Données Catégorielles
Applications

B. G.

14 Février 2003

Sommaire

1 Préliminaires, objectifs

La régression logistique est souvent présentée comme :

- un modèle analogue à la régression linéaire pour des données binaires ou polytomiques (Guyon, Amemya).

  un cas particulier de modèle loglinéaire (Christensen).

  une analyse des associations dans un tableau de contingence (Agresti, Andersen).

Ces différents points de vue apportent des éclairages intéressants sur le traitement statistique mais ils sont souvent accompagnés d'ambiguïtés, en particulier sur les objectifs de l'utilisateur ; ambiguïtés sur lesquelles les auteurs insistent, mais que les lecteurs ont tendance à négliger.

- ♦ S'agit-il d'expliquer des variables endogènes à partir de variables explicatives, ou de comparer des distributions (effectifs) de sous-groupes de la population?
- ♦ S'agit-il de rechercher les variables exogènes (ou de classement en sous-groupes) qui résument ou expliquent le mieux la répartition de la variable endogène?
- ♦ S'agit-il d'associer les lignes et les colonnes (ou les différentes dimensions) d'un tableau de contingence et d'évaluer leur dépendance?
- ♦ Quelle est la population étudiée, qu'appelle-t-on réponse, facteur ("covariate")?
- ♦ Qu'entend-on par association conditionnelle, homogène, par homogénéïté marginale, indépendance conditionnelle?
- ♦ S'agit-il d'ajustement, de prévision?
- ♦ Quelle sont les lois de probabilité dans l'échantillon observé?
- ♦ Estime-t-on une probabilité (un risque), un effectif, l'espérance d'un effectif, une probabilité conditionnelle?

Du point de vue technique (statistique et numérique), les modèles sont assez faciles à estimer puisque les lois de probabilités en cause appartiennent à la famille exponentielle; les vraisemblances sont généralement simples, les tests ont des propriétés asymptotiques connues, les méthodes d'estimation sont unifiées (modèles linéaires généralisés).

Par contre l'application de ces modèles et la *pertinence des conclusions* qu'on en tire ("interprétation"), n'est pas toujours claire, bien que les praticiens aient souvent acquis, par l'expérience, des règles de conduite qui s'appliquent bien à leur problème habituel et leur évite des méprises.

Notre point de vue, dans les exemples qui seront traités, est celui du statisticien appliqué qui fait confiance aux statisticiens probabilistes pour tout ce qui concerne les propriétés

plus ou moins complexes des statistiques qui sont calculées. Il ne conteste pas les résultats fournis par l'ordinateur, mais il sait choisir dans la multitude de statistiques calculées celles qui correspondent aux objectifs qu'il s'est fixés.

Il veut savoir si ses données, et les informations qu'il détient sur elles, sont bien traitables, et avec quel modèle ; ensuite, il veut savoir ce que lui apporte son modèle estimé et surtout ce qu'il ne lui apporte pas. Un médecin dirait que le traitement doit-être adapté au patient (penser aux contre-indications), et efficace pour la maladie qu'on désire soigner ; encore qu'en statistique il est plus exact de parler de symtômes, de diagnostic que de traitement.

Nous partirons de l'examen des tableaux de contingence comme le font les grands classiques (Agresti, Andersen, Christensen), ce qui permettra de bien décrire les conditions d'expérimentation (construction de l'échantillon), les différentes lois de probabilités en jeu et quelques tests simples.

Références :

AGRESTI, A. (1990). *Categorical Data Analysis.* Wiley.

AGRESTI, A. (1996). *An Introduction to Categorical Data Analysis.* Wiley.

AMEMEYA, T. (1985),*Advanced Econometrics*" Blackwell.

ANDERSEN, E. B. (1980). *Discrete Statistical Models with Social Science Applications.* North Holland.

CHRISTENSEN, R. (1990). *Log-linear models.* Springer.

GUYON, X, (2001) *Statistique et économétrie, Du modèle linéaire... aux modèles non-linéaires*, Ellipses.

McCULLAG (1989) *Generalized Linear Models*, Chapman and Hall.

