) = I^{-1}(\hat{\boldsymbol{\beta}})$ ($= I_o^{-1}(\hat{\boldsymbol{\beta}})$ pour l'algo de Newton-Raphson, $= I_e^{-1}(\hat{\boldsymbol{\beta}})$ pour l'algo Fisher scoring).
On estime donc l'erreur-type asymptotique d'un estimateur $\hat{\beta}_j$, notée $\sigma(\hat{\beta}_j)$, par la racine de l'élément (j, j) sur la diagonale de $I^{-1}(\hat{\boldsymbol{\beta}})$.

- **Intervalle de confiance de Wald de niveau $1-\alpha$ pour un paramètre** β_j : $\hat{\beta}_j \pm z_{\alpha/2}\hat{\sigma}(\hat{\beta}_j)$.
- **Test de Wald sur un paramètre** $\{H_0 : \beta_j = 0\}$ ou $\{H_0$: il n'y a pas de lien entre Y et $x_j\}$

  Statistique du test : $Z_w = \frac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)} \xrightarrow[H_0]{\text{asymp.}} N(0,1)$,

  ou encore : $Z_w^2 = \frac{\hat{\beta}_j^2}{\hat{\sigma}^2(\hat{\beta}_j)} \xrightarrow[H_0]{\text{asymp.}} \chi_1^2$.
- **Test de rapport de vraisemblance sur plusieurs paramètres** :
  $\{H_0 : \beta_j = 0$ pour tout $j \in \mathcal{B}\}$ ou $\{H_0$: Les modèles M_C et M_R sont équivalents$\}$
  où M_C est le modèle complet comprenant les paramètres β_j avec $j \in \mathcal{B}$ et M_R est le modèle réduit dans lequel ces paramètres sont fixés à zéro.
  Statistique du test :
  $LR = 2\{max\ln(L_C) - max\ln(L_R)\} \xrightarrow[H_0]{\text{asymp.}} \chi_b^2$
  où $max\ln(L_C)$ est la log-vraisemblance maximale de M_C,
  $max\ln(L_R)$ est la log-vraisemblance maximale de M_R et
  b est le nombre de β_j dans l'ensemble $\mathcal{B}$, soit le nombre de paramètres de M_C - nombre de paramètres de M_R.

Prédiction de Y

- **Valeur prédite** : $\hat{\mu} = g^{-1}(\boldsymbol{x}^t\hat{\boldsymbol{\beta}})$
- **Intervalle de confiance de Wald de niveau $1-\alpha$ pour μ** :
  $\mu \in [g^{-1}(\hat{\eta} - z_{\alpha/2}\ \hat{\sigma}(\hat{\eta}))\ ,\ g^{-1}(\hat{\eta} + z_{\alpha/2}\ \hat{\sigma}(\hat{\eta}))]$
  où $\hat{\eta} = \boldsymbol{x}^t\hat{\boldsymbol{\beta}}$ et $\hat{\sigma}(\hat{\eta}) = \sqrt{\boldsymbol{x}^t\hat{\sigma}^2(\hat{\boldsymbol{\beta}})\boldsymbol{x}}$.

Validation du modèle

Définition de la déviance :

Déviance standardisée : $D^* = 2\{max\ln(L_S) - max\ln(L_P)\}$
où $max\ln(L_S)$ est la log-vraisemblance maximale d'un modèle saturé (modèle comportant un paramètre pour chaque observation $\Rightarrow \hat{\mu}_i = y_i$, et $max\ln(L_P)$ est la log-vraisemblance maximale du modèle proposé.

Déviance (non standardisée) : $D = \phi D^*$.

Si $\phi = 1$, comme en régression Poisson et logistique, alors les deux déviances sont égales et la statistique du test de rapport de vraisemblance mentionné ci-dessus peut se calculer par $LR = D_R - D_C$ où D_R est la déviance du modèle réduit et D_C est la déviance du modèle complet.

Exemples de formules de déviance :

	reg Poisson lien ln	reg logistique lien logit
D^*	$2\sum_{i=1}^{n}\left(y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right)$	$2\sum_{i=1}^{n}\left(y_i \ln\left(\frac{y_i}{\hat{\pi}_i}\right) + (1-y_i)\ln\left(\frac{1-y_i}{1-\hat{\pi}_i}\right)\right)$

Définition de la statistique khi-deux de Pearson :

Stat khi-deux de Pearson standardisée : $X^{2\,*} = X^2/\phi$

Stat khi-deux de Pearson (non standardisée) : $X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$

où $V(\hat{\mu}_i) = n_i\hat{\pi}_i(1-\hat{\pi}_i)$ si $Y_i \sim Binomiale(n_i, \pi_i)$
et $V(\hat{\mu}_i) = \hat{\mu}_i$ si $Y_i \sim Poisson(\mu_i)$.

Statistiques d'ajustement du modèle :

Sous $\{H_0$: le modèle s'ajuste bien aux données$\}$ et à la condition que les valeurs prédites par le modèle ne soient pas trop petites, on a que

$$D^* \text{ et } X^{2\,*} \xrightarrow[H_0]{\text{asymp.}} \chi^2_{n-(p+1)}$$

où n représente le nombre de données utilisées pour ajuster le modèle et $p+1$ le nombre de paramètres dans le modèle à valider.

