Analyse de données catégorielles

Ce cours est librement inspiré du livre de Agresti ([1]), ainsi que du livre [2]. Il reprend par ailleurs des parties et des graphiques issues du cours de Patrick Taffé (qui vient de disparaître du net, mais que l'on peut encore trouver si on cherche bien).

1 Variables catégorielles

Elles sont en fait de types assez variés et on peut les organiser selon les catégories suivantes.

1.1 Distinction entre variable réponse et variable explicative

Les modèles statistiques nécessitent la distinction entre la variable réponse (variable dépendante) et les variables explicatives (indépendantes). Cette partie du cours s'intéressera uniquement aux méthodes statistiques adaptées au cas d'un variable réponse (dépendante) catégorielle, les variables explicatives pouvant être de n'importe quel type.

1.2 Distinction entre variables nominales, ordinales et intervalles

Une **variable nominale** ne fournit pas d'information qui permette de comparer quantitativement les éléments de l'échantillon (sujets). Elle attribue à chaque sujet une étiquette.

Exemples : Le sexe (masculin ou féminin) La religion (catholique, protestant, Musulman ...). La profession (avocat, professeur, plombier ...).

Ici, l'ordre des modalités n'a pas de sens, et l'analyse statistique ne doit pas dépendre de cet ordre.

Beaucoup de variables catégorielles sont en général ordonnées, bien que la distance entre catégories soit inconnue (i.e. non quantifiable). Ce sont les **variables ordinales.**

Une variable de type **intervalle** est quand à elle de type numérique (exemple pression artérielle, durée de vie etc.) et il est possible de calculer une distance entre les modalités. C'est la façon dont on mesure la variable qui définit son type.

Exemple : la variable "éducation" :
- Nominale : privé ou public.
- Ordinale : Niveau d'études atteint (brevet, bac, licence, master etc.).
- Intervalle : Nombre d'années d'études après le bac.

Les méthodes pour variables nominales peuvent s'appliquer aux deux autres types. Les méthodes pour variables ordinales peuvent s'appliquer au type intervalle (si le nombre de catégories est faible).

Nous allons parler des méthodes pour variables nominales et ordinales, qui peuvent aussi s'appliquer aux variables de type intervalle :
- Si le nombre de valeurs prises est faible.
- Après avoir effectué un regroupement des valeurs en catégories.

1.3 Distinction entre variables continues et discrètes

La distinction s'effectue principalement à l'aide du nombre de valeurs pouvant être prises par la variable. En effet, dans la pratique, les moyens de mesures utilisent une discrétisation liée à la précision des instruments. Des **variables discrètes** prenant un grand nombre de valeurs peuvent être considérées comme **continues**.

Dans la suite de ce cours, nous considérerons les variables à réponses discrètes suivantes :
- Variables nominales.
- Variables ordinales.
- Variables discrètes de type intervalle avec peu de valeurs.
- Variables continues groupées par catégories.

1.4 Distinction entre quantitatif et qualitatif

Les variables nominales sont **qualitatives** (les catégories diffèrent en qualité mais pas en quantité). Les variables de type intervalle sont **quantitatives** (différents niveaux correspondent à différentes quantités de la caractéristique d'intérêt). Les variables ordinales sont "entre les deux". On peut les voir comme des variables qualitatives (utiliser des méthodes pour variables nominales) mais on les associe plus souvent au type intervalle. En effet, elles sont en général liées à une variable continue sous-jacente qu'il est impossible de mesurer. Leur bonne gestion demande une bonne connaissance du problème (expertise), mais on peut utiliser une grande variété de méthodes pour les analyser.

2 Distributions pour données catégorielles

2.1 Distribution binomiale

De nombreuses applications considèrent le cas où on observe un nombre fixé n d'observations binaires (succès-échec), soit $y_1, y_2, ..., y_n$ i.i.d. tq $P(Y_i = 1) = \pi$ et $P(Y_i = 0) = 1 - \pi$. C'est la loi de Bernoulli.

- Essais identiques : la probabilité de succès π est la même pour chaque essai.
- Essais indépendants : les variables $\{Y_i\}$ sont indépendantes.

La variable $Y = \sum_{i=1}^{n} Y_i$ est distribuée selon une loi Binomiale $bin(n, \pi)$ tq :

$$p(y) = \begin{pmatrix} n \\ y \end{pmatrix} \pi^y (1-\pi)^{n-y}, \qquad y = 0, 1, 2, ..., n.$$

avec : $\begin{pmatrix} n \\ y \end{pmatrix} = \frac{n!}{y!(n-y)!}$. De plus :

$$E(Y_i) = \pi \quad \text{et} \quad var(Y_i) = \pi(1-\pi).$$

Et donc

$$E(Y) = n\pi \quad \text{et} \quad var(Y) = n\pi(1-\pi).$$

Dans la pratique, il n'est pas toujours garantit que des observations binaires successives soient indépendantes. On utilise dans ce cas d'autres distributions. C'est le cas lorsque l'on fait des tirages de Bernoulli dans une population finie : On utilise alors la loi hypergéométrique.

2.2 Distribution multinomiale

C'est le cas ou la réponse Y peut avoir plus de deux valeurs. Soit $y_{ij} = 1$ si le résultat de l'essai i appartient à la catégorie j, et $y_{ij} = 0$ sinon. Alors $y_i = (y_{i1}, y_{i2,}..., y_{ic})$ est un essai multinomial avec $\sum_{j=1}^{n} y_{ij} = 1$. Soit $n_j = \sum_{i=1}^{n} y_{ij}$. Le vecteur $(n_1, n_2, ..., n_c)$ a une distribution multinomiale.

Soit $\pi_j = P(Y_{ij} = 1)$. La loi multinomiale est :

$$p(n_1, n_2, ..., n_{c-1}) = \left(\frac{n!}{n_1!n_2!...n_c!}\right) \pi_1^{n_1}\pi_2^{n_2}...\pi_c^{n_c}.$$

La distribution binomiale est un cas particulier avec $c = 2$. De plus :

$$E(n_j) = n\pi_j, \quad var(n_j) = n\pi_j(1 - \pi_j), \quad cov(n_j, n_k) = -n\pi_j\pi_k.$$

2.3 Distribution de Poisson

C'est le cas où la variable d'intérêt n'est pas le résultat d'une somme finie de tests (ex : nombre de morts sur les routes en une semaine en Italie). Il n'y a pas de limite n à la valeur de y). La loi de probabilité de Poisson est :

$$p(y) = \frac{e^{-\mu}\mu^y}{y!}, \quad y = 0, 1, 2, ...$$

De plus, $E(Y) = var(Y) = \mu$. Cette loi est utilisée pour compter des événements survenant aléatoirement dans le temps ou dans l'espace. C'est aussi une approximation de la loi binomiale pour n grand et π petit, avec $\mu = n\pi$.

Exemple : Si chacun des 50 millions de conducteurs en Italie a la probabilité 0.000002 de mourir dans un accident de la route cette semaine, alors le nombre de morts Y a la fin de la semaine est distribué selon une $Bin(50000000, 0.000002)$, ou approximativement selon une loi de Poisson avec $\mu = n\pi = 50000000 \times 0.000002 = 100$.

Remarque : La moyenne est égale à la variance. Cela implique que la variation est plus grande lorsque la moyenne est grande.

2.4 Sur-dispersion

Dans la pratique, le comptage d'événements présente souvent une variabilité plus grande que celle prédite par les lois binomiales ou de Poisson. Ce phénomène est appelé **surdispersion**.

En effet, nous avons supposé dans l'exemple précédent que chaque personne avait la même probabilité de mourir dans un accident de la route pendant la semaine. En réalité, cette probabilité varie selon de multiples facteurs tels que : nbre de km parcourus, port de la ceinture de sécurité, localisation géographique, etc.

Cela induit des variations plus grandes que celles prédites par le modèle de Poisson, qui est souvent trop simple pour représenter ce type de variables. On utilise souvent à la place le modèle négatif-binomial, qui permet à la variance d'être supérieure à la moyenne.

Si l'on suppose des distributions binomiales ou multinomiales, le phénomène de surdispersion peut survenir lorsque la "vraie" distribution est en fait un mélange de différentes distributions binomiales dont les paramètres sont liés à un phénomène non mesuré.

2.5 Lien entre modèle Poissonien et multinomial

En Italie cette semaine : Soit y_1 = nombre de personnes mortes en voitures, y_2 = nombre de personnes mortes en avion, et y_3 = nombre de personnes mortes en train. Un modèle de Poisson pour (Y_1, Y_2, Y_3) considère ces variables comme indépendantes de paramètres (μ_1, μ_2, μ_3). La loi jointe des $\{Y_i\}$est le produit des densités. Le total $n = \sum Y_i$ a une loi de Poisson de paramètre $\sum \mu_i$.

Le modèle Poissonien suppose que n est aléatoire et non fixé. Si l'on suppose un modèle Poissonien et que l'on conditionne par rapport à n, $\{Y_i\}$ n'a plus une loi de Poisson car chaque $Y_i < n$. On a :

$$\begin{aligned} & P\left[(Y_1 = n_1, Y_2 = n_2, ..., Y_c = n_c) \mid \sum Y_i = n\right] \\ & = \frac{P\left(Y_1 = n_1, Y_2 = n_2, ..., Y_c = n_c\right)}{P\left(\sum Y_i = n\right)} \\ & = \frac{\prod_i \left[\exp(-\mu_i)\mu_i^{n_i}/n_i!\right]}{\exp\left(-\sum \mu_j\right)\left(\sum \mu_j\right)^n/n!} = \frac{n!}{\prod_i n_i!}\prod_i \pi_i^{n_i}, \end{aligned}$$

avec $\pi_i = \mu_i / \sum \mu_j$. Il s'agit de la distribution multinomiale $(n, \{\pi_i\})$.

3 Description de tables de contingence

3.1 Structure probabiliste pour tables de contingence

3.1.1 Tables de contingence et distributions

Soit X et Y deux variables à réponses categorielles, X et Y possédant respectivement I et J catégories. Chaque individu est associé à une des $I \times J$ combinaisons de (X, Y). La distribution de (X, Y) est représentée par une table à I entrées pour la catégorie X, et J entrées pour la catégorie Y, chaque cellule représentant une des $(I \times J)$ combinaisons. Cette table s'appelle une table de contingence (Karl pearson 1904) ou une table de classification croisée ou table $I \times J$.

La table 1 donne un exemple de table de contingence, extraite d'un rapport analysant la relation entre la prise régulière d'aspirine et l'occurrence d'un infarctus du myocarde chez les physiciens. L'étude de 5 ans a analysé de manière aléatoire (randomized study) 11034 physiciens prenant soit de l'aspirine, soit un placebo (sans savoir lequel des deux), en cherchant à vérifier l'hypothèse selon laquelle la prise d'aspirine réduirait l'occurrence d'infarctus.

Notons π_{ij} la probabilité que $P(X = i, Y = j)$. La distribution π_{ij} est la **distribution jointe** de X et Y. Les **distributions marginales** sont les totaux par ligne et colonne

	Myocardial Infarction		
	Fatal attack	nonfatal attack	no Attack
Placebo	18	171	10.845
Aspirin	5	99	10.933

TAB. 1 – Classification croisée de l'usage d'aspirine et de la présence d'un infarctus du myocarde. Source : Preliminary report : Findings from the aspirin component of the ongoing Physicians' Health Study. New Engl. J. Med. 318 : 262-264 1988.

des valeurs de π_{ij}. On note π_{i+} la distribution marginale en ligne et π_{+j} la distribution marginale en colonne, le + indiquant sur quel indice porte la somme :

$$P(X=i)=\pi_{i+}=\sum_j \pi_{ij} \quad \text{et} \quad P(Y=j)=\pi_{+j}=\sum_i \pi_{ij}.$$

Avec $\sum_i \pi_{i+}=\sum_j \pi_{+j}=\sum_i\sum_j \pi_{ij}=1$. La distribution marginale apporte une information sur une seule variable.

En général : Y =Réponse, et X =variable explicative.

Pour une catégorie fixée de X, Y a une certaine distribution. On cherche à étudier ses variations en fonction des variations de X. On note $\pi_{j|i}$ la probabilité pour un individu fixé d'appartenir à la catégorie j $(Y=j)$ sachant que $X=i$:

$$\pi_{j|i}=P\left(Y=j \mid X=i\right)=\frac{\pi_{ij}}{\pi_{i+}}.$$

On appelle **distribution conditionnelle** de y sachant $X=i$ le vecteur de probabilité $(\pi_{1|i},...,\pi_{J|i})$.