2 Préliminaires, exemples

Pour fixer les idées sur le type de questions que nous examinerons, nous proposons quelques exemples de données catégorielles comme on en rencontre souvent. Ces exemples seront repris un par un dans les sections suivantes, pour définir la réalité qui est sous-jacente à ces tableaux de données, la lecture "intelligente" qu'on peut en faire, les informations particulières qu'on espère en tirer.

Commençons par citer la présentation historique de Fisher, à propos du goût du thé, bien qu'il soit moins simple qu'il ne paraît (FISHER, R. A. ,1935, *The Design of Experiment*. Oliver and Boyd).

Il s'agit de savoir s'il est possible de distinguer, en buvant du thé, si le lait a été versé dans la tasse avant ou après le thé. L'expérience consiste à présenter 8 tasses de thé. Dans quatre tasses le lait a été versé avant le thé, dans les quatre autres il a été versé après. On demande à quelqu'un de distinguer les deux types de tasses de thé.

Le tableau de contingence des réponses se présente sous la forme:

	a effectivement été versé en premier		
semble avoir été versé en premier▼	le lait	le thé	*total*
le lait	3	1	4
le thé	1	3	4
total	4	4	8

(Les exemples suivants ont été aimablement fournis par des collègues de l'université Paris I)

Le premier exemple est le résultat d'une enquête sur l'existence du père Noël.

	âges des enfants interrogés				
croit au Père Noël ▼	3 ans	4 ans	5 ans	6 ans	*total*
oui	30	13	15	5	63
non	5	10	12	28	55
total	35	23	27	33	118

Le deuxième exemple porte sur la qualité de l'accueil des étudiants par l'administration, dans unité de formation.

accueil ▼	1ère année	2ème année	total
mauvais	74	31	105
moyen	122	77	199
bon	18	43	61
très bon	4	1	5
total	218	152	370

Le troisième exemple décrit les résultats des étudiants inscrits dans un DESS et provenant de filières différentes.

	Paris		Province		
filière à l'entrée▶	Economie	Mass-Mst	Economie	Mass-Mst	total
note < 12	3	2	5	1	11
note ≥ 12	6	4	8	4	22
total	9	6	13	5	33

Le quatrième exemple examine les conditions de réussite au concours interne de Premier Surveillant pour deux générations de surveillants, hommes et femmes (Administration Pénitentiaire). Par exemple, les hommes entrés comme élèves-surveillants dans les années 80 à 84 et qui ont été reçus au concours de Premier Surveillant sont répartis selon l'âge et l'ancienneté (dans l'année où ils sont reçus) :

Hommes (80-84)	**ancienneté** (en années)			
tranches d'âge ▼	1-8	9-10	11 &+	*total*
moins de 30 ans	83	29	0	112
31-35 ans	80	71	43	194
36-40 ans	45	39	52	136
41 ans et plus	39	24	21	84
total	247	163	116	526

Le cinquième exemple montre la gravité des accidents impliquant deux véhicules, selon l'âge du conducteur sensé être responsable, et son sexe (département du Nord).

Hommes	**caractère de gravité de l'accident**				
	aucun tué	1 tué	2 tués	*plus de 2 tués*	total
âge < 18 ans	3712	74	4	0	3790
18-24 ans	5990	286	27	5	6308
25-64 ans	7642	517	34	10	8203
âge ≥ 65 ans	155	29	0	0	184
total	17499	906	65	15	18485

Femmes	**caractère de gravité de l'accident**				
	aucun tué	1 tué	2 tués	plus de 2 tués	total
âge < 18 ans	1489	28	1	0	1518
18-24 ans	1687	53	2	0	1742
25-64 ans	2663	95	1	4	2763
âge ≥ 65 ans	89	8	0	0	97
total	5928	184	4	4	6120

Le sixième exemple rend compte d'un essai clinique portant sur 238 patients à risques (ayant déjà eu des polypes adénomateux retirés lors d'une première coloscopie). Au bout

d'un an, on compare la récidive selon qu'ils ont pris de l'aspirine ou un placebo. On évalue la récidive des adénomes par la somme des diamètres des adénomes retirés lors de la deuxième coloscopie ("burden"). On compare alors les groupes de patients par leur âge, leur sexe, leurs antécédents personnels (polypes) et le nombre d'adénomes à la coloscopie initiale.