En régression logistique et Poisson, on calcule ces statistiques sur les données groupées pour avoir des valeurs prédites plus grandes et ainsi augmenter la fiabilité de la loi asymptotique (qui devient $\chi^2_{C-(p+1)}$ où C est le nombre de lignes du jeu de données groupées). Dans ces cas, les formules des statistiques d'ajustement sont :

Stat	régression Poisson	régression logistique
D^*	$2\sum_{c=1}^{C}\left(y_c^* \ln\left(\frac{y_c^*}{\hat{\mu}_c^*}\right) - (y_c^* - \hat{\mu}_c^*)\right)$	$2\sum_{c=1}^{C}\left(y_c^* \ln\left(\frac{y_c^*}{n_c^*\hat{\pi}_c}\right) + (n_c^* - y_c^*)\ln\left(\frac{(n_c^* - y_c^*)}{n_c^*(1-\hat{\pi}_c)}\right)\right)$
$X^{2\,*}$	$\sum_{c=1}^{C}\frac{(y_c^* - \hat{\mu}_c^*)^2}{\hat{\mu}_c^*}$	$\sum_{c=1}^{C}\frac{(y_c^* - n_c^*\hat{\pi}_c)^2}{n_c^*\hat{\pi}_c(1-\hat{\pi}_c)}$

Étude des valeurs prédites et des résidus : Encore une fois, en régression logistique et Poisson, on utilise des valeurs prédites et des résidus calculés avec les données groupées pour valider le modèle.

Postulats du modèle :

- les observations de la variable réponse Y sont indépendantes ;
- la relation entre la variable réponse Y et les variables explicatives est bien modélisée par la fonction de lien $g()$ choisie ;
- la variable réponse suit bien la distribution $\mathcal{L}$ choisie.

Graphiques pour étudier la justesse de la fonction de lien :

- valeurs observées en fonction des valeurs prédites :
  OK si linéaire ;
- valeurs observées et prédites en fonction d'un x_j :
  OK si les valeurs prédites sont proches des valeurs observées.

Résidus :

- résidus bruts : $y_i - \hat{\mu}_i$;
- résidus de Pearson : r_{P_i} défini tel que $\sum_{i=1}^{n} r_{P_i}^2 = X^2$;
- résidus de Déviance : $r_{D_i} = signe(y_i - \mu_i)\sqrt{d_i}$
  avec d_i défini tel que $\sum_{i=1}^{n} d_i = D$.

Sous $\{H_0$: le modèle s'ajuste bien aux données$\}$, et à la condition que les valeurs prédites par le modèle ne soient pas trop petites, ces deux derniers types de résidus suivent asymptotiquement une loi normale d'espérance nulle et de variance constante, mais pas unitaire.

Il existe une version standardisée pour ces deux types de résidus (formule non donnée en classe), ici notés sr_{P_i} et sr_{D_i}, tel que sous H_0 ces deux types de résidus suivent asymptotiquement une $N(0, 1)$.

Étude de la justesse de la distribution :

- Graphique des résidus de Pearson ou de déviance en fonction des valeurs prédites :
  OK si points répartis de façon aléatoire, homogène ;
- Étude de la normalité des résidus de Pearson ou de déviance :
  OK si normalité plausible.

Correction pour sur ou sous dispersion

- Matrice de variance-covariance asymptotique de $\hat{\boldsymbol{\beta}}$ corrigée : $\hat{\sigma}^2(\hat{\boldsymbol{\beta}})_{corr} = \hat{\sigma}^2(\hat{\boldsymbol{\beta}})\hat{\phi}$;
- Statistique corrigée du test de Wald : $Z_{w,corr} = \frac{Z_w}{\sqrt{\hat{\phi}}} = \frac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)\sqrt{\hat{\phi}}}$;
- Statistique corrigée du test de rapport de vraisemblance : $LR_{corr} = \frac{LR}{\hat{\phi}}$;

avec un facteur correctif basé sur la déviance : $\hat{\phi} = D/(C-(p+1))$
ou sur la statistique khi-deux de Pearson : $\hat{\phi} = X^2/(C-(p+1))$.

Régression logistique polytomique

La variable réponse Y est nominale à J modalités ($J > 2$).
Notons $\pi_k(\boldsymbol{x}) = P(Y = k \mid \boldsymbol{x})$ la probabilité que Y soit égale à sa modalité k, sachant $\boldsymbol{x}$, pour $k = 1, \ldots, J$.

Réponse nominale : modèle logit généralisé :

Modèle : $\ln\left(\frac{\pi_k(\boldsymbol{x})}{\pi_J(\boldsymbol{x})}\right) = \boldsymbol{x}^t\boldsymbol{\beta}_k$ pour $k = 1, \ldots, J-1$.

Valeurs prédites : $\hat{\pi}_k(\boldsymbol{x}) = \frac{exp(\boldsymbol{x}^t\hat{\boldsymbol{\beta}}_k)}{1+\sum_{h=1}^{J-1} exp(\boldsymbol{x}^t\hat{\boldsymbol{\beta}}_h)}$ pour $k = 1, \ldots, J-1$
et $\hat{\pi}_J(\boldsymbol{x}) = \frac{1}{1+\sum_{h=1}^{J-1} exp(\boldsymbol{x}^t\hat{\boldsymbol{\beta}}_h)}$.

Réponse ordinale : modèle à rapports de cotes proportionnels :

Modèle : $logit(P(Y \leq k \mid \boldsymbol{x})) = \beta_{0_k} + \boldsymbol{x}_{-0}^t\boldsymbol{\beta}_{-0}$ pour $k = 1, \ldots, J-1$,
où $\boldsymbol{x}_{-0}$ est le vecteur $\boldsymbol{x}$ auquel on a enlevé le premier élément unitaire et $\boldsymbol{\beta}_{-0}$ est le vecteur des paramètres $\boldsymbol{\beta}$ excluant l'ordonnée à l'origine.

Valeurs prédites : $\hat{P}(Y \leq k \mid \boldsymbol{x}) = \frac{exp(\hat{\beta}_{0_k} + \boldsymbol{x}_{-0}^t\hat{\boldsymbol{\beta}}_{-0})}{1+exp(\hat{\beta}_{0_k} + \boldsymbol{x}_{-0}^t\hat{\boldsymbol{\beta}}_{-0})}$ pour $k = 1, \ldots, J-1$
et $\hat{P}(Y \leq J \mid \boldsymbol{x}) = 1$, on a donc $\hat{\pi}_k(\boldsymbol{x}) = \hat{P}(Y \leq k \mid \boldsymbol{x}) - \hat{P}(Y \leq k-1 \mid \boldsymbol{x})$.