En pratique bien entendu la distribution jointe, les distributions marginales et conditionnelles doivent être estimées à partir des réalisations des variables (X,Y). Notons n le nombre total d'observations et N_{ij} le nombre d'observations pour lesquelles $X=i$ et $Y=j$. Les réalisations de N_{ij}, notées par n_{ij}, peuvent être mises dans un tableau avec I lignes et J colonnes. Ce tableau est appelé un **tableau de contingence** $I\times J$, de dimension 2 ou à deux entrées. La distribution conjointe est estimée d'une façon naturelle par les fréquences :

$$p_{ij}=\frac{N_{ij}}{n}.$$

La **distribution marginale** de X s'estime par les fréquences marginales :

$$p_{i+}=\sum_{j=1}^{J} p_{ij} \qquad \forall 1\leqslant i\leqslant I.$$

Et il en est de même pour la loi marginale de Y. On utilisera fréquemment les notations suivantes :

$$n_{i+}=\sum_{j=1}^{J} n_{ij} \qquad\qquad n_{+j}=\sum_{i=1}^{I} n_{ij}$$

qui permettent d'écrire

$$p_{i+} = \frac{n_{i+}}{n} \qquad \text{et} \qquad p_{+j} = \frac{n_{+j}}{n}.$$

A partir du tableau de contingence, nous pouvons estimer facilement la distribution conditionnelle de Y, étant donné $X = i$ de la manière suivante :

$$p_{j|i} = \frac{n_{ij}}{n_{i+}}.$$

Les notions et définitions présentées ci dessus se généralisent facilement au cas d'un tableau de contingence à plusieurs entrées, et serviront dans l'exposé détaillé des tests d'indépendances. En effet, de l'usage des tables de contingence découle la mise au point de critères statistiques permettant de juger de l'importance de la liaison entre deux variables.

	Colonne	
Ligne	j	Total
i	n_{ij} $\left(p_{ij} = \frac{n_{ij}}{N}\right)$	n_{i+} $\left(p_{i+} = \frac{n_{i+}}{N}\right)$
Total	n_{+j} $\left(p_{+j} = \frac{n_{+j}}{N}\right)$	N (1.0)

Tab. 2 – Notations pour les tables de contingence.

	Colonne		
Ligne	1	2	Total
1	π_{11} $(\pi_{1\|1})$	π_{12} $(\pi_{2\|1})$	π_{1+} (1.0)
2	π_{21} $(\pi_{1\|2})$	π_{22} $(\pi_{2\|2})$	π_{2+} (1.0)
Total	π_{+1}	π_{+2}	1.0

Tab. 3 – Notations pour les probabilités jointes, conditionnelles et marginales.

3.1.2 Sensibilité et spécificité

Dans le cas d'une étude diagnostic, ces termes font référence à un diagnostic correct.
– Sensibilité = Le sujet est malade et le diagnostic est positif.
– Spécificité = Le sujet est sain et le diagnostic est négatif.

Cancer du poumon	Diagnostic		
	Positif	Negatif	Total
Oui	0.82	0.18	1.0
Non	0.01	0.99	1.0

TAB. 4 – Distributions conditionnelles estimées pour le diagnostic du cancer du poumon. Source : Data from W. Lawrence et al., J. Natl. Cancer Inst. 90 : 1792-1800 1998.

Pour une table 2×2 du format de l'exemple, la sensibilité est $\pi_{1|1}$ tandis que la spécificité est $\pi_{2|2}$.

Dans l'exemple de la table 4 : Sensibilité estimée = 0.82, Spécificité estimée = 0.99.

3.1.3 Indépendance de variable catégorielles

La distribution conditionnelle de Y sachant X s'écrit en fonction de la distribution jointe :

$$\pi_{j|i} = \pi_{ij}/\pi_{i+} \quad \text{pour tout } i, j.$$

On dit que deux variables catégorielles sont indépendantes si la distribution jointe est le produit des distributions marginales.

$$\pi_{ij} = \pi_{i+}\pi_{+j} \quad \forall i = 1, ..., I \quad \text{and} \quad j = 1, ..., J.$$

Pour une table de contingence, cela signifie que la réponse en colonne est identique quelque soit la ligne considérée. Lorsque X et Y sont indépendantes :

$$\pi_{j|i} = \pi_{ij}/\pi_{i+} = (\pi_{i+}\pi_{+j})/\pi_{i+} = \pi_{+j} \quad \text{pour } i = 1, ..., I.$$

Chacune des distributions conditionnelles de Y est identique à la distribution marginale. On peut donc dire que 2 variables sont indépendantes lorsque $\{\pi_{j|1} = ... = \pi_{j|I}, \text{ for } j = 1, ...J\}$; i.e. la probabilité de la réponse (colonne) est identique pour chaque ligne. On dit aussi qu'il y a homogeneité des distributions conditionnelles.

3.1.4 Distributions de Poisson, binomiales et multinomiales

Les distributions introduites précédemment peuvent s'étendre au cas des effectifs dans une table de contingence. En effet, un modèle Poissonien considère les effectifs Y_{ij} comme des variables indépendantes Poisson(μ_{ij}). La probabilité jointe d'observer les effectifs n_{ij} est donc le produit des probabilités $P(Y_{ij} = n_{ij})$ pour chaque cellule du tableau :

$$\prod_i \prod_j \exp(-\mu_{ij})\, \mu_{ij}^{n_{ij}}/n_{ij}!$$

Lorsque la taille totale n de l'échantillon est fixée, mais pas les effectifs des totaux des lignes et colonnes, alors c'est le modèle multinomial qui s'applique, les effectifs des IJ cellules étant les valeurs prises. La probabilité jointe d'observer les effectifs n_{ij} est donc :

$$[n!/(n_{11}!...n_{IJ}!)] \prod_i \prod_j \pi_{ij}^{n_{ij}}.$$

Souvent, les valeurs de la réponse Y sont observées séparément pour chaque valeur de X. Dans ce cas là, le total des lignes est fixé (par simplicité, on utilise la notation $n_i = n_{i+}$). On suppose alors que les n_i observations de Y pour $X = i$ sont indépendantes de probabilités $(\pi_{1|i}, ..., \pi_{J|i})$. Les effectifs n_{ij} pour $j = 1...J$ vérifient donc $\sum_j n_{ij} = n_i$ et sont répartis selon la distribution multinomiale :

$$\frac{n_i!}{\prod_j n_{ij}!} \prod_j \pi_{j|i}^{n_{ij}}. \tag{1}$$

Si les tirages sont indépendants selon les différents niveaux de la variable X, alors la probabilité jointe des effectifs du tableau total est le produit des multinomiales (1). Il s'agit du schéma multinomial indépendant.

Il arrive que les totaux en lignes et en colonnes soient fixés naturellement par l'expérience. On est alors dans le cas moins fréquent du schéma hypergéométrique.

3.1.5 Exemple des ceintures de sécurité

Des chercheurs du Massachussets ont analysé des conducteurs impliqués dans un accident de la route, en étudiant la relation entre le port de la ceinture (oui,non) et les dégats de l'accident (mort,survie). Chaque accident à venir sera donc enregistré et noté dans une table de contingence (cf table 5).

	Dégats de l'accident	
Usage de la ceinture	Mort	Survie
Oui		
Non		

Tab. 5 – Usage de la ceinture de sécurité et dégâts de l'accident.

La taille de l'échantillon est donc aléatoire : On peut traiter les effectifs comme des variables de Poisson indépendantes de paramètres $(\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22})$.

Supposons maintenant que les chercheurs tirent aléatoirement 200 accidents parmi les archives de la police. La taille n de l'échantillon est alors fixée. Il est possible de modéliser les effectifs des 4 cellules par une loi multinomiale $(200, (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}))$.

Supposons maintenant que les registres d'accidents soient séparés selon qu'il y ait eu mort d'homme ou non. Les chercheurs peuvent donc tirer aléatoirement 100 accidents ayant entrainés la mort, et 100 autres non. En procédant ainsi, on fixe le total des colonnes à 100, et on peut donc considérer que les effectifs de chaque colonne sont issus de lois binomiales indépendantes.

Encore une autre approche, plus traditionnelle, consiste à prendre 200 sujets et d'en assigner aléatoirement 100 d'entre eux au port de la ceinture, les 100 autres n'en portant pas. On force ensuite les 200 sujets à avoir un accident ... Le total des lignes est ainsi fixé, et les effectifs de chaque ligne seront donc indépendants issus d'un schéma binomial.

Rq : le design de l'expérience dépend de ce que l'on étudie, et du type de résultats que l'on veut obtenir.

3.1.6 Les différents types d'études

	Cancer du Poumon	
Fumeur	Cas	Témoins
Oui	688	650
Non	21	59
Total	709	709

TAB. 6 – Classification croisée entre le fait d'être fumeur, et la présence d'un cancer du poumon.

A la table 6 sont présentés les résultats d'une étude des liens entre le cancer du poumon et le fait d'être fumeur (def ici : fumer au moins une cigarette par jour l'année précédent la question). Dans 20 hôpitaux anglais ont été identifiés 709 patients atteints d'un cancer des poumons : Ce sont les cas étudiés. Pour chacun d'entre eux, 709 patients n'ayant pas de cancer ont aussi été interrogés sur leur comportement vis à vis de la cigarette. Ce sont les témoins.

En général, la variable "présence du cancer" est la variable réponse, et le fait de fumer la variable explicative. Dans ce cas précis, la distribution marginale de la var cancer est fixée, et c'est le fait d'avoir été fumeur qui est la variable réponse. Ce type d'étude s'appelle une **étude cas-témoin**.

Exemple d'objectif : Comparer les fumeurs et les non-fumeurs en termes de proportion de cancer. Ce type d'étude nous donne au contraire la distribution de fumeurs selon la variable cancer. Pour ceux ayant un cancer la proportion de fumeurs est $688/709 = 0.970$, mais seulement de $650/709 = 0.917$ pour les témoins. Ce type d'étude ne permet pas d'estimer la probabilité de cancer selon le nombre de cigarettes fumées.

Autre type d'étude : Les sujets sont sélectionnés aléatoirement dans la population des jeunes de 20 ans, puis on mesure le taux de cancers du poumon 60 ans après pour les fumeurs et les non-fumeurs. Ce type d'étude est prospectif et il en existe deux sortes :

- **Essais cliniques :** Les sujets sont assignés aléatoirement à la catégorie fumeur/non-fumeur.
- **Cohortes :** Les sujets font eux mêmes le choix de fumer ou non.

4 Comparaisons de deux proportions

Cas où la réponse est binaire (succes/échec) et deux groupes sont étudiés (Table 2x2). Les groupes sont en ligne, les catégories de Y en colonnes.

4.1 Difference de proportions

Pour la ligne i, $\pi_{1|i}$ est la proba conditionnelle $P(Y = 1)$, avec $\pi_{2|i} = 1 - \pi_{1|i}$. On note $\pi_i = \pi_{1|i}$. La comparaison la plus basique est la différence $\pi_1 - \pi_2$ qui vaut 0 lorsque les lignes ont une distribution conditionnelle identique. Il est équivalent de comparer les proportions de succès que les proportions d'échecs.

4.2 Risque relatif

Si on étudie l'effet d'un traitement médical sur la survie des patients, la différence entre 0.010 et 0.001 peut avoir plus de signification qu'une différence entre 0.410 et 0.401. Le risque relatif est défini par :

$$\pi_1/\pi_2.$$

L'indépendance est atteinte lorsque le risque relatif est proche de 1. Ici : $0.010/0.001 = 10$ et $0.410/0.401 = 1.02$.

4.3 Rapport de côtes (odds ratio)

Pour une proba de succès π, la **côte** est définie par

$$\Omega = \pi/(1 - \pi).$$

Quand $\Omega > 1$, un succès est plus probable qu'un échec. Considérons les tables 2x2. Pour la ligne i, la côte est $\Omega_i = \pi_i/(1 - \pi_i)$. Le **rapport des côtes** est défini par :

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_1/(1 - \pi_1)}{\pi_2/(1 - \pi_2)}. \tag{2}$$

Dans le cas de distribution jointes pour (X, Y), la définition équivalente est :

$$\theta = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}.$$

4.4 Propriétés du rapport de côtes

On a $\theta > 0$ avec $\theta = 1$ correspondant à l'indépendance de X et Y. Lorsque $1 < \theta < \infty$, les sujets de la ligne 1 ont un succès plus probable que ceux de la ligne 2 ($\pi_1 > \pi_2$).