	Aspirine			**Placebo**		
burden ▶	$< 6\,mm$	$\geq 6\,mm$	total	$< 6\,mm$	$\geq 6\,mm$	*total*
âge ▼						
≤ 56	50	3	53	34	7	41
56-64	39	6	45	32	7	39
≥ 65	23	4	27	22	10	32
sexe ▼						
femmes	36	1	37	32	4	36
hommes	76	12	88	56	20	76
antécédents ▼						
non	88	9	97	69	12	81
oui	24	4	28	19	12	31
initial ▼						
nombre ≤ 2	87	6	93	72	9	81
nombre > 2	25	7	32	16	15	31
Ensemble	112	13	125	88	24	112

Le septième exemple compare les résultats de deux recensements (1990 et 1999), sur la composition des habitants des 80 quartiers de Paris. On retient pour chaque quartier, comme critères de classement, le genre, la nationalité des habitants, et une variable "illustrative" issue d'une classification exogène définissant l'activité d'un quartier par le type de catégorie socio-professionnelle qui y est majoritaire. Exemples de quartiers :

quartier	**Saint-Ambroise** (professions intermédiaires)					
année	1990			1999		
nationalité	française	étrangère	total	française	étrangère	*total*
hommes	12770	3213	15983	12891	2515	15406
femmes	14975	2509	17484	14473	2289	16762
total	27745	5722	33467	27364	4804	32168

quartier	**Croulebarbe** (cadres actifs)					
année	1990			1999		
nationalité	française	étrangère	total	française	étrangère	*total*
hommes	7762	999	8761	7862	877	8739
femmes	10252	954	11206	9873	914	10787
total	18014	1953	19967	17735	1791	19526

quartier	**Chaillot** (cadres retraités)					
année	1990			1999		
nationalité	française	étrangère	*total*	française	étrangère	*total*
hommes	7687	2176	9863	7968	1910	9878
femmes	9437	2457	11894	9060	2275	11335
total	17124	4633	21757	17028	4185	21213

quartier	**Goutte-d'Or** (ouvriers)					
année	1990			1999		
nationalité	française	étrangère	total	française	étrangère	*total*
hommes	9221	5392	14613	10251	4427	14678
femmes	9821	3792	13613	10121	3725	13846
total	19042	9184	28226	20372	8152	28524

Le huitième exemple est un extrait du recensement 1990 pour deux arrondissements de Paris (5ème et 13ème) dont on compare la répartition des *actifs* par catégories socio-professionnelles, par leur âge et leur niveau d'études (population réduite à la tranche 20-59 ans).

Distributions marginales :

	classes d'âge				
résidence ▼	20-29	30-39	40-49	50-59	*total*
5ème Ardt	1672	1953	1827	1224	6676
13ème Ardt	4930	5919	5249	3361	19390
total	6602	7872	7007	4585	26066

	niveau d'études				
résidence ▼	<=brevet	bac	1e cycle	2e cycle	*total*
5ème Ardt	2107	898	814	2857	6676
13ème Ardt	9234	3047	2307	4802	19390
total	11341	3945	3121	7659	26066

	catégories socio-professionnelles					
résidence ▼	ouv	emp	proi	pic	cpis	*total*
5ème Ardt	473	1313	1352	488	3050	6676
13ème Ardt	2610	5467	4580	939	5794	19390
total	3083	6780	5932	1427	8844	26066

	distance du lieu de résidence au lieu de travail					
résidence ▼	0-4 km	5-9 km	10-14	15-19	20-24	*total*
5ème Ardt	4866	1079	345	148	130	6676
13ème Ardt	11733	5138	1427	432	420	19390
total	16599	6217	1772	980	550	26066

Le neuvième exemple montre comment on peut construire un indicateur synthétique (score), qui permette de classer des individus (clients) en partant de ce qu'on sait sur eux, pour prévoir leur comportement. Il s'agit de détecter si des clients qui se présentent sont susceptibles de ne pas pouvoir rembourser un prêt. L'exemple est inspiré d'une étude faite par un organisme de crédit. Le score sera construit à partir d'un échantillon de bons payeurs et d'un échantillon de mauvais payeurs issus du fichier des clients connus.