Note : On peut construire un modèle à risques partiellement proportionnels en rendant certains paramètres β devant des variables explicatives distincts et non communs entre les $J-1$ équations du modèle.

Annexe A

Rappels en probabilité et statistique

A.1 Définitions en probabilité

A.1.1 Probabilité conditionnelle

La probabilité conditionnelle que l'événement A se réalise, sachant qu'un autre événement B s'est réalisé, est définie par :

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

où $P(A \cap B)$ est la probabilité que les événements A et B se réalisent conjointement.

A.1.2 Théorème de Bayes

Soit A et B deux événements, le théorème de Bayes dit que :

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^C)P(A^C)}$$

où A^C est le complémentaire de l'événement A, donc $P(A^C) = 1 - P(A)$ (Ross, 2007, section 3.3).

A.2 Rappels concernant certaines distributions

A.2.1 Loi normale

Soit X une variable aléatoire continue de densité donnée par

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right],$$

pour $x \in \mathbb{R}$, $\mu \in \mathbb{R}$ et $0 < \sigma^2 < \infty$, on dit que X suit une loi normale de paramètres μ et σ^2, ce que l'on note $X \sim \mathcal{N}(\mu, \sigma^2)$ (Casella et Berger, 2002, section 3.3).

Le graphe de la densité de la loi normale $\mathcal{N}(\mu, \sigma^2)$ (figure A.1) est en forme de cloche symétrique centrée à μ et avec points d'inflexion $\mu - \sigma$ et $\mu + \sigma$.

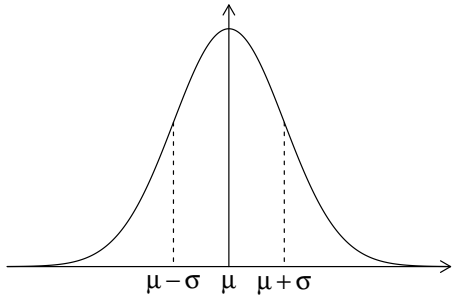


FIGURE A.1 – Densité $\mathcal{N}(\mu, \sigma^2)$.

L'espérance et la variance de X sont en fait les paramètres de la loi :

$$\begin{aligned} E(X) &= \mu \\ Var(X) &= \sigma^2. \end{aligned}$$

Propriétés de la loi normale

1. Toute variable aléatoire normale peut être ramenée à une variable aléatoire normale standard, ou centrée réduite, notée Z, par une simple

transformation. Si $X \sim \mathcal{N}(\mu, \sigma^2)$, alors

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1).$$

2. Soient $X_1, \ldots X_n$ n variables indépendantes telles que $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. Soient $a_1, \ldots, a_n$ et $b_1, \ldots, b_n$ des constantes. Alors,

$$\sum_{i=1}^{n}(a_i X_i + b_i) \sim \mathcal{N}\left(\sum_{i=1}^{n}(a_i \mu_i + b_i), \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$$

Cette propriété signifie que toute combinaison linéaire de variables aléatoires normales suit aussi une loi normale (Casella et Berger, 2002, corolaire 4.6.10).

Calcul de probabilités

On note Φ la fonction de répartition de la loi normale standard $\mathcal{N}(0, 1)$. Si Z est une variable aléatoire $\mathcal{N}(0, 1)$, alors

$$\begin{aligned} \Phi(z) &= P(Z \leq z) \\ &= \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{x^2}{2}\right] dx \end{aligned}$$

On trouve les valeurs de Φ dans des tables, comme celle de l'annexe C.1.

A.2.2 Théorème Limite Central (TLC)

Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes et identiquement distribuées de loi quelconque, mais avec moyenne $\mu \in \mathbb{R}$ et variance $0 < \sigma^2 < \infty$. Définissons la somme $S_n = \sum_{i=1}^{n} X_i$ et la moyenne $\bar{X} = \sum_{i=1}^{n} X_i / n$. Alors on a que

$$\frac{(S_n - n\mu)}{\sqrt{n}\sigma} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}$$

converge en probabilité vers une variable aléatoire normale standard $\mathcal{N}(0, 1)$ (Casella et Berger, 2002, théorème 5.5.15).

A.2.3 Loi du khi-deux

Soit X une variable aléatoire suivant une loi du khi-deux, aussi nommée khi-carré (en anglais chi-square), à d degrés de liberté. On note $X \sim \chi^2_d$. La densité de cette variable aléatoire est (voir figure A.2) :

$$f(x) = \begin{cases} \frac{1}{2^{d/2}\Gamma(d/2)} x^{d/2-1} e^{-x/2} & \text{si } x > 0, \\ 0 & \text{sinon.} \end{cases}$$

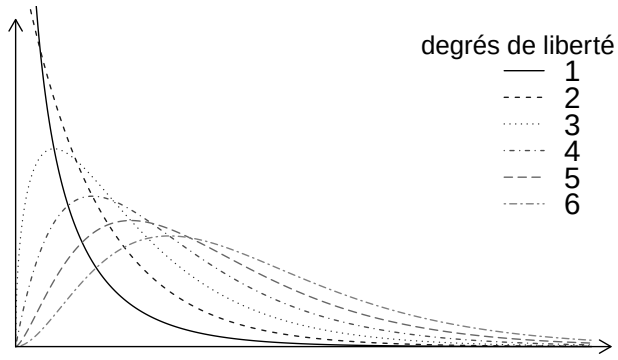


FIGURE A.2 – Densité χ^2_d pour quelques valeurs de degrés de liberté d.

L'espérance et la variance de X sont :

$$\begin{aligned} E(X) &= d \\ Var(X) &= 2d. \end{aligned}$$

Quelques liens entre la loi du khi-deux et les autres lois :
- si $X \sim \chi^2_d$, alors ceci est équivalent à $X \sim Gamma(d/2, 2)$;
- si $X \sim \mathcal{N}(0,1)$ alors $X^2 \sim \chi^2_1$.