Remarque : Si $\theta = 4$, la côte (odds) de succès est 4 fois plus grande pour la ligne 1 que pour la ligne 2. Cela ne veut pas dire que $\pi_1 = 4 \times \pi_2$, qui correspond à l'interprétation d'un "risque relatif" égal à 4.

Il est souvent plus pratique d'étudier $\log \theta$. L'indépendance correspond à $\log \theta = 0$, et le log odd ratios est symétrique par rapport à cette valeur. Deux valeurs pour identiques de $\log \theta$ mais de signes opposés ont la même force d'association (exemple $\log 4 = 1.39$ et $\log 0.25 = -1.39$). L'odd ratio ne chnage pas lorsque l'on change l'orientation de la table. Il

n'est donc pas nécessaire d'identifier la variable réponse pour calculer l'odds ratio, même si la définition 2 utilise $\pi_i = P(Y = 1|X = i)$. Ainsi on a :

$$\begin{aligned}\theta &= \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = \frac{P(Y = 1|X = 1)/P(Y = 2|X = 1)}{P(Y = 1|X = 2)/P(Y = 2|X = 2)} \\ &= \frac{P(X = 1|Y = 1)/P(X = 2|Y = 1)}{P(X = 1|Y = 2)/P(X = 2|Y = 2)}.\end{aligned}$$

Pour une table de contingence contenant des fréquences, l'estimation de l'odd-ratio est :

$$\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

4.4.1 Exemple Aspirine et infarctus :

Reprenons la table 1 et fusionnons les infarctus ayant entraîné la mort et ceux qui ne l'ont pas entraîné. Parmi les 11034 physiciens prenant le placebo, 189 ont eu un infarctus, soit une proportion de $189/11034 = 0.0171$. Parmi ceux prenant de l'aspirine, la proportion est de $104/11037 = 0.0094$. La différence des proportions est de $0.0171 - 0.0094 = 0.0077$, et le risque relatif est $0.0171/0.0094 = 1.82$. La proportion d'infarctus chez ceux prenant un placebo est 1.82 fois la proportion chez ceux prenant de l'aspirine. L'estimation des odds ratio est $(189 \times 10933)/(10845 \times 104) = 1.83$. La côte (odds) d'un infarctus pour ceux prenant un placebo est 1.83 fois la côte pour ceux prenant de l'aspirine.

4.4.2 Etudes cas-témoins et Odds Ratios :

Dans ce cas, il n'est en général pas possible d'estimer $P(Y = j|X = i)$, mais il est souvent possible d'estimer l'OR. En effet, illustrons le en examinant à nouveau la table 6. Il s'agit de deux échantillons binomiaux de $X =$ fumeur pour des niveaux fixés de $Y =$ présence d'un cancer du poumon. On peut estimer la probabilité qu'un sujet soit un fumeur, sachant qu'il ait un cancer du poumon. Soit 688/709 pour les cas, et 650/709 pour les témoins. On ne peut pas estimer la probabilité d'avoir un cancer sachant le fait de fumer, ce qui serait pourtant plus intéressant. On ne peut donc pas estimer la différence de proportion, ni le rapport des probabilités de présence d'un cancer. La différence des proportions et le risque relatif sont limités à la comparaison des probabilités d'être un fumeur. On peut par contre calculer le rapport de côtes (odds-ratio) :

$$\frac{(688/709)/(21/709)}{(650/709)/(59/709)} = \frac{688 \times 59}{650 \times 21} = 3.$$

On peut donc dire que l'estimation de la probabilité de présence d'un cancer pour les fumeurs est 3 fois plus élevée que pour les non-fumeurs.

L'Odds Ratio comme mesure d'association : L'OR s'interprète comme une mesure d'association. S'il est supérieur à 1 la relation est croissante, et décroissante s'il est inférieur à 1. Lorsqu'il est égal à 1 il n'y a pas d'association. La valeur de l'odds ratio indique la direction ainsi que la force de l'association.

L'Odds Ratio comme mesure du risque relatif (RR) :

$$\text{odds ratio} = \text{relative risks} \left(\frac{1-\pi_2}{1-\pi_1} \right)$$

Lorsque la prévalence de l'événement à expliquer est faible (π_1 et π_2 sont petites), l'odds ratio fournit une approximation du risque relatif. C'est le cas de l'exemple "aspirine et infarctus" où le risque relatif (1.82) et l'odds-ration (1.83) sont très proches.

4.5 Association partielle dans une table 2×2

Dans les études expérimentales (observational studies), des variables extérieures peuvent avoir un impact sur la relation entre X et Y. Il faut donc contrôler ces facteurs de confusion (confoundig factors) pour être sûr que l'effet observé n'est pas un effet indirect de la covariable. On peut contrôler les effets des covariables en assignant de façon aléatoire les sujets à l'exposition des différentes modalités de la variable X, mais cela n'est pas toujours possible.

Exemple : Si l'on étudie les effets de la fumée passive dans un couple (i.e. un non-fumeur qui vit au quotidien avec un fumeur) sur la présence d'un cancer du poumon, une étude peut s'intéresser à comparer les personnes non-fumeur mariées à un(e) fumeur avec les personnes non-fumeur mariées à un(e) non-fumeur. Il est possible que les conjoints de non-fumeurs soient plus jeunes en moyenne que les conjoints de fumeurs, et les jeunes ont moins de cancer. Une proportion plus faible cancer chez les époux de non-fumeurs peut simplement représenter le fait qu'ils sont en moyenne plus jeunes, et non pas les effets de la fumée passive.

Il faut donc contrôler les covariables succeptibles d'avoir une influence (ici par exemple : age ou CSP).

4.6 Tables partielles

On peut contrôler les effets d'une covariable Z en étudiant la relation XY pour des niveaux fixés de Z. On construit des tables 2×2 en extrayant des "tranches" de la table à 3 entrées XYZ pour des niveaux fixés de Z. Ce sont des **tables partielles**. On peut combiner ces tables partielles dans une unique table appelée **table marginale XY** en sommant les fréquences quelque soit la valeur de Z. Ces tables ignorent Z et ne la contrôlent pas. Elles ne contiennent aucune information sur Z.

Les associations contenues dans un table partielle sont appellées **associations conditionnelles** et peuvent être différentes de celles contenues dans une table marginale. Il peut être dangereux d'analyser uniquement les associations dans les tables partielles.

5 Extension aux tables $I \times J$

Dans le cas d'une table 2×2, une unique valeur (odds-ratios) peut résumer l'association entre les variables, mais cela n'est généralement pas possible pour les tables $I \times J$ sans perdre de l'information.

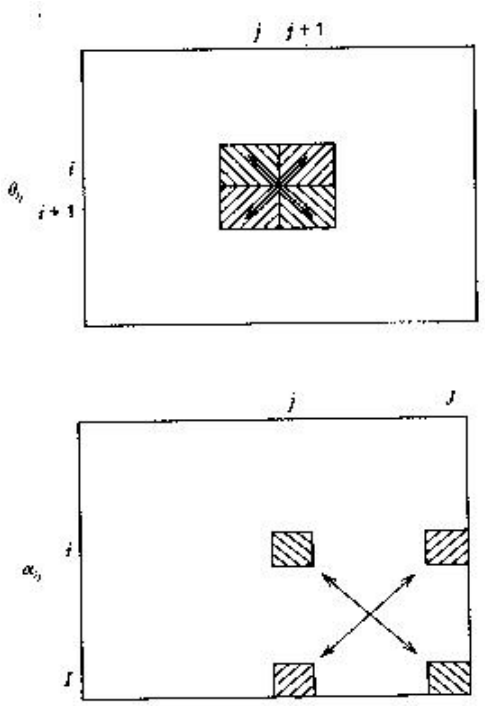


FIG. 1 – Odds ratios locaux.

5.1 Odds ratios pour tables $I \times J$

Si l'on considère toutes les associations possibles, il y a redondance de l'information. On considère généralement les $(I-1)(J-1)$ odds-rations locaux suivants :

$$\theta_{ij} = \frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{i,j+1}\pi_{i+1,j}} \qquad I = 1, ..., I-1 \qquad J = 1, ..., J-1.$$

Les cellules utilisées dans ce cas là sont adjacentes. Dans la table 1 l'odds ratio local estimé est de 2.08 pour les deux premières colonnes et de 1.74 pour les deux dernières. Ce qui signifie que dans chaque cas le cas le plus grave est plus probable pour le groupe prenant le placebo (fatal contre non-fatal, et non-fatal contre "pas d'infarctus"). Le produit des deux odds-ratios locaux est 3.63, qui est l'odds ratio entre la première colonne et la dernière.

On peut définir d'autres odds-ratios locaux :

$$\alpha_{ij} = \frac{\pi_{ij}\pi_{IJ}}{\pi_{iJ}\pi_{Ij}} \qquad I = 1, ..., I-1 \qquad J = 1, ..., J-1.$$

Une illustration de ces deux possibilités est donnée à la figure 1.

5.2 Coefficient d'association entre variables ordinales

Si les variables X et Y sont ordinales, il existe des coefficients d'association particuliers. Supposons que les numéros des lignes et des colonnes correspondent à l'ordre "naturel", c'est à dire que si $i < i'$, la modalité i vaut "moins" que la modalité i'. Le coefficient introduit

dans ce paragraphe mesure si de "grandes" valeurs de X ont tendance à être réalisés pour de "grandes" valeurs de Y. C'est donc une mesure de monoticité pour des variables qualitatives ordinales.

Le coefficient le plus souvent utilisé pour mesurer des dépendances monotones entre les lignes et les colonnes d'un tableau de contingence est le γ de Goodman et Kruskall. Prenons deux couples (X_1, X_2) et (Y_1, Y_2) indépendants et ayant la même distribution que (X, Y). On appelle C la probabilité que les deux couples soient en accord positif, et D la probabilité qu'ils soient en accord négatif. On définit

$$C = P\left(X_1 > X_2 \text{ et } Y_1 > Y_2\right) + P\left(X_2 > X_1 \text{ et } Y_2 > Y_1\right)$$

et

$$D = P\left(X_1 < X_2 \text{ et } Y_1 > Y_2\right) + P\left(X_2 > X_1 \text{ et } Y_2 < Y_1\right).$$

Le coefficient γ est donné par

$$\gamma = \frac{C - D}{C + D}.$$

Ce coefficient est compris entre -1 et 1 et si X et Y sont indépendantes, alors $\gamma = 0$. Plus γ est grand, plus il y a un lien entre les 2 variables qualitatives. Si γ est négatif, alors la liaison est negative.

5.3 Coefficient kappa de Cohen

La dernière mesure d'association traitée est le kappa de Cohen. Ce coefficient s'applique uniquement dans un contexte bien particulier, ce qui met cette section un peu à l'écart. Supposons que plusieurs juges ou observateurs doivent classer N objets. Ce classement se fait, au moins partiellement, sur la base de critères subjectifs. Ce qui implique que les juges pourront classer les objets différemment. Le coefficient κ mesure maintenant le degré d'accord entre les juges. Il est défini d'une façon générale par :

$$\kappa = \frac{p_A - p_E}{1 - p_E}$$

où p_A est la probabilité que les observateurs émettent le même jugement, et p_E est la probabilité que les juges soient d'accord au cas où ils jugeraient au hasard. Le coefficient κ est compris entre 0 et 1. En cas d'accord parfait entre les juges, κ égal 1. S'ils jugent de façon aléatoire, κ est égal à 0. Plus il y a de catégories, plus il est difficile d'atteindre un accord parfait. Par exemple, un κ égal à 0.5 est plutot bon si la variable Y a 5 catégories.