Les bons payeurs		
situation familiale	sans enfants	enfants
personne seule	1280	280
couple	1080	1360

Les mauvais payeurs		
situation familiale	sans enfants	enfants
personne seule	1120	320
couple	680	1560

Ces exemples vont être explorés dans un premier temps pour en dégager les traits essentiels et les questions qu'on est amené à se poser. Ils seront repris ensuite pour être analysés en détail avec des modèles statistiques.

3 Premiers pas

Nous allons reprendre les exemples précédents pour bien définir les conditions d'expérimentation. Ces conditions d'expérimentation ne sont pas toujours explicites, mais il est indispensable de les connaître pour bien comprendre l'information que peut apporter le tableau des données et pour choisir le modèle statistique qui doit être utilisé.

Exemple 0 : le goût du thé

L'objectif de cette expérience est d'évaluer la pertinence de l'affirmation du goûteur, en comparant sa réponse à celle qu'on aurait obtenue en répondant au hasard.

On remarque d'abord que ce tableau ne présente pas beaucoup de "degrés de liberté". L'expérience porte sur 8 tasses classées en deux groupes de 4. L'épreuve consiste à classer les tasses. La répartition des huit tasses (4 nombres) est aléatoire mais soumise à des contraintes : les totaux en ligne et colonne sont fixés (égaux à 4).

	a effectivement été versé en premier		
semble avoir été versé en premier▼	le lait	le thé	*total*
le lait	3	1	4
le thé	1	3	4
total	4	4	8

L'épeuve est donc résumée par une seule variable aléatoire : par exemple, le nombre de tasses classées par le goûteur sous la rubrique *thé effectivement versé en premier,* et qui sont correctement classées. En effet les contraintes marginales permettent de déduire le tableau complet à partir du nombre k (=3) inscrit dans la case (1,1), comme d'ailleurs on aurait pu le faire à partir de chacune des trois autres.

Ce nombre est la réalisation d'une variable aléatoire qui suit une loi *hypergéométrique* dans **ce cas particulier** où les tasses sont présentées au hasard, où l'avis du goûteur n'est pas influencé par les tasses précédemment testées, où toutes les situations sont équiprobables ; en somme, où on n'y voit goutte.

Par exemple, la probabilité de se trouver avec les nombres qui figurent dans le tableau est, en tenant compte des hypothèses précédentes :

$$\frac{C_4^3.C_4^1}{C_8^4} = 0.23$$

En examinant les probabilités des 5 réponses possibles :

$$\frac{C_4^k.C_4^{4-k}}{C_8^4} \; pour \; k = 0, 1, ...4$$

on trouve :

k	0	1	2	3	4
probabilité	0.014	0.229	0.514	0.229	0.014

Peut-on estimer que la réponse n'est pas donnée au hasard ?

Cet exemple, très simple, met en évidence une difficulté qu'on rencontrera souvent dans les expérimentations : il faut définir précisément ce qui est aléatoire, les dépendances, les variables contrôlées, le *protocole d'expérience.*

Dans le cas particulier étudié, la réponse à l'épreuve globale est la somme des résultats de huit petites épreuves du type (oui/non) mais qui doivent finalement aboutir à quatre de chaque type. C'est donc que ces huit petites épreuves ne sont pas indépendantes.

L'expérience aurait pu être menée de façon différente, par exemple en ne donnant pas à l'avance le nombre de tasses de chaque type ; la seule contrainte porterait alors sur le nombre total de tasses. La réponse du goûteur dans la case (1,1) pourrait alors aller de 0 à 8, et suivre une loi binomiale.

Cet exemple initial, traité dans de nombreux ouvrages n'a été présenté que pour montrer, sur une expérience simple et des données élémentaires, l'importance des conditions dans lesquelles s'est effectuée cette l'expérience.

Dans les exemples suivants nous partirons d'expériences mieux définies.

3.1 Exemple 1 : le Père Noël

Ce nouvel exemple fera penser à une enquête sociologique sur l'insécurité, un sondage électoral, une enquête sur la qualité d'une lessive ou d'un ordinateur, sur l'efficacité d'une médecine exotique, etc.

La population étudiée est un échantillon de 118 enfants d'une école, dans la tranche d'âge 3-6 ans.