Une propriété intéressante de la loi du khi-deux :
si $X_1, \ldots, X_n$ sont des variables indépendantes telles que $X_i \sim \chi^2_{d_i}$ pour $i = 1, ..., n$, alors

$$\sum_{i=1}^{n} X_i \sim \chi^2_{\left(\sum_{i=1}^{n} d_i\right)}.$$

Ces informations peuvent être retrouvées dans Casella et Berger (2002, section 3.3 et lemme 5.3.2). L'annexe C.2 contient une table des quantiles de la loi du khi-deux selon ses degrés de liberté.

A.2.4 Famille exponentielle

Certains résultats utilisés dans ce cours sont vrais uniquement pour des distributions de la famille exponentielle. Casella et Berger (2002, section 3.4) définissent ce qu'est la famille exponentielle. Ce qui importe de savoir ici, c'est que les distributions binomiale, Poisson, multinomiale et normale en font partie.

A.3 Vraisemblance

Soit $X_1, \ldots, X_n$ un échantillon aléatoire de taille n, de fonction de masse ou de densité conjointe $f(\boldsymbol{x}|\boldsymbol{\theta})$. Conditionnellement à ce que $X = (X_1, \ldots, X_n)$ soit observé et prenne la valeur $x = (x_1, \ldots, x_n)$, la fonction de vraisemblance est définie par (Casella et Berger, 2002, section 6.3.1) :

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = f(\boldsymbol{x}|\boldsymbol{\theta}).$$

La différence entre la vraisemblance $L(\boldsymbol{\theta}|\boldsymbol{x})$ et la densité $f(\boldsymbol{x}|\boldsymbol{\theta})$ est que la première est une fonction du vecteur de paramètres alors que la deuxième est une fonction des observations x.

Si les variables aléatoires X_1 à X_n sont indépendantes, on a que

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}).$$

A.3.1 Définition de la fonction score et des matrices d'information espérée et observée

À partir de la vraisemblance d'un vecteur de paramètres, on peut définir les quantités suivantes (Casella et Berger, 2002, section 10.3.2), qui nous seront utiles pour obtenir des estimateurs des paramètres (en particulier dans un modèle linéaire généralisé) et construire des tests.

fonction score : vecteur des dérivées partielles de la log-vraisemlance par rapport aux paramètres :

$$S(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ln L(\boldsymbol{\theta}|\boldsymbol{x}).$$

Matrice d'information observée : moins la matrice des dérivées secondes de la log-vraisemblance par rapport aux paramètres :

$$I_o(\boldsymbol{\theta}) = -\frac{\partial^2}{\partial \theta_j \partial \theta_{j'}} \ln L(\boldsymbol{\theta}|\boldsymbol{x}).$$

Matrice d'information espérée (aussi appelée matrice d'information de Fisher ou simplement matrice d'information) : Pour présenter cette matrice, il faut d'abord voir le score défini ci-dessus comme une statistique. Pour une valeur donnée de $\boldsymbol{\theta}$, le score dépend de l'échantillon, et est donc une variable aléatoire. Étant donné que nous considérons maintenant $\boldsymbol{\theta}$ connu, nous allons remplacer dans la formule du score $L(\boldsymbol{\theta}|\boldsymbol{x})$ par $f(\boldsymbol{x}|\boldsymbol{\theta})$. De plus, nous allons remplacer dans cette expression le petit $\boldsymbol{x}$ par un grand $\boldsymbol{X}$, car le score est maintenant vu comme une nouvelle variable aléatoire fonction des variables aléatoires X_1 à X_n. On écrit donc maintenant le score ainsi :

$$S(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ln f(\boldsymbol{X}|\boldsymbol{\theta}).$$

La matrice d'information espérée se définit ainsi :

$$I_e(\boldsymbol{\theta}) = E(S(\boldsymbol{\theta})^2) = E\left(\left(\frac{\partial}{\partial \boldsymbol{\theta}} \ln f(\boldsymbol{X}|\boldsymbol{\theta})\right)^2\right)$$

Si X suit une distribution de la famille exponentielle, cette matrice se simplifie à :

$$I_e(\boldsymbol{\theta}) = -E\left(\frac{\partial^2}{\partial \theta_j \partial \theta_{j'}} \ln f(\boldsymbol{X}|\boldsymbol{\theta})\right).$$

L'inégalité de Cramér-Rao (Casella et Berger, 2002, section 7.3.2) dit que la matrice de variance covariance d'un estimateur sans biais de $\boldsymbol{\theta}$, noté $\boldsymbol{\theta}^*$, est supérieure ou égale à l'inverse de la matrice d'information espérée :

$$Var(\boldsymbol{\theta}^*) \geq I_e(\boldsymbol{\theta})^{-1}.$$

A.3.2 Estimateur du maximum de vraisemblance

L'estimateur du maximum de vraisemblance de $\boldsymbol{\theta}$ est le point en lequel la vraisemblance $L(\boldsymbol{\theta}|\boldsymbol{x})$ est maximisée. Il sera noté $\hat{\boldsymbol{\theta}}$. Dériver une formule pour un estimateur du maximum de vraisemblance fait intervenir des notions de calcul différentiel et intégral. Maximiser la vraisemblance revient à maximiser la log-vraisemblance puisque le logarithme naturel est une fonction monotone croissante. Ainsi, l'estimateur du maximum de vraisemblance de $\boldsymbol{\theta}$ est un point en lequel la fonction score définie ci-dessus (dérivée première de la log-vraisemblance par rapport à $\boldsymbol{\theta}$) vaut zéro. Cependant, cette condition n'est pas suffisante pour affirmer qu'il s'agit bien d'un maximum, et non d'un minimum ou encore d'un point de selle. Il faut donc étudier les dérivées secondes de la log-vraisemblance. Cependant, nous ne ferons pas de calculs du genre dans ce cours. Nous allons simplement utiliser des résultats que vous avez probablement déjà prouvés dans d'autres cours. Si ce n'est pas le cas, vous pouvez trouver les preuves de ces résultats dans un livre de base en statistique mathématique tel Casella et Berger (2002) ou Hogg *et al.* (2005).