6 Inférence dans les tables de contingence

6.1 Intervalle de confiance pour les paramètres d'association

6.1.1 Estimation par intervalle des odds-ratios

L'odds ratio estimé $\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$ d'une table 2×2 vaut 0 ou ∞ si un des $n_{ij} = 0$, et indéfini si deux d'entre eux sont égaux à 0. La probabilité que cela arrive n'est pas nulle, et $E(\hat{\theta})$ et

$Var(\hat{\theta})$ ne sont donc pas définies. On peut remplacer $\hat{\theta}$ par

$$\tilde{\theta} = \frac{(n_{11}+0.5)(n_{22}+0.5)}{(n_{12}+0.5)(n_{21}+0.5)},$$

qui est moins biaisé que $\hat{\theta}$ et qui converge asymptotiquement vers une loi normale centrée en θ. Pour construire un intervalle de confiance pour θ, on utilise plutot $log(\hat{\theta})$ et on montre que :

$$\hat{\sigma}\left(log(\hat{\theta})\right) = \left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right)^{1/2}.$$

En utilisant l'approximation asymptotique par la loi normale, on a alors

$$log(\hat{\theta}) \pm z_{\alpha/2}\hat{\sigma}\left(log(\hat{\theta})\right)$$

qui est intervale de confiance pour $\log(\theta)$.

6.1.2 Estimation par intervalles des différences de proportions

On peut considérer ici que les échantillons y_i de chacun des groupes comparés sont indépendants de loi binomiale (n_i, π_i). On a $\hat{\pi}_i = \frac{y_i}{n_i}$ d'espérance π_i et de variance $\frac{\pi_i(1-\pi_i)}{n_i}$. On a $\hat{\pi}_1$ et $\hat{\pi}_2$ indépendants, et donc :

$$E\left(\hat{\pi}_1 - \hat{\pi}_2\right) = \pi_1 - \pi_2,$$

et

$$\hat{\sigma}\left(\hat{\pi}_1 - \hat{\pi}_2\right) = \left[\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}\right]^{1/2}.$$

Pour estimer l'écart type, on remplace π_1 et π_2 par leurs estimations, et on obtient l'intervalle de confiance suivant pour $\pi_1 - \pi_2$:

$$\left(\hat{\pi}_1 - \hat{\pi}_2\right) \pm z_{\alpha/2}\hat{\sigma}\left(\hat{\pi}_1 - \hat{\pi}_2\right).$$

6.1.3 Estimation par intervalles pour le risque relatif (RR)

Le risque relatif estimé est $r = \frac{\hat{\pi}_1}{\hat{\pi}_2}$. On utilise ici aussi le $\log(r)$ qui converge plus vite vers la loi normale. On a

$$\hat{\sigma}\left(\log(r)\right) = \left(\frac{1-\pi_1}{n_1} + \frac{1-\pi_2}{n_2}\right)^{1/2}.$$

On peut donc produire un intervalle de confiance pour $\log\left(\frac{\pi_1}{\pi_2}\right)$:

$$\log(r) \pm z_{\alpha/2}\hat{\sigma}\left(\log(r)\right).$$

6.2 Tests d'indépendance pour tables de contingence

Dans cette section, nous allons tester l'indépendance de deux variables qualitatives X (J modalités) et Y (I modalités), ce qui revient à tester l'hypothèse nulle suivante :

$$H_0 : \pi_{ij} = \pi_{i.}\pi_{.j} \qquad \forall 1 \leqslant i \leqslant I, 1 \leqslant j \leqslant J$$

où à tester

$$H_0 : \pi_{j|i} = \pi_{j.} \qquad \forall 1 \leqslant i \leqslant I, 1 \leqslant j \leqslant J.$$

L'hypothèse d'indépendance de X et Y est donc équivalente à l'hypothèse que les distributions conditionnelles soient identiques aux distributions marginales. Dans le langage non mathématique, on dit que l'on teste l'indépendance entre les lignes et les colonnes.

6.2.1 Test du khi-deux

La statistique de test habituellement utilisée pour tester cette H_0 est bien évidemment la statistique du khi-carré introduite en 1900 par Pearson, et définie par :

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

où les n_{ij} sont les différents éléments du tableau de contingence et $\hat{\mu}_{ij} = \frac{n_{i.}n_{.j}}{n}$. La valeur de $\hat{\mu}_{ij}$ est parfois considérée comme la valeur attendue de N_{ij} sous l'hypothèse d'indépendance. La loi asymptotique de X_P^2 sous H_0 et pour $n \to \infty$, est bien connue comme étant une loi du khi-carré avec $df = (I-1)(J-1)$ degrés de liberté. Nous rejetons donc H_0 si la valeur de X_P^2 est plus grande que $\chi^2_{1-\alpha}(df)$, le quantile $(1-\alpha)$ d'une distribution $\chi^2(df)$.

Le test du χ^2 est un test statistique qui aide simplement à prendre une décision quant à la probabilité que les variables soient indépendantes l'une de l'autre ou non dans la population. La valeur du χ^2 n'est pas elle-même un bon indice de la taille ou de l'importance de l'association entre les traits mesurés.

6.2.2 Test du rapport de vraisemblance

Le second test calculé à partir des tables de contingence est celui du rapport de Vraisemblance. Sa statistique de test est :

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\left(\log n_{ij} - \log \hat{\mu}_{ij}\right).$$

Sous l'hypothèse d'indépendance, la statistique calculée suit elle aussi une loi de χ^2 avec le même nombre de degré de liberté que précédemment. Plus les valeurs de X_P^2 et G^2 sont élevées, plus il y a évidence d'une dépendance. On ne peut par contre rien dire sur la force de cette dépendance.

6.2.3 Approfondir un test du chi-deux

Les statistiques X_P^2 et G^2 ne renseignent pas sur la force de l'association. Il faut donc analyser plus finement cette association. On peut par exemple étudier les résidus de Pearson définis par :

$$e_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}}}.$$

Les résidus représentent la contribution de chaque cellule de la table à la statistique X_P^2 car on a $\sum_i \sum_j e_{ij}^2 = X_P^2$. Plus la contribution est importante, plus cette cellule repésente un éloignement par rapport à l'hypothèse d'indépendance. De plus, sous H_0 les $\{e_{ij}\}$ sont asymptotiquement gaussiens centrés. On peut aussi utiliser les résidus de Pearson standardisés définis par :

$$\frac{n_{ij} - \hat{\mu}_{ij}}{\left[\hat{\mu}_{ij}\left(1 - p_{i+}\right)\left(1 - p_{+j}\right)\right]^{1/2}}.$$

Un résidu standardisé supérieur à 2 ou 3 indique une cellule pour laquelle H_0 n'est pas vérifiée.

6.2.4 Limitations du test du chi-deux

Ce test nécessite de grands échantillons, et doit être complété par des analyses sur les résidus et sur les odds-ratios pour comprendre la teneur de l'association. En outre, ce test ne tient pas compte de l'ordre éventuel des modalités de X ou Y. Les statistiques X_P^2 et G^2 ne changent pas quand on modifie l'ordre des colonnes ou des lignes. Si une des variables est ordinale alors il faut utiliser d'autres types de statistiques.

Il est important de noter qu'il existe un test **exact** pour les tables 2×2 et les petits échantillons qui est nommé **test exact de Fisher**.

7 Le modèle linéaire généralisé

Le modèle linéaire généralisé constitue un cadre général permettant d'englober une grande quantité des modèles disponibles. Cette famille de modèles permet d'étudier la liaison entre une variable dépendante ou **réponse** Y et un ensemble de variables explicatives ou **prédicteurs** $X_1, ..., X_K$. Elle englobe le modèle linéaire général (régression multiple, analyse de la variance et analyse de la covariance), le modèle log-linéaire et des techniques de modélisation telles que la régression logistique ou la régression de Poisson. Les modèles linéaires généralisés sont formés de trois composantes :

- La variable de réponse Y, composante aléatoire à laquelle est associée une loi de probabilité.
- Les variables explicatives $X_1, ..., X_K$ utilisées comme prédicteurs dans le modèle, définissent sous forme d'une combinaison linéaire la composante déterministe.
- Le lien décrivant la relation fonctionnelle entre la combinaison linéaire des variables $X_1, ..., X_K$ et l'espérance mathématique de la variable de réponse Y.

Notons $(Y_1, ..., Y_n)$ un échantillon aléatoire de taille n de la variable de réponse Y, les variables aléatoires $Y_1, ..., Y_n$ étant supposées indépendantes. Dans certaines applications chaque variable Y_i est binaire ; on supposera alors que la composante aléatoire est distribuée

selon une loi binomiale. Dans d'autres applications chaque réponse est un effectif distribué selon une loi de Poisson. Si chaque observation provient d'une variable continue, on peut supposer une distribution normale de la composante aléatoire. Les effectifs d'une table de contingence sont en général modélisés par une loi de Poisson.

Concernant la composante déterministe, exprimée sous forme linéaire $\beta_0 + \beta_1 X_1 + ... + \beta_K X_K$, elle précise quels sont les prédicteurs servant à décrire la moyenne de Y que l'on note μ. Certaines des variables X_j peuvent se déduire de variables initiales utilisées dans le modèle. Par exemple on pourra utiliser $X_3 = X_1 \times X_2$ de façon à étudier l'interaction entre X_1 et X_2.

La troisième composante d'un modèle linéaire généralisé est le lien entre la composante aléatoire et la composante déterministe. Elle spécifie comment l'espérance mathématique de Y, notée μ, est liée au prédicteur linéaire construit à partir des variables explicatives. On peut modéliser une fonction monotone $g(\mu)$ de l'espérance. On a alors :

$$g(\mu) = \beta_0 + \beta_1 X_1 + ... + \beta_K X_K.$$

La fonction g est appelée fonction de lien. La fonction de lien $g(\mu) = \log(\mu)$ permet par exemple de modéliser le logarithme de l'espérance et donne lieu aux modèles log-linéaires.

La fonction de lien $g(\mu) = \log \frac{\mu}{1-\mu}$ modélise le logarithme du rapport des chances. Elle est appelée logit et est adaptée au cas où μ est compris entre 0 et 1. C'est ce que l'on appelle la régression logistique.

Le choix du modèle linéaire généralisé dépend de la nature des données que l'on souhaite étudier. Le tableau 7 résume ces différents cas.

Composante aléatoire	Lien $g(\mu)$	Nature des variables de la composante déterministe	Modèle
Normale	Identité	Quantitatives	Régression
Normale	Identité	Qualitatives	Analyse de la variance
Normale	Identité	Mixtes	Analyse de la covariance
Binomiale	Logit	Mixtes	Régression logistique
Poisson	Log	Mixtes	Modèles log-linéaires
Multinomiale	Logit généralisé	Mixtes	Modèles à réponses multinomiales

TAB. 7 – Types de modèles couverts par le modèle linéaire généralisé.

L'avantage des modèles linéaires généralisés est de fournir un cadre théorique adapté aussi bien à la modélisation de variables quantitatives que qualitatives. L'unification des différents types de modèles a permis de disposer de toute une batterie de techniques statistiques rigoureuses permettant d'aider le statisticien dans son choix de modèle. On dispose notamment de plusieurs tests statistiques permettant de tester l'adéquation du modèle aux données.

Deux statistiques sont utiles pour juger de cette adéquation :
- La déviance normalisée
- La statistique du khi-deux de Pearson.

Un modèle linéaire généralisé est défini par la loi de probabilité $f(y, \theta)$ de la réponse Y et la nature de la fonction de lien g reliant l'espérance μ de Y aux variables explicatives $X_1, ..., X_K$, $g(\mu_i) = x_i'\beta$.

On note b l'estimation du maximum de vraisemblance de β. Pour mesurer l'adéquation du modèle étudié aux données, on construit tout d'abord un modèle saturé (basé sur la même loi de probabilité et la même fonction de lien) contenant autant de variables explicatives indépendantes que de données : ce modèle permet de reconstruire parfaitement les données. On note $b_{\max}$ l'estimation du vecteur des paramètres β pour ce modèle.

L'adéquation du modèle étudié est alors mesuré à l'aide de la statistique D^* appelée **déviance normalisée** :

$$D^* = 2 \log \lambda = 2 \log \left(\frac{L(b_{\max}; y)}{L(b; y)} \right),$$

où L est la vraisemblance du modèle. On démontre que lorsque le modèle étudié est exact, la déviance normalisée D^* suit approximativement une loi de khi-deux à $n - K$ degrés de liberté.