A.4 Tests asymptotiques usuels

Soit θ un paramètre scalaire. On cherche à mener le test suivant sur ce paramètre :

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \text{ au choix } \left\langle \begin{array}{llll} \theta \neq \theta_0 & \text{ou} & & \text{test biltéral} \\ \theta > \theta_0 & \text{ou} & \theta < \theta_0 & \text{test unilatéral.} \end{array} \right.$$

Nous utilisons fréquemment dans ce cours trois types de test asymptotiques usuels. Il s'agit des tests de Wald, score et du rapport de vraisemblance.

A.4.1 Test de Wald

Les tests de Wald sont les plus simples à construire parmi les 3 trois types de tests présentés dans cette section. Ils portent ce nom en l'honneur du statisticien hongrois Abraham Wald. La statistique de ce test est définie de façon très générale comme suit :

$$\frac{\theta^* - \theta_0}{se(\theta^*)} \qquad \xrightarrow[H_0]{\text{asymp.}} \qquad N(0,1)$$

où θ^* est un estimateur ponctuel de θ et $se(\theta^*)$ est un estimateur de l'erreur-type de θ^*. On note parfois $\hat{\sigma}(\theta^*)$ au lieu de $se(\theta^*)$. Cette autre notation met vraiment en évidence le fait que l'erreur-type est estimée. Souvent, on utilise $\theta^* = \hat{\theta}$, soit l'estimateur du maximum de vraisemblance de θ.

A.4.2 Test score

Les tests score (en anglais score tests), aussi appelés « tests du multiplicateur de Lagrange », sont dus au statisticien C.R. Rao. Ils se basent sur la statistique de test suivante :

$$\frac{S(\theta_0)}{\sqrt{I(\theta_0)}} \qquad \xrightarrow[H_0]{\text{asymp.}} \qquad N(0,1)$$

où $S(\theta_0)$ est la fonction score calculée au point $\theta = \theta_0$ et $I(\theta_0)$ est la matrice d'information espérée (ici de dimension 1×1) calculée au point $\theta = \theta_0$. Ce test comporte moins d'approximation que le test de Wald puisque l'erreur-type se trouvant au dénominateur n'est pas estimée.

A.4.3 Test du rapport de vraisemblance

Pour le test du rapport de vraisemblance, nous allons nous ramener à un cas plus général où $\boldsymbol{\theta}$ est un vecteur de paramètres. On cherche à tester :

$$H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0 \qquad \text{versus} \qquad H_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0^c$$

où $\boldsymbol{\Theta}_0$ représente un sous-ensemble de l'espace $\boldsymbol{\Theta}$ des valeurs possible de $\boldsymbol{\theta}$ et $\boldsymbol{\Theta}_0^c$ est le complément de ce sous-ensemble. On aurait pu noter $\boldsymbol{\Theta}_0^c$ de la façon suivante : $\boldsymbol{\Theta}/\boldsymbol{\Theta}_0$. Notez que ce test possède uniquement une forme bilatérale.

Le rapport de vraisemblance est défini comme suit :

$$\Lambda(\boldsymbol{x}) = \frac{\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} L(\boldsymbol{\theta}|\boldsymbol{x})}{\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\boldsymbol{\theta}|\boldsymbol{x})}$$

Si $\hat{\boldsymbol{\theta}}$, le maximum de vraisemblance de $\boldsymbol{\theta}$ existe, et que $\hat{\boldsymbol{\theta}}_0$ représente le maximum de vraisemblance restreint sous l'espace $\boldsymbol{\Theta}_0$, alors le rapport de vraisemblance peut s'écrire plus simplement :

$$\Lambda(\boldsymbol{x}) = \frac{L(\hat{\boldsymbol{\theta}}_0|\boldsymbol{x})}{L(\hat{\boldsymbol{\theta}}|\boldsymbol{x})}$$

Un test du rapport de vraisemblance est un test dont la région critique a la forme $\{\boldsymbol{x} : \Lambda(\boldsymbol{x}) \leq c\}$, où c est un nombre tel que $0 \leq c \leq 1$.

Pour effectuer un test de rapport de vraisemblance, on utilise la statistique suivante :

$$\begin{aligned} LR &= -2 \ln \Lambda(\boldsymbol{X}) \\ &= -2 \ln \left(\frac{L(\boldsymbol{\theta}_0|\boldsymbol{X})}{L(\hat{\boldsymbol{\theta}}|\boldsymbol{X})} \right). \end{aligned}$$

Sous H_0, lorsque $n \to \infty$, cette statistique suit une loi du khi-deux à d degrés de liberté (Casella et Berger, 2002, théorème 10.3.3) :

$$LR \quad \xrightarrow[H_0]{\text{asymp.}} \quad \chi^2_d.$$

Le nombre de degrés de liberté d est la différence entre le nombre de paramètres libres sous l'espace $\boldsymbol{\Theta}$ (en d'autres mots la dimension de $\boldsymbol{\Theta}$) et le nombre de paramètres libres sous $\boldsymbol{\Theta}_0$ (en d'autres mots la dimension de $\boldsymbol{\Theta}_0$).

Le seuil observé du test se calcule donc ainsi : $P(\chi^2_d \geq lr)$ où lr est la valeur observée de la statistique de test LR.

A.5 Intervalles de confiance

Les estimations ponctuelles comportent toujours une incertitude. Un intervalle de confiance estime un paramètre tout en informant sur l'incertitude de cette estimation. On définit un intervalle de confiance de niveau de confiance $1-\alpha$ du paramètre scalaire θ par les bornes L et U telles que

$$P(L \leq \theta \leq U) = 1 - \alpha.$$

On rapporte habituellement un intervalle de confiance comme suit :
$\theta \in [L, U]$.

Il existe un lien entre les intervalles de confiance et un test bilatéral sur un paramètre :

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta \neq \theta_0.$$

De toute statistique utilisable pour confronter ces hypothèses, on peut tirer un intervalle de confiance correspondant tel qu'on rejettera H_0 si et seulement si θ_0 n'appartient pas à l'intervalle $[L, U]$.

A.5.1 Intervalle de confiance de Wald

Prenons la statistique du test de Wald :

$$\frac{\theta^* - \theta_0}{se(\theta^*)} \qquad \xrightarrow[H_0]{\text{asymp.}} \qquad N(0,1).$$

Plaçons-nous dans le cas général et non sous l'hypothèse nulle H_0. On a alors que

$$\frac{\theta^* - \theta}{se(\theta^*)} \qquad \xrightarrow{\text{asymp.}} \qquad N(0,1).$$

On sait donc que

$$P\left(-z_{\alpha/2} \leq \frac{\theta^* - \theta}{se(\theta^*)} \leq z_{\alpha/2}\right) \approx 1 - \alpha.$$

Dans cette équation, isolons θ. Nous obtenons

$$P\left(\theta^* - z_{\alpha/2} se(\theta^*) \leq \theta \leq \theta^* + z_{\alpha/2} se(\theta^*)\right) \approx 1 - \alpha.$$

Ainsi, $[\theta^* - z_{\alpha/2} se(\theta^*), \theta^* + z_{\alpha/2} se(\theta^*)]$ est l'intervalle de confiance de Wald de niveau $1-\alpha$ pour θ.

A.5.2 Intervalles de confiance sans forme algébrique

Dans le cas de l'intervalle de confiance de Wald, les calculs étaient directs et une forme algébrique simple a été obtenue. Ce n'est pas toujours facile d'isoler θ. C'est même parfois impossible, mais une solution peut toujours être trouvée numériquement. Ainsi, il est toujours possible de calculer numériquement l'intervalle de confiance correspondant à un test du rapport de vraisemblance. Il s'agit des valeurs de θ pour lesquelles la statistique du test, notée LR, est inférieure ou égale à $\chi^2_{1,\alpha}$, car $P(LR \leq \chi^2_{1,\alpha}) \approx 1 - \alpha$.

Ainsi, pour trouver numériquement les bornes de l'intervalle de confiance, on peut calculer la statistique du test LR pour un grand nombre de valeurs du paramètre θ, couvrant tout le support des valeurs possibles de θ. La plus petite valeur de θ essayée pour laquelle $LR \leq \chi^2_{1,\alpha}$ est la borne inférieure de l'intervalle et la plus grande valeur de θ avec $LR \leq \chi^2_{1,\alpha}$ est la borne supérieure.

Annexe B

Quelques études dans le domaine médical

Les variables catégoriques sont fréquentes dans le domaine médical, en particulier en épidémiologie. L'épidémiologie est, selon l'Office québécois de la langue française, la branche de la médecine qui étudie les divers facteurs conditionnant l'apparition, la fréquence, le mode de diffusion et l'évolution des maladies affectant des groupes d'individus. Dans ce domaine, il n'est donc pas rare de séparer des sujets d'une étude en groupes selon leur exposition à certains facteurs de risque d'une maladie. Le but est d'étudier l'effet des facteurs de risque sur la survenue de la maladie. Dans le traitement statistique des données recueillies lors de ces études, les facteurs de risque sont représentés par des variables explicatives X_j et la survenue de la maladie par une variable réponse Y. Cette variable est souvent catégorique (par exemple une variable dichotomique : développer ou non la maladie).

Les informations dans cette section ne sont pas propres à l'analyse de données catégoriques, elles sont plus générales. De plus, elles sont utiles à la compréhension de matière dans plus d'un chapitre des notes. C'est pourquoi cette section est placée en annexe plutôt que d'être intégré à un chapitre.

Pour plus d'information sur les types d'études épidémiologiques, Rothman *et al.* (2008, chapitre 6) est une bonne référence.

B.1 Caractéristiques des études

Pour décrire les différentes études, nous allons utiliser les caractéristiques suivantes :

Expérimentales versus observationnelles :

- **Les études expérimentales** :
  L'exposition des sujets aux facteurs de risque est contrôlée : ils sont assignés à un groupe d'exposition par les chercheurs.
- **Les études observationnelles** :
  L'exposition des sujets aux facteurs de risque n'est pas contrôlée par les chercheurs. C'est ce qui survient dans la vie des sujets qui va déterminer leur appartenance aux groupes d'exposition au facteur de risque.

Aspect temporel :

- **Les études prospectives** :
  Les sujets sont suivis dans le temps. Au début de l'étude, l'exposition des sujets aux facteurs de risque est connue.
- **Les études rétrospectives** :
  Les chercheurs déterminent l'exposition des sujets aux facteurs de risque en cherchant dans leur passé, soit en leur posant des questions ou en tirant de l'information dans leur dossier médical.

Type d'échantillonnage :

- **Échantillonnage simple** :
  On tire un seul échantillon à partir de la population cible de l'étude.
- **Échantillonnage multiple** :
  La population est divisée en sous-populations (parfois appelées strates).
  On effectue un échantillonnage distinct dans chacune des sous-populations.

B.2 Types d'études

Les méthodes statistiques vues dans ce cours servent fréquemment à l'analyse de données recueillies lors d'études des types suivants :

1. **L'étude transversale** (cross-sectionnal study) :
   Ce genre d'étude est simplement une enquête dans laquelle on détermine les valeurs des variables explicatives X (exposition aux facteurs de risque) et de la variable réponse Y (survenue de la maladie) à partir des réponses des participants.

   Exemple : Pour étudier le lien entre la cigarette (X) sur la survenue d'un infarctus du myocarde(Y), on pourrait simplement échantillonner aléatoirement des sujets dans la population et leur demander de participer à un sondage. On demanderait aux individus s'ils sont fumeurs ou s'ils ont déjà été fumeurs et s'ils ont déjà été victimes d'un infarctus.