La statistique du **khi-deux de Pearson** est définie par

$$\chi^2 = \sum \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)},$$

où V représente la variance. Cette statistique est elle aussi distribuée selon une loi de khi-deux à $n - K$ degrés de liberté si le modèle étudié est exact. Cette mesure de qualité synthétise en fait les résidus de Pearson dont on donne l'expression pour un individu i :

$$r_{P_i} = \frac{y_i - \hat{\mu}_i}{V(\hat{\mu}_i)}.$$

Ils permettent de contrôler l'erreur faite sur chacun des coefficients estimés, et d'établir un diagnostic quand à la validité de celui ci.

Le cadre théorique développé pour les modèles linéaires généralisés permet d'appréhender une grande variété d'approches sensiblement différentes. Les statistiques précédentes permettent de faire un choix de modèle basé sur des critères statistiques rigoureux. Les méthodes abordées dans la section précédente, désignées par le terme générique de méthodes descriptives multidimensionnelles, ne possèdent pas toutes cet arsenal théorique malgré leur avantages.

Avant de décrire la régression logistique, on accordera une attention particulière au modèle Log-Linéaire qui permet de rechercher des interactions entre un certain nombre de variables qualitatives.

7.1 Le modèle Log-linéaire

Le modèle linéaire usuel essaie de prédire ou d'expliquer une variable Y, mesurée sur une échelle continue, à partir d'un ensemble de K variables explicatives $X_1, ..., X_K$ qui peuvent être continues ou catégorielles. L'usage du modèle Log-linéaire est quand à lui plus approprié pour rechercher des relations entre un certain nombre de variables qualitatives. Il

a la particularité de ne pas nécessiter, a priori, de distinction entre la variable à expliquer et les variables explicatives. On parlera plutôt de modèle d'association que de régression.

Considérant que le $k^{ième}$ facteur peut prendre un total de I_k modalités, on construit la table de contingence à K entrées à partir de ces K facteurs. Un tableau de contingence $I_1 \times I_2 \times ... \times I_k$ est ainsi obtenu. L'idée du modèle Log-linéaire est d'expliquer les logarithmes des valeurs attendues des effectifs à l'aide des niveaux correspondants des facteurs et des interactions entre ces niveaux. La fonction de lien utilisé dans ce cas là est la fonction log.

8 La régression logistique

C'est une des méthodes les plus utilisées dans le cas où la variable réponse Y est dichotomique. Lorsqu'elle est polytomique, on parlera plutôt de régression logistique multinomiale mais le principe est identique. Dans son acception classique du terme, la régression logistique fait référence au cas où toutes les variables explicatives sont qualitatives. Il est cependant possible d'utiliser la régression logistique lorsque les variables explicatives sont qualitatives et quantitatives. On se place ici dans le cadre du modèle linéaire généralisé en faisant référence au cas où la fonction de lien utilisée est la fonction logit :

$$g(x) = \frac{x}{1-x},$$

qui est l'inverse de la fonction logistique $F(x) = \frac{e^x}{1+e^x}$. La régression logistique s'attache à modéliser la probabilité π_i que Y_i soit égale à 1 ($\pi_{ij} = P(Y_i = j)$ dans le cas polytomique). C'est à dire :

$$\begin{aligned} \pi_i &= P(Y_i = 1 \mid x_i) = E(Y_i \mid x_i) \\ &= F(x_i\beta), \end{aligned}$$

où $x_i\beta$ est le prédicteur linéaire. La fonction logistique est bien adaptée à la modélisation de probabilités, car elle prend ses valeurs entre 0 et 1 selon une courbe en S. La nécessité d'un modèle particulier est justifiée notamment par le fait que l'utilisation d'un modèle de régression classique n'est pas adéquate (une probabilité doit être comprise entre 0 et 1).

8.1 Pourquoi la régression logistique

Lorsque la variable dépendante n'est pas quantitative mais qualitative ou catégorielle le modèle de régression linéaire n'est pas approprié. Lorsque le nombre d'attributs est de deux l'on parle de variable dichotomique, e.g. le sexe « mâle » ou « femelle », tandis que s'il est supérieur à deux l'on a une variable polytomique, e.g. une pression « haute », « normale » ou « basse ». On a représenté, ci-dessous, différents graphes illustrant les différences fondamentales entre variable qualitative et variable quantitative. Dans la première figure (Figure 2), la variable dépendante est la maladie coronarienne. Cette variable peut prendre les attributs « oui » ou « non » de sorte qu'il n'est pas possible d'écrire une relation directement entre la maladie coronarienne et l'âge. Dans le second graphe (Figure 3), la variable dépendante est quantitative, il s'agit de la taille, de sorte qu'il est possible d'établir directement une relation (linéaire ou pas) entre la taille et l'âge. Le troisième graphe (Figure 4), illustre l'hypothèse de Normalité souvent adoptée lorsque la variable dépendante est quantitative.

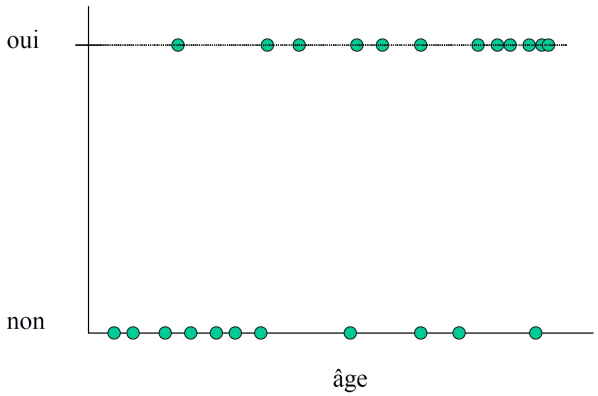


FIG. 2 – Présence ou absence d'une maladie coronarienne (variable CHD de l'exercice en annexe).

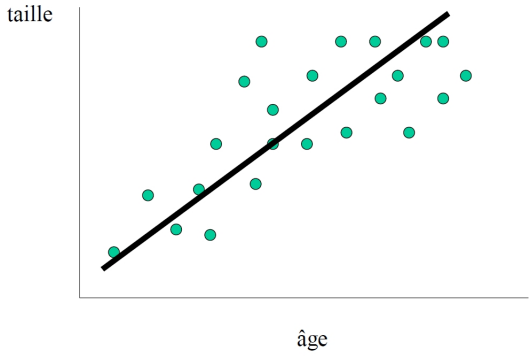


FIG. 3 – Relation entre taille et âge chez les enfants.

lorsque la variable dépendante est qualitative elle n'admet pas d'échelle de mesure naturelle et on modélise par conséquent sa probabilité de prendre tel ou tel attribut. Dans le graphique 5, on a regroupé les données concernant l'âge en catégories et calculé dans chacune de ces catégories le pourcentage de personnes souffrant d'une maladie coronarienne :

On constate que l'on a une relation sigmoïdale, i.e. en forme de S, entre la proportion de maladie coronarienne et l'âge. On en déduit que pour modéliser la probabilité de maladie coronarienne en fonction de l'âge il faudra utiliser une relation sigmoïdale. En effet, une probabilité étant par définition comprise entre 0 et 1 le modèle linéaire n'est bien entendu pas approprié (puisqu'il ne limite pas les valeurs de notre probabilité au domaine compris entre 0 et 1) et la relation est forcément non-linéaire :

8.2 Régression logistique binomiale

Lorsque l'on est en présence d'une variable réponse binaire Y (à valeur 0 ou 1) et d'une variable explicative X, on note $\pi(x) = P(Y = 1 \mid X = x) = 1 - P(Y = 0 \mid X = x)$. Le

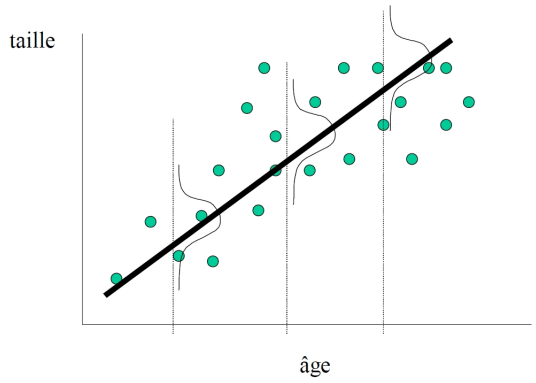


FIG. 4 – Relation entre taille et âge chez les enfants : hypothèse de Normalité.

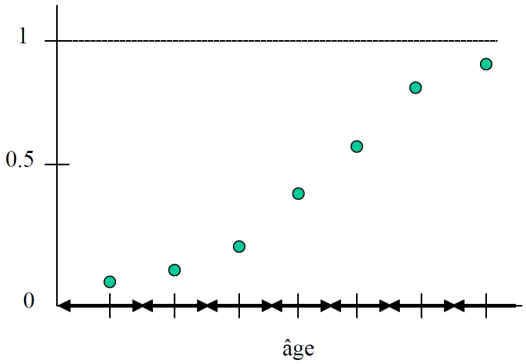


FIG. 5 – Pourcentage de personnes souffrant d'une maladie coronarienne par catégorie d'âge.

modèle de régression logistique consiste à poser

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}.$$

De manière équivalente, on suppose que le log du rapport des côtes (odds ratios dans la littérature anglaise), aussi appelé logit, respecte la relation linéaire suivante :

$$\text{logit}[\pi(x)] = \log \frac{\pi(x)}{1 - \pi(x)} = \beta_0 + \beta_1 x.$$

Le modèle logistique s'écrit donc :

$$E(Y \mid x, \beta_0, \beta_1) = P(Y = 1 \mid X = x, \beta_0, \beta_1) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}.$$

Plusieurs prédicteurs peuvent être pris en compte. On notera dans ce cas x_i la valeur (ou la modalité) prise par le prédicteur X_i.

Remarque : Un choix intuitif pour modéliser une probabilité est d'utiliser une fonction de répartition (en S). Lorsque cette fonction est celle de la loi logistique (de la forme $\frac{\exp(x)}{1+\exp(x)}$) on obtient le modèle de régression logistique. Si l'on utilise la fonction de répartition de la loi normale, on obtient le modèle **probit** (cf figure 8).

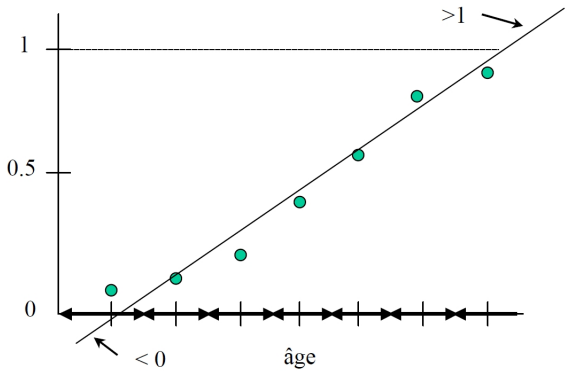


FIG. 6 – Pourcentage de personnes souffrant d'une maladie coronarienne par catégorie d'âge : relation linéaire.

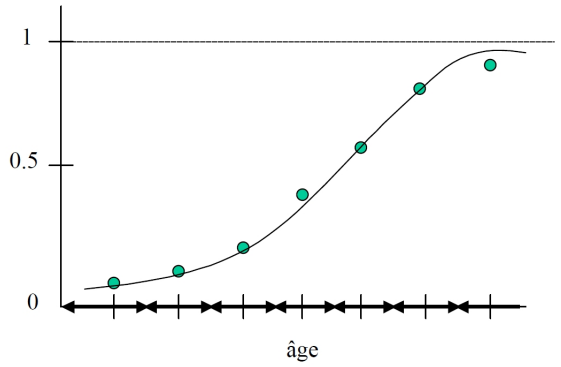


FIG. 7 – Pourcentage de personnes souffrant d'une maladie coronarienne par catégorie d'âge : relation sigmoïdale.