   Caractéristiques :
   - Étude observationnelle ;
   - Étude rétrospective ;
   - Typiquement, l'échantillonnage est simple (tel que suggéré dans l'exemple), mais il peut aussi être multiple. S'il est multiple, les sous-populations sont formées en fonction de l'exposition des individus à un ou des facteurs de risque. Pour ce faire, il faut connaître d'avance l'exposition de tous les individus de la population à ces facteurs de risque.

2. **L'étude cas-témoins** (case-control study) :
   Ce type d'étude est typiquement utilisé pour étudier les facteurs de risque d'une maladie rare. Elle a beaucoup de points en commun avec une étude transversale, car les variables X et Y sont mesurées à un moment précis dans le temps. Cependant, on suréchantillonne les sujets malades (appelés les cas) par rapport aux sujets non malades (appelés les témoins). On s'assure ainsi d'avoir suffisamment de cas, sans pour autant devoir inclure un très grand nombre de sujets dans l'échantillon et ainsi réduire les coûts de l'étude. L'échantillonnage est donc multiple, mais les sous-populations sont déterminées par Y et non par X,

comme lorsqu'on utilise un échantillonnage multiple dans une étude transversale. Étant donné que l'on doit connaître Y afin de procéder à l'échantillonnage, il est parfois difficile de former un échantillon réellement représentatif de la population à l'étude.

Exemple : Pour étudier l'effet de la cigarette sur la survenue d'un infarctus du myocarde, on pourrait échantillonner des cas parmi des patients d'un hôpital ayant subi un infarctus. Les témoins pourraient être n'importe quels autres patients de l'hôpital n'ayant pas subi d'infarctus. Par un questionnaire, on leur demanderait s'ils sont ou ont déjà été fumeurs.

Caractéristiques :
- Étude observationnelle ;
- Étude rétrospective ;
- Échantillonnage toujours multiple, avec deux sous-populations : les cas (malades) versus les témoins (non malades). On procède donc à deux échantillonnages distincts : un parmi les cas et un autre parmi les témoins. Deux tailles d'échantillons sont préétablies. On sélectionne proportionnellement un plus grand nombre de cas que l'on en retrouve dans la population en général. Ainsi, la proportion de malades parmi les participants à l'étude est plus grande que la proportion de malades dans la population en général. Le suréchantillonnage des cas induit un biais, dont on doit tenir compte lors de l'analyse statistique.

3. **L'étude longitudinale** :
   Une étude longitudinale se déroule dans le temps, contrairement à une étude transversale qui se réalise à un moment précis. On distingue ici deux types d'étude longitudinale.

   (a) **L'essai clinique** (clinical trial) :
   On assigne les sujets à un groupe d'exposition aux facteurs de risque (variables X_i). On observe ensuite les sujets pendants un certain temps afin de mesurer la variable Y.

Exemple : Afin de déterminer si la prise quotidienne d'aspirine réduit le risque d'infarctus du myocarde, on pourrait mener un essai clinique. Tous les sujets de l'étude prendraient une pilule par jour pendant 5 ans. Pour certains, cette pilule contiendrait de l'aspirine, tandis que pour d'autres, il s'agirait simplement d'un placébo. Les sujets seraient assignés de façon aléatoire à un des deux groupes (aspirine ou placebo). À la fin de l'étude, on dénombrerait les participants victimes ou non d'un infarctus pendant la durée de l'étude.

Caractéristiques :
- Étude expérimentale ;
- Étude prospective ;
- Échantillonnage multiple, car on fixe le nombre de sujets par groupe d'exposition aux facteurs de risque.

Remarque : L'utilisation probablement la plus connue des essais cliniques est celle pour le développement de nouveaux médicaments par des compagnies pharmaceutiques.

(b) **L'étude de cohorte** (cohort study) :
Au début de l'étude, on mesure l'exposition des sujets aux facteurs de risque, sans les contrôler. On doit parfois attendre plusieurs années avant d'observer la survenue de la maladie Y. Ces études sont longues et souvent coûteuses, mais elles sont effectuées en général avec un échantillon qui représente bien la population à l'étude. Elles fournissent beaucoup d'informations et permettent d'effectuer des inférences fiables si les tailles d'échantillons sont suffisantes.

Exemple : Dans l'exemple de la cigarette et de l'infarctus du myocarde, si on avait d'abord échantillonné des individus de la population et qu'on les avait suivis dans le temps, il aurait s'agit d'une étude de cohorte. Puisque subir un infarctus est un événement rare, il faudrait inclure un très grand nombre de sujets dans l'étude pour espérer observer suffisamment de cas pour assurer la validité de l'inférence statistique.

Caractéristiques :
- Étude observationnelle ;
- Étude typiquement prospective, mais peut aussi être rétrospective (historique) ;
- L'échantillonnage est parfois simple (tel que suggéré dans l'exemple) et parfois multiple, comme dans un essai clinique. S'il est multiple, les sous-populations sont formées en fonction de l'exposition des individus à un ou des facteurs de risque.

Annexe C

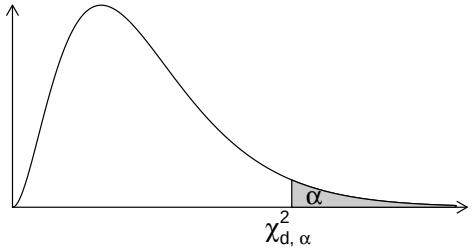
Tables de loi

C.1 Table de la loi normale

Fonction de répartition $\Phi(z) = P(Z \leq z)$ de la loi normale standard $\mathcal{N}(0,1)$

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9031	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9958	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

C.2 Table de quantiles de la loi khi-deux



Degrés de liberté	Probabilité α								
	0.99	**0.95**	**0.90**	**0.75**	**0.50**	**0.25**	**0.10**	**0.05**	**0.01**
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68	24.72
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03	26.22
13	4.107	5.892	7.042	9.299	12.340	15.98	19.81	22.36	27.69
14	4.660	6.571	7.790	10.165	13.339	17.12	21.06	23.68	29.14
15	5.229	7.261	8.547	11.037	14.339	18.25	22.31	25.00	30.58
16	5.812	7.962	9.312	11.912	15.338	19.37	23.54	26.30	32.00
17	6.408	8.672	10.085	12.792	16.338	20.49	24.77	27.59	33.41
18	7.015	9.390	10.865	13.675	17.338	21.60	25.99	28.87	34.80
19	7.633	10.117	11.651	14.562	18.338	22.72	27.20	30.14	36.19
20	8.260	10.851	12.443	15.452	19.337	23.83	28.41	31.41	37.57
22	9.542	12.338	14.041	17.240	21.337	26.04	30.81	33.92	40.29
24	10.856	13.848	15.659	19.037	23.337	28.24	33.20	36.42	42.98
26	12.198	15.379	17.292	20.843	25.336	30.43	35.56	38.89	45.64
28	13.565	16.928	18.939	22.657	27.336	32.62	37.92	41.34	48.28
30	14.953	18.493	20.599	24.478	29.336	34.80	40.26	43.77	50.89
40	22.164	26.509	29.051	33.660	39.335	45.62	51.80	55.76	63.69
50	27.707	34.764	37.689	42.942	49.335	56.33	63.17	67.50	76.15
60	37.485	43.188	46.459	52.294	59.335	66.98	74.40	79.08	88.38

Bibliographie

Agresti, A. (1992). A survey of exact inference for contingency tables. *Statistical Science*, 7:131–177.

Agresti, A. (2002). *Categorical Data Analysis*. Wiley. Second edition.

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*. Wiley. Second edition.

Agresti, A. et Coull, B. A. (1998). Approximate is better than exact for interval estimation of binomial proportions. *The American Statistician*, 52(2):119–126.

Beaucage, C. et Viger, Y. B. (1996). *Épidémiologie appliquée : Une initiation à la lecture critique de la littérature en sciences de la santé*. Gaëtan Morin Éditeur.

Bishop, Y. M., Fienberg, S. et Holland, P. (1975). *Discrete Multivariate Analysis : Theory and Practice.* M.I.T. Press. réimprimé par Springer-Verlag en 2007.

Breslow, N. et Day, N. E. (1980). *Statistical Methods in Cancer Research, Vol. I : The Analysis of Case-Control Studies*. Lyon : IARC.

Cameron, A. et Trivedi, P. (1998). *Regression Analysis of Count Data*. Cambridge University Press.

Casella, G. et Berger, R. L. (2002). *Statistical Inference*. Duxbury. Second edition.

Clopper, C. J. et Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26:404–413.

Conover, W. J. (1999). *Practical Nonparametric Statistics*. John Wiley & Sons, Inc. Third edition.

Duan, N. (1983). Smearing estimate : A nonparametric retransformation method. *Journal of the American Statistical Association*, 78(383):605–610.

Fleiss, J. L., Levin, B. et Paik, M. C. (2003). *Statistical Methods for Rates and Proportions*. John Wiley and Sons. Third edition.

Freeman, G. H. et Halton, J. H. (1951). Note on an exact treatment of contingency, goodness of fit and other problems of significance. *Biometrika*, 38:141–149.

Friendly, M. (1994). Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, 89:190–200.

Friendly, M. (2000). *Visualizing Categorical Data*. SAS Press.

Gramenzi, A., Gentile, A., Fasoli, M., D'Avanzo, B., Negri, E., Parazzini, F. et Vecchia, C. L. (1989). Smoking and myocardial infarction in women : A case-control study from northern italy. *Journal of Epidemiology and Community Health*, 43:214–217.

GSS (2012). General social survey. [En ligne], http://www3.norc.org/gss+website/, Accès aux données : http://sda.berkeley.edu/cgi-bin/hsda?harcsda+gss10nw.

Hines, W. W., Montgomery, D. C., Goldsman, D. M. et Borror, C. M. (2012). *Probabilités et statistique pour ingénieurs*. Chenelière Éducation. 2e édition.

Hogg, R. V., McKean, J. W. et allen T. Craig (2005). *Introduction to mathematical Statistics*. Prentice Hall. Sixth edition.

Hosmer, D. et Lemeshow, S. (2000). *Applied Logistic Regression*. John Wiley and Sons, second édition.

Landis, R. J., Heyman, E. R. et Koch, G. G. (1978). Average partial association in three-way contingency tables : A review and discussion of alternative tests. *International Statistical Review*, 46:237–254.

Lebart, L., Morineau, A. et Piron, M. (1997). *Statistique exploratoire multidimentionnelle*. Dunod. Deuxième édition.

Lohr, S. L. (2009). *Sampling : Design and Analysis*. Cengage Learning. Second edition.

McCullagh, P. et Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall. Second edition.

Moore, D. (2003). *The Basic Practice of Statistics*. W. H. Freeman. Third edition.

Nelder, J. A. et Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A*, 135(3):370–384.

Ross, S. M. (2007). *Initiation aux probabilités*. Presses Polytechniques et Universitaires Romandes. Traduction de la septième édition américaine.

Rothman, K. J., Greenland, S. et Lash, T. L. (2008). *Modern Epidemiology*. Lippincott Williams & Wilkins. Third edition.

Shoukri, M. M. (2010). *Measures of Interobserver Agreement and Reliability*. Chapman & Hall/CRC Press. Second edition.

Steering Committee of the Physicians' Health Study Research Group (1989). Final report on the aspirin component of the ongoing physicians' health study. *The New England journal of medicine*, 321:129–135.

Sun, X. et Yang, Z. (2008). Generalized mcnemar's test for homogeneity of the marginal distributions. *In SAS Global Forum*. Disponible en ligne : http://www2.sas.com/proceedings/forum2008/382-2008.pdf.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22:209–212.

Zelen, M. (1971). The analysis of several 2×2 contingency tables. *Biometrika*, 58:129–137.