8.3 Régression logistique multinomiale ou polytomique

Lorsque Y représente une réponse qualitative à J catégories (avec J>2), on utilisera la régression logistique multinomiale, appelée aussi régression logistique polytomique. Celle ci consiste à effectuer J-1 régressions logistiques binomiales correspondant aux combinaisons de la catégorie de référence avec les J-1 autres catégories. Dans le cas d'une variable Y à 4 catégories, en prenant comme catégorie de référence celle correspondant à la catégorie n°4, on effectuera donc 3 régressions logistiques binomiales différentes :

$$\log\left(\frac{P\left(Y=1 \mid X=x\right)}{P\left(Y=4 \mid X=x\right)}\right)=\log\left(\frac{\pi_1\left(x\right)}{\pi_4\left(x\right)}\right)=\beta_{01}+\beta_{11}x.$$
$$\log\left(\frac{P\left(Y=2 \mid X=x\right)}{P\left(Y=4 \mid X=x\right)}\right)=\log\left(\frac{\pi_2\left(x\right)}{\pi_4\left(x\right)}\right)=\beta_{02}+\beta_{12}x.$$
$$\log\left(\frac{P\left(Y=3 \mid X=x\right)}{P\left(Y=4 \mid X=x\right)}\right)=\log\left(\frac{\pi_3\left(x\right)}{\pi_4\left(x\right)}\right)=\beta_{03}+\beta_{13}x.$$

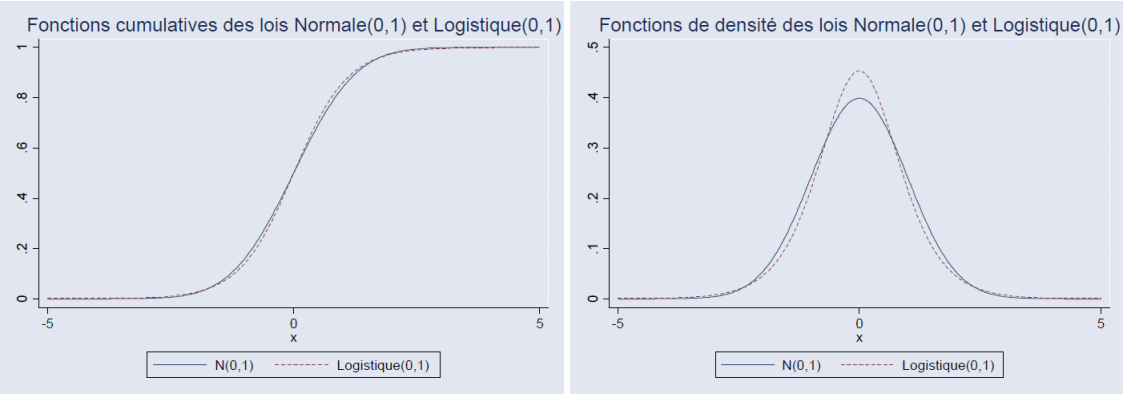


FIG. 8 – Fonction de densité et de répartition pour les lois normale et logistique.

Lorsqu'il y a K prédicteurs, on note $x = (x_1, ..., x_K)$ une valeur de $X = (X_1, ..., X_K)$. Le modèle logistique a alors pour expression :

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + ... + \beta_K x_K)}{1 + \exp(\beta_0 + \beta_1 x_1 + ... + \beta_K x_K)}$$

Estimer un modèle de régression revient à estimer les coefficients β de ce modèle.

8.4 Estimation et tests dans le modèle logistique

8.4.1 Estimation du modèle

On utilise généralement la méthode du maximum de vraisemblance. Lorsque les observations $y_i, i = 1, ..., n$ sont supposées indépendantes, la vraisemblance s'écrit :

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} [P(Y = 1 \mid x, \beta_0, \beta_1)]^{y_i} [1 - P(Y = 1 \mid x, \beta_0, \beta_1)]^{1-y_i}.$$

Remarque : Lorsque l'on est en présence de mesures répétées pour chaque individu, l'hypothèse d'indépendance des données n'est pas plausible. Il faut alors utiliser d'autres méthodes tenant compte de la corrélation des données (modèle marginal avec GEE, modèle logistique conditionnel, modèle mixte).

8.4.2 Tests de significativité des coefficients

Pour tester la significativité d'un ou plusieurs coefficients, par ex. $Ho : \beta_k = 0$ versus $Ha : \beta_k \neq 0$, on utilisera soit le test de Wald W, soit le test du rapport de vraisemblance LR. Dans le cas où l'on veut tester la significativité d'un seul coefficient ces statistiques s'écrivent :

$$W = \frac{\hat{\beta}_k}{\hat{\sigma}\left(\hat{\beta}_k\right)} \longrightarrow N(0, 1),$$

$$LR = -2 \log\left(\frac{L_{H_0}}{L_{H_a}}\right) \longrightarrow \chi^2(1),$$

tandis que si l'on veut tester la significativité de plusieurs coefficients, par ex. $Ho : \beta_1 = \beta_2 = ... = \beta_M = 0$, alors elles s'écrivent :

$$W = \hat{\beta}'(\hat{\sigma}\left(\hat{\beta}\right))^{-1}\hat{\beta} \longrightarrow \chi^2(M),$$
$$LR = -2\log\left(\frac{L_{H_0}}{L_{H_a}}\right) \longrightarrow \chi^2(M),$$

où L_{H_0} est la vraisemblance évaluée sous la contrainte Ho et L_{H_a} la vraisemblance non contrainte. La statistique de Wald fait intervenir les expressions matricielles suivantes :

$$\hat{\sigma}\left(\hat{\beta}\right) = \left(X'VX\right)^{-1}, V = \begin{bmatrix} \hat{p}_1(1-\hat{p}_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \hat{p}_n(1-\hat{p}_n) \end{bmatrix} \text{ et } X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}.$$

8.5 Logit et odds ratios

Prenons le cas d'un modèle comportant une seule variable explicative dichotomique, cad une covariable x prenant 2 valeurs 0 et 1 (ex : sexe). On rappelle que si $p_1 = P(Y = 1 \mid x = 1)$ et $p_0 = P(Y = 1 \mid x = 0)$, alors l'odds ratio est défini par :

$$OR = \frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}}.$$

Le modèle logistique précise que

$$\log it\left[P(Y = 1 \mid x)\right] = \beta_0 + \beta_1 x.$$

On a donc

$$OR = \frac{\exp^{\log it[P(Y=1|x=1)]}}{\exp^{\log it[P(Y=1|x=0)]}} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}.$$

De sorte que dans un modèle logistique, l'exponentielle du coefficient d'une variable explicative s'interprète comme son odds ratio.

8.6 Interprétation des coefficients

Nous avons vu que dans le cas d'un modèle comportant une seule variable explicative dichotomique l'exponentielle du coefficient de cette variable s'interprétait comme un Odds Ratio. Voyons ce qui se passe lorsque la variable explicative admet plusieurs catégories, i.e. elle est polytomique, ou qu'elle est continue, ou encore que le modèle incorpore d'autres co-variables ainsi que des interactions.

8.6.1 Le cas d'un modèle additif, i.e. sans interactions

Un modèle est additif lorsque les co-variables $x_1, x_2, \ldots, x_p$ entrent dans le modèle de manière additive sans faire intervenir le produit d'une variable avec une autre. Dans le cas de la régression logistique, le modèle est additif sur l'échelle « logit », mais multiplicatif lorsqu'on considère la probabilité. Considérons le modèle :

$$\log it\left[P(Y = 1 \mid x_1, ..., x_p)\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p.$$

Où β_0 est la constante du modèle. Pour illustrer, considérons le modèle suivant :

$$\log it\left[P\left(Y=1 \mid \hat{a}ge, sexe\right)\right] = \beta_0 + \beta_1 \hat{a}ge + \beta_2 sexe.$$

où les variables explicatives sont l'âge et le sexe. Il s'agit d'un modèle additif car il n'y a pas d'interaction (de produit) entre les variables âge et sexe. Autrement dit, dans ce modèle on postule que l'effet de l'âge et du sexe sont indépendants (sur l'échelle logit). Graphiquement, cette hypothèse implique que la droite représentant l'effet de l'âge est simplement translatée sur une distance β_2 lorsqu'on passe d'un genre à l'autre (cf figure9). Dans cet exemple, le vieillissement a le même effet chez les hommes et chez les femmes, mais le niveau absolu du risque est différent (les deux droites ne sont pas superposées). Autrement dit, un accroissement unitaire de l'âge augmentera le logit du même montant quel que soit le genre, et l' Odds Ratio associé à la variable âge sera le même pour les hommes et les femmes.

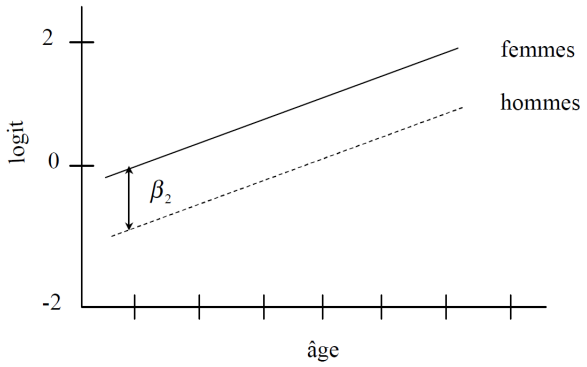


FIG. 9 – Relation entre le logit et l'âge chez les femmes et les hommes dans un modèle additif.

8.6.2 La constante du modèle

La constante du modèle s'interprète comme « l'effet » de la catégorie de référence. Autrement dit, β_0 permet de calculer la probabilité de y lorsque toutes les co-variables $x_1, x_2, \ldots, x_p$ sont nulles. Revenons à notre exemple d'un modèle contenant l'âge et le sexe comme variables explicatives Nous avons arbitrairement choisi de coder les valeurs de la variable sexe = 0 pour les femmes et sexe = 1 pour les hommes, de sorte que β_0 s'interprète comme le logit de la probabilité d'une femme d'âge 0. En effet, la probabilité $P\left(Y=1 \mid \hat{a}ge, sexe\right)$, e.g. d'être malade en fonction de son âge et sexe, s'écrit :

$$P\left(Y=1 \mid \hat{a}ge, sexe\right) = \frac{e^{\beta_0+\beta_1\hat{a}ge+\beta_2 sexe}}{1+e^{\beta_0+\beta_1\hat{a}ge+\beta_2 sexe}}$$

De sorte que pour une femme d'âge 0 on obtient :

$$P\left(Y=1 \mid \hat{a}ge=0, sexe=0\right) = \frac{e^{\beta_0}}{1+e^{\beta_0}},$$

sa probabilité ne dépendant que de β_0. Pour un homme d'âge 0, en revanche, la probabilité dépend aussi de β_1

$$P\left(Y=1 \mid \hat{a}ge=0, sexe=1\right)=\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}.$$

8.6.3 Coefficient d'une variable explicative dichotomique

Lorsque la variable explicative est dichotomique l'exponentielle du coefficient de cette variable s'interprète comme l'Odds Ratio (OR) associé au passage de la catégorie de référence 0 à la catégorie 1. Ainsi, dans notre exemple, lorsque la variable sexe passe de 0 à 1, on a

$$OR=\frac{\exp^{\log it[P(y=1|\hat{a}ge,sexe=1)]}}{\exp^{\log it[P(y=1|\hat{a}ge,sexe=0)]}}=\frac{e^{\beta_0+\beta_1\hat{a}ge+\beta_2}}{e^{\beta_0+\beta_1\hat{a}ge}}=e^{\beta_2}.$$

Il s'agit d'un Odds Ratio ajusté puisque modèle comporte en plus de la variable d'exposition sexe la variable explicative âge. Remarquons que l'Odds Ratio ajusté est en général différent de celui non ajusté, même si son calcul ne fait pas intervenir directement la variable âge, car l'estimation de β_2 dépend de celle de β_1.

8.6.4 Coefficient d'une variable explicative polytomique

Lorsque la variable explicative est polytomique, i.e. elle admet plus de deux catégories, on choisi l'une des catégories comme référence et l'on calcule des Odds Ratios pour les autres catégories par rapport à cette référence. Considérons par exemple la variable éducation comportant 3 niveaux : 1 pour niveau « fin de scolarité », 2 pour « apprentissage » et 3 pour « études supérieures ». Pour représenter une telle variable l'on considérera un modèle avec, en plus de la constante, deux variables « indicatrice » ou « dummy » prenant la valeur 1 si l'individu possède l'attribut et 0 sinon :

- $D_1=1$ si apprentissage, et 0 sinon.
- $D_2=1$ si études supérieures, et 0 sinon.

Le logit s'écrit $\log it\left[P\left(Y=1 \mid \textit{éducation}\right)\right]=\beta_0+\beta_1 D_1+\beta_2 D_2$. L'Odds Ratio associé au passage de la catégorie 1 « fin de scolarité » à la catégorie 2 « apprentissage » est :

$$OR=\frac{\exp^{\log it[P(y=1|\textit{éducation}=2)]}}{\exp^{\log it[P(y=1|\textit{éducation}=1)]}}=\frac{e^{\beta_0+\beta_1}}{e^{\beta_0}}=e^{\beta_1}.$$

Tandis que celui associé associé au passage de la catégorie 1 « fin de scolarité » à la catégorie 3 « études supérieures » est :

$$OR=\frac{\exp^{\log it[P(y=1|\textit{éducation}=3)]}}{\exp^{\log it[P(y=1|\textit{éducation}=1)]}}=\frac{e^{\beta_0+\beta_{21}}}{e^{\beta_0}}=e^{\beta_2}.$$

8.6.5 Coefficient d'une variable explicative continue

Lorsque la variable explicative est continue on calcule un Odds Ratio associé à un accroissement unitaire. Par exemple, considérons la variable âge mesurée en années et supposons que la personne soit d'âge x. Le vieillissement d'une année est associé à un Odds Ratio donné par l'expression :

$$OR=\frac{\exp^{\log it[P(y=1|\hat{a}ge=x+1)]}}{\exp^{\log it[P(y=1|\hat{a}ge=x)]}}=\frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1 x}}=e^{\beta_1}.$$

8.7 Sélection du modèle logistique

Le choix d'un modèle logistique est une opération complexe qui doit s'effectuer pas à pas. Lorsque l'on est dans un cadre de la régression logistique multinomiale, le nombre de coefficients à contrôler est très important, et le choix du "bon" modèle doit se faire d'autant plus soigneusement. Afin de sélectionner un modèle, il faut procéder par ordre et comparer ceux ci deux à deux, en considérant des suites de modèles dit emboîtés. Un modèle M_1 est emboîté dans un autre modèle M_2 lorsque toutes les variables prises en compte dans M_1 se retrouvent aussi dans M_2. On peut aussi dire que le modèle M_1 est "plus petit" que le modèle M_2, ou qu'il est moins complexe (c.a.d. avec moins de variables).

8.8 Critères de qualité du modèle

Plusieurs indicateurs permettent d'obtenir des renseignements sur le poids des variables explicatives et sur la qualité du modèle choisi. De manière globale, la qualité d'un modèle est mesurée par la vraisemblance : plus celle ci est grande, plus le modèle est adapté aux données. Dans le cadre de la régression logistique, il est d'usage d'utiliser à la place la quantité appelée déviance, souvent notée Λ :

$$\Lambda = -2Log\,(L)\,,$$

où L est la vraisemblance (Likelihood en anglais). Cette définition est différente de celle donnée précédemment, mais pour retrouver le terme D^* il suffit de retrancher la déviance du modèle saturé.

Sous l'hypothèse que deux modèles sont emboîtés, la différence entre la déviance d'un modèle M_1 et celle d'un modèle M_2 "plus petit" est donc une valeur positive, qui suit une loi du χ^2 dont le degré de liberté est la différence entre le nombre de paramètres des modèles considérés. Lorsque le gain en terme de déviance (c.a.d. le gain en explication des données obtenu avec le modèle le plus complet par rapport au modèle le "plus petit") est faible, la p-value est élevée. En effet, plus la p-value est élevée, plus la différence des déviances est susceptible de suivre une loi de χ^2, et plus les modèles sont proches. On acceptera donc des modèles qui apportent un gain non négligeable en terme de déviance, relativement au nombre de paramètres utilisés, c'est à dire des modèles dont la p-value est faible.

La dernière mesure de qualité de modèle que nous utiliserons est le critère AIC pour "Akaike Information Criterion". Celui ci est défini de la manière suivante :

$$AIC = -2\log\,(L) + 2k,$$

où L est la Vraisemblance du modèle, et k le nombre de paramètre de celui ci. Ce critère permet de construire un classement de modèles statistiques tenant compte du principe de parcimonie. Les meilleurs ajustements correspondent aux plus faibles valeurs.

8.9 Critères de sélection de variables

La première étape lorsque l'on souhaite sélectionner un modèle pertinent, est de repérer les variables explicatives qui ont une forte influence sur la variable à étudier (ici la classification des sons). Plusieurs possibilités existent pour effectuer cette étape, et nous en avons déjà présenté quelques unes dans l'étude des tables de contingence. Nous allons voir

qu'étudier le gain en Deviance peut se ramener dans certains cas aux tests du maximum de vraisemblance appliqués aux tables de contingences.

Un autre critère communément utilisé est celui du test de Wald. Lorsque l'on estime un modèle logistique additif incluant toutes les variables explicatives, on peut mesurer une quantité égale au carré du rapport du coefficient estimé sur l'erreur commise. Cette quantité suit un $\chi^2(1)$ lorsque le coefficient considéré est égal à 0. L'usage de ce test est donc de sélectionner les variables susceptibles d'être exclues du modèle final, correspondant à des p-values élevées, c'est à dire les variables avec une statistique de Wald faible (et donc une erreur de mesure importante). Ce test est à utiliser en combinaison avec le critère des Deviances car il peut induire des erreurs d'appréciations.

8.10 Tests de significativité des variables explicatives

Il est possible de comparer le poids des différentes variables susceptibles d'entrer dans le modèle. Cela peut être fait par le biais des tables de contingences et des tests d'indépendance de type χ^2, mais il est aussi possible d'utiliser plusieurs régressions logistiques pour chaque variable susceptible d'intervenir dans le modèle. Il suffit ensuite les comparer au modèle ne faisant intervenir que les constantes.

Les valeurs obtenues en faisant ce test correspondent exactement au test d'indépendance basé sur la vraisemblance présenté plus haut. Ces tests, en association avec le test d'indépendance du χ^2, constituent une mesure fiable de l'influence des variables sur la classification. Une autre façon de repérer les variables pouvant être exclues du modèle final est de regarder la statistique de Wald associée aux coefficients estimés pour ces variables pour le modèle complet (incluant toutes les variables candidates)

Le coefficient de Wald est obtenu en calculant le carré du rapport coefficient/erreur standardisée, et suit une loi de khi-deux à 1 degré de liberté lorsque le coefficient associé est nul. De grandes p-value correspondent donc aux coefficients que l'on peut considérer comme nuls et susceptibles d'être exclus du modèle final.

8.11 Recherche d'interactions

Afin de détecter les interactions susceptibles d'avoir de l'influence, il est d'usage de comparer la déviance du modèle avec interaction avec la déviance du modèle additif associé. Dans la mesure où il s'agit de modèles emboîtés, la différence des déviances suit là encore une loi du χ^2. De manière plus générale, la différence des déviances permet de tester l'apport explicatif d'une suite de modèle emboîtés. Un modèle ne modifiant que très peu la déviance n'apportera que peu d'information mais aura pour conséquence d'ajouter des variables, ce qui peut nuire à la qualité d'estimation des paramètres. On choisira donc un modèle le plus parcimonieux possible, faisant un compromis entre la part d'information expliquée et la complexité.

Références

[1] Alan Agresti. *Categorical data analysis. 2nd ed.* Wiley Series in Probability and Mathematical Statistics., Chichester, 2002.

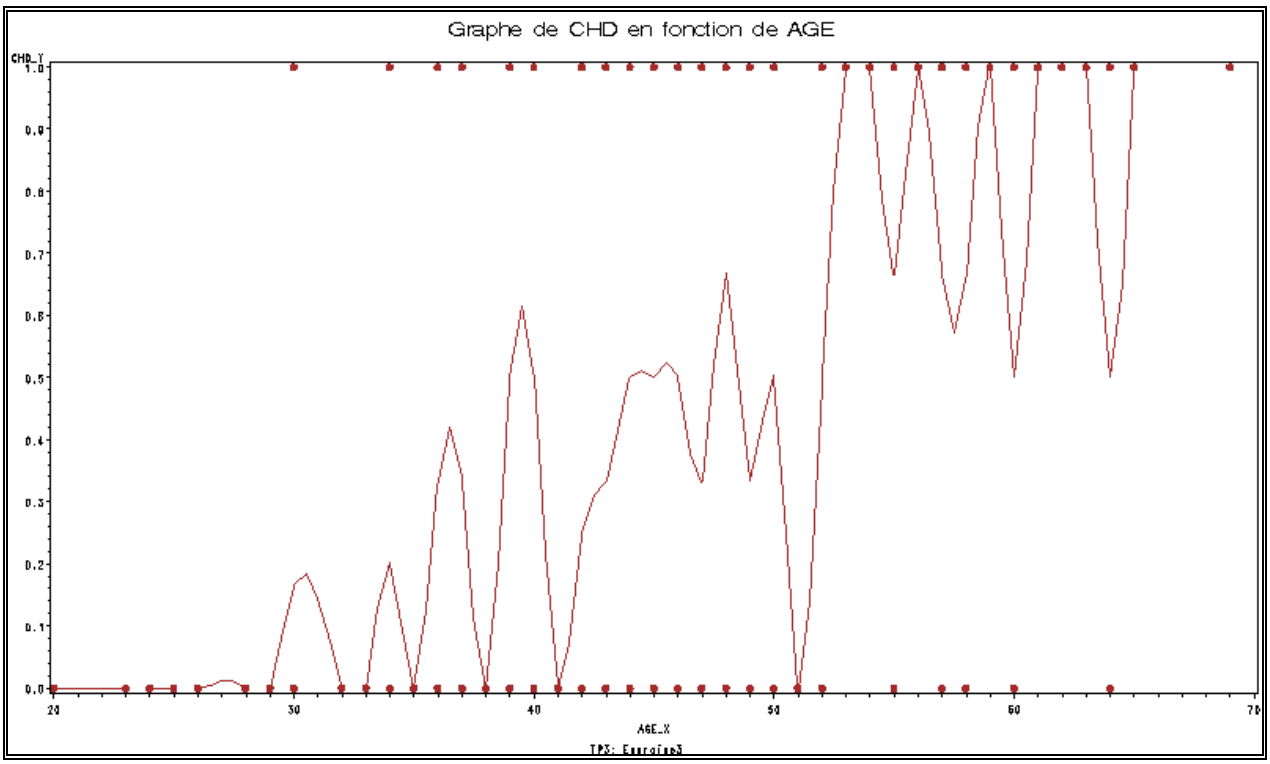
[2] André Carlier, Gilles Celeux, Alice Gueguen, Abdallah Mkhadri, Jean-Pierre Nakache, Jean-François Petiot, and Jean-Christophe Turlot. *Analyse discriminante sur variables qualitatives. Préface de Jean-Jacques Daudin.* Polytechnica, Paris, 1994.

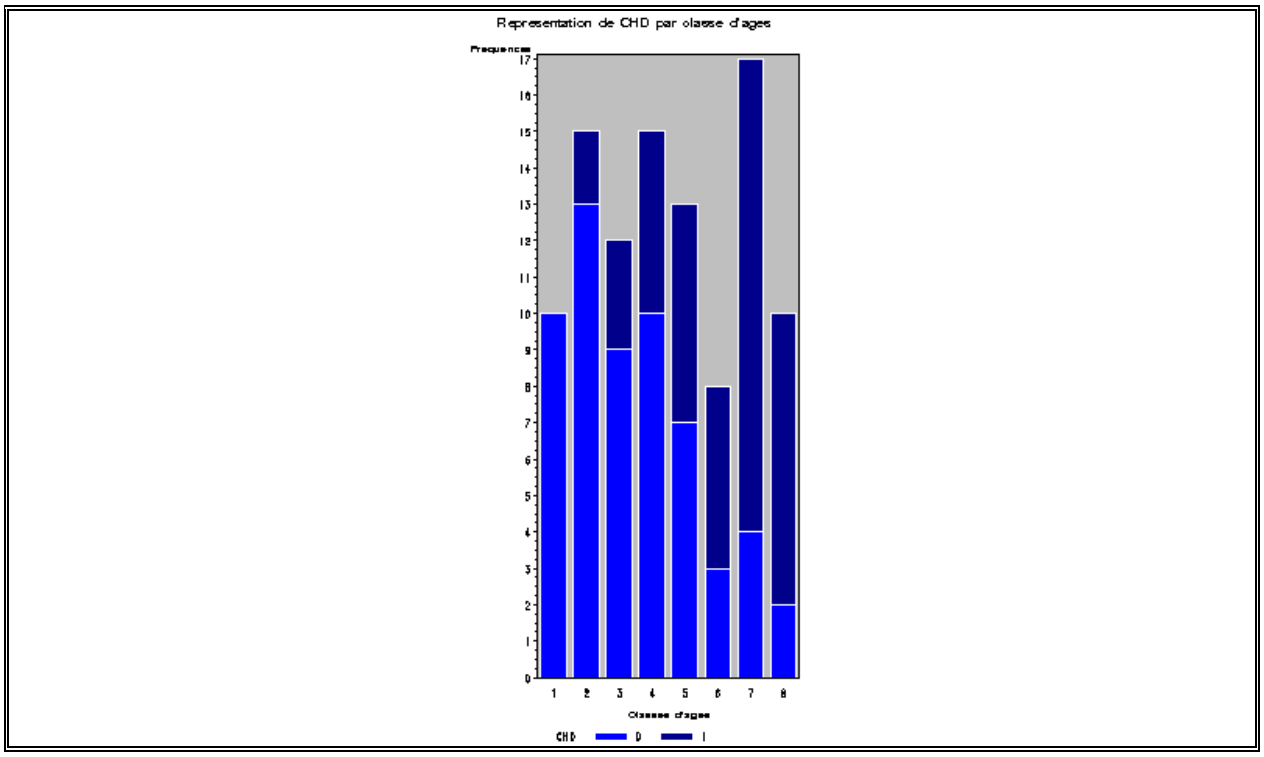
[3] D.D. Hosmer and S. Lemeshow. *Applied logistic regression.* John Wiley & son, 1989.

Code et sorties de l'exercice 3 du TP3 (exemple du cours)

Ce jeu de données est analysé dans Hosmer D. W. & Lemeshow S. (2000) : Applied logistic regression, 2nd edition, John Wiley & Sons, New York

```
/************************/
/*    exercice3 séance3 */
/************************/

options pagesize=38 linesize=78 nodate; title; footnote 'TP3: Exercice3';

/* Sample data set: heart */

/* Tri de la table heart suivant les valeurs 0 ou 1 de CHD */
proc sort data=malib.heart out=HeartByCHD;
by CHD;
run;

/* Affichage de la table heart: 1 page par modalité de CHD */
proc print data=HeartByCHD noobs;
var AGE AGRP;
by CHD;
run;

/* On supprime HeartByCHD dans la Work */
proc delete data=HeartByCHD; run;

/* Graphique de CHD en fonction de AGE */
proc gplot data=malib.heart;
title 'Graphe de CHD en fonction de AGE';
symbol color=brown value=dot;
plot CHD*AGE;
label CHD='CHD_Y'
            AGE='AGE_X';
run;
quit;

/* Représentation des modalités de CHD par classe d'âges */
goptions reset=all border cback=white colors=(blue darkblue) ctext=black;

title c=black f=swissb h=1.2 j=center 'Representation de CHD par classe
d''ages';

axis1       label=(h=1 c=black f=swissb 'Frequences');
axis2 label=(h=1 c=black f=swissb 'Classes d''ages');

proc gchart data=malib.heart;
vbar AGRP / subgroup=CHD raxis=axis1 maxis=axis2 caxis=black coutline=white
cframe=ligr;
run;
quit;
```


Graphe de CHD en fonction de AGE
CHD_Y
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
20
30
40
50
60
70
AGE_X


Representation de CHD par classe d'ages
Frequences
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
1
2
3
4
5
6
7
8
Classes d'ages
CHD
0
1

```
title ' ';

/* Création de formats personnalisés. Utile pour discrétiser une variable
continue ou recoder une variable discrète */
proc format;
value fAGRP 1='20-29'
                        2='30-34'
                        3='35-39'
                        4='40-44'
                        5='45-49'
                        6='50-54'
                        7='55-59'
                        8='60-69';
value fCHD  0='Absent'  1='Present';
run;

title;
proc tabulate data=malib.heart format=8.1;
class AGRP CHD;
format AGRP fAGRP. CHD fCHD.;
label AGRP='AgeGroup';
table AGRP ALL,N*(ALL CHD) / box='Frequency Table of AgeGroup by CHD' RTS=12;
keylabel ALL='Total' N='Frequency';
run;
```

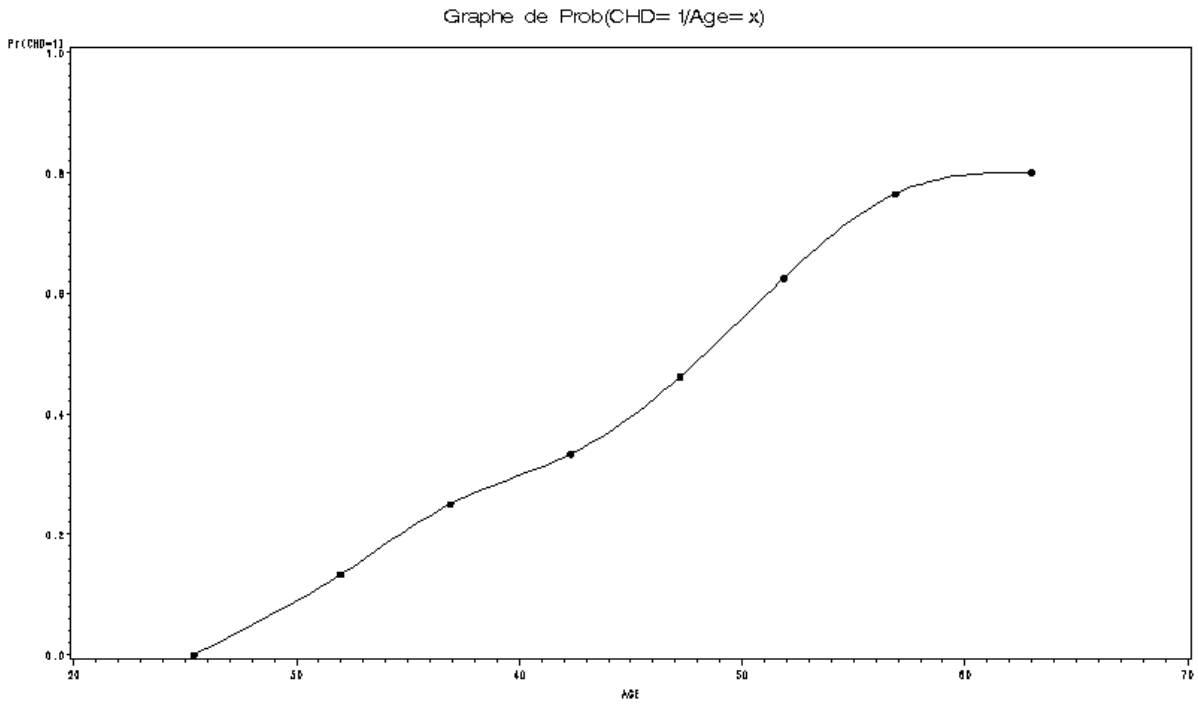
Frequency Table of AgeGroup by CHD	Frequency		
	Total	CHD	
		Absent	Present
AgeGroup	10	10	.
20-29			
30-34	15	13	2
35-39	12	9	3
40-44	15	10	5
45-49	13	7	6
50-54	8	3	5
55-59	17	4	13
60-69	10	2	8
Total	100	58	42

```
proc means data=malib.heart mean maxdec=1;
class AGRP;
var CHD AGE;
output out=StatUniv mean=CHDmean AGEmean;
run;
```

La procédure MEANS

AGRP	N Obs	Variable	Moyenne
1	10	CHD	0.0
		AGE	25.4
2	15	CHD	0.1
		AGE	32.0
3	12	CHD	0.3
		AGE	36.9
4	15	CHD	0.3
		AGE	42.3
5	13	CHD	0.5
		AGE	47.2
6	8	CHD	0.6
		AGE	51.9
7	17	CHD	0.8
		AGE	56.9
8	10	CHD	0.8
		AGE	63.0

```
proc print data=StatUniv; run;

/* L'instruction attrib sert à définir les attributs d'une variable: étiquette
+ format + informat */
/* format=n.d affiche un nombre sur n colonnes avec d décimales */
data StatUniv;
set Statuniv;
attrib AGRP label='Classe d''ages' format=2.;
attrib CHDmean label='CHD moyen' format=3.1;
attrib AGEmean label='Age moyen' format=4.1;
keep AGRP AGEmean CHDmean;
where AGRP ne .;
run;

proc print data=StatUniv noobs label; run;

/* Le graphe suivant fait apparaître une courbe en forme de S (ou sigmoïde)
                  */
/* ce qui est caractéristique d'une fonction exp(ax)/(1+exp(ax)) du modèle
logistique  */
goptions reset=all border cback=white ctext=black;

proc gplot data=StatUniv;
title 'Graphe de Prob(CHD=1/Age=x)';
symbol interpol=splines /*join*/ value=dot;
label CHDmean='Pr(CHD=1)'
            AGEmean='AGE';
plot CHDmean*AGEmean / vaxis=0 to 1 by 0.2;
run;
quit;

proc delete data=StatUniv; run;
```



```
title 'Construction du modèle logistique';

proc logistic data=malib.heart;
model CHD=AGE;
output out=sorties predprobs=individual;
run;
quit;
```

Construction du modèle logistique

The LOGISTIC Procedure

Informations sur le modèle	
Data Set	MALIB.HEART
Response Variable	CHD
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	100
Number of Observations Used	100

Profil de réponse		
Valeur ordonnée	**CHD**	**Fréquence totale**
1	0	58
2	1	42

Probability modeled is CHD=0.

État de convergence du modèle
Convergence criterion (GCONV=1E-8) satisfied.

Statistiques d'ajustement du modèle		
Critère	**Coordonnée à l'origine uniquement**	**Coordonnée à l'origine et covariables**
AIC	138.058	105.951
SC	140.664	111.162
-2 Log L	136.058	101.951

Test de l'hypothèse nulle globale : BETA=0			
Test	**Khi 2**	**DF**	**Pr > Khi 2**
Likelihood Ratio	34.1070	1	<.0001
Score	30.1452	1	<.0001
Wald	23.2690	1	<.0001

Analyse des estimations de la vraisemblance maximum					
Paramètre	**DF**	**Estimation**	**Erreur std**	**Khi 2 de Wald**	**Pr > Khi 2**
Intercept	1	5.9756	1.2198	23.9987	<.0001
AGE	1	-0.1241	0.0257	23.2690	<.0001

Estimations des rapports de cotes			
Effet	**Point Estimate**	**95% Limites de confiance de Wald**	
AGE	0.883	0.840	0.929

Association des probabilités prédites et des réponses observées			
Percent Concordant	81.1	**Somers' D**	0.641

Association des probabilités prédites et des réponses observées			
Percent Discordant	17.0	**Gamma**	0.654
Percent Tied	2.0	**Tau-a**	0.316
Pairs	2436	**c**	0.821

```
/* Evaluation de la qualité prédictive du modèle en croisant les from-into */
proc freq data=sorties;
title 'Evaluation de la qualité du modèle';
table _from_*_into_ / nopercent norow nocol chisq;
run;
```

Evaluation de la qualité du modèle

La procédure FREQ

FREQUENCE	Table de _FROM_ par _INTO_		
FROM(Formatted Value of the Observed Response)	**_INTO_(Formatted Value of the Predicted Response)**		**Total**
	0	**1**	
0	47	11	58
1	15	27	42
Total	62	38	100

Statistiques pour table de _FROM_ par _INTO_

Statistique	DF	Valeur	Proba.
Khi-2	1	21.2366	<.0001
Test du rapport de vraisemblance	1	21.7215	<.0001
Continuité Adj. Khi-2	1	19.3566	<.0001
Khi-2 de Mantel-Haenszel	1	21.0243	<.0001

Statistique	DF	Valeur	Proba.
Coefficient Phi		0.4608	
Coefficient de contingence		0.4185	
V de Cramer		0.4608	

Test exact de Fisher	
Cellule (1,1) Fréquence (F)	47
Pr <= F unilatérale à gauche	1.0000
Pr >= F unilatérale à droite	4.501E-06
Table de probabilité (P)	3.962E-06
Pr <= P bilatéral	7.391E-06

Taille de l'échantillon = 100