L'objectif est d'évaluer les réponses des enfants (oui/non). La réponse de chaque enfant est aléatoire. Les enfants sont classés par leur âge : *l'âge structure la population.*

L'âge n'est pas aléatoire dans l'expérience, c'est une variable exogène, de classement, dite "extérieure" au modèle bien qu'elle intervienne pour distinguer des sous-populations d'enfants qui n'ont pas nécessairement le même comportement, c'est-à-dire pas la même probabilité de réponse.

On suppose que les réponses sont indépendantes au sens où la réponse d'un enfant n'a pas d'influence sur la réponse d'un autre. Il s'agit d'**indépendance en probabilité**.

Par contre, il est possible que les réponses dépendent en moyenne de l'âge de l'enfant. Il s'agit d'une *liaison* entre l'âge et le paramètre *espérance mathématique* de la loi de probabilité de réponse de chaque enfant.

	âges des enfants interrogés				
croit au Père Noël ▼	3 ans	4 ans	5 ans	6 ans	*total*
oui	30	13	15	5	63
non	5	10	12	28	55
total	35	23	27	33	118

Dans ces conditions, pour chaque âge (colonne du tableau), on postulera que la loi du nombre de réponses *oui* est binomiale de paramètres (p_i, n_i) où n_i est l'effectif correspondant à l'âge i, qui se trouve dans la dernière ligne du tableau, et p_i est la probabilité de répondre *oui*, commune aux enfants d'âge i. Il y a ainsi 4 paramètres p_i à estimer.

On peut désirer les évaluer ; tester s'ils sont tous identiques ce qui voudrait dire que l'âge n'intervient pas dans les réponses ; tester l'égalité de certains d'entre eux, etc.

Si, au lieu de travailler sur chaque colonne, on considére le tableau dans son ensemble, la répartition des 118 enfants interrogés est décrite par huit nombres aléatoires dont la somme est égale à 118 (une contrainte). Il y a donc 7 paramètres. La loi conjointe de ces huit nombres est une loi multinomiale.

C'est une description instantanée. On peut alors se demander si les reponses dépendent ou non de l'âge, c'est-à-dire si les répartitions dans les lignes et dans les colonnes sont indépendantes. Dans ce cas, il y a des contraintes supplémentaires sur les paramètres de la loi multinomiale : la probabilité dans chaque case est le produit des probabilités marginales (lignes, colonnes), et il ne reste plus que 3 paramètres à estimer :

$$(nb\_lignes - 1) \times (nb\_colonnes - 1) = 3$$

Une autre forme d'expérimentation aurait consisté à construire l'échantillon à partir des croyances : par exemple, en retenant les 63 premiers enfants qui croient au Père Noël et les 55 premiers qui n'y croient pas. On peut trouver plus naturel de fixer les effectifs à (60, 60) plutôt que (63, 55) qui semble artificiel. Dans les enquêtes, il arrive souvent que certaines réponses ne soient pas utilisables ; les nombres de réponses retenues ne sont donc pas égaux même si au départ il y avait le même nombre de réponses souhaitées. Il est donc important que les résultats ne soient pas étroitement liés aux effectifs des sous-populations.

Le tableau de résultats aurait la même allure mais il y aurait deux contraintes, au lieu d'une, correspondant à la fixation des totaux par ligne (63, 55), au lieu du total global (118). La réponse aléatoire est l'âge et il s'agit alors de deux lois multinomiales indépendantes avec $2 \times 3 = 6$ paramètres. La répartition du tableau *dans son ensemble* n'obéit plus à une loi multinomiale. Mais on verra que la paramétrisation en termes de rapport des chances (*odds ratios*) dans la régression logistique rend très proches les deux points de vue.

Ces deux visions du tableau correspondent à des structures aléatoires différentes, donc une forme d'expérimentation différente, et on les distingue souvent dans la présentation des nombres en insistant sur l'importance respective des lignes ou des colonnes. Dans le premier cas on s'intéresse de préférence aux distributions des colonnes et dans le second cas à celle des lignes des lignes comme ci dessous.

Répartitions (pourcentages) en colonne

	âges des enfants interrogés				
croit ▼	3 ans	4 ans	5 ans	6 ans	*total*
oui	30 **(85.7%)**	13 **(56.5%%)**	15 **(55.6%)**	5 **(15.2%)**	63 (53.4%)
non	5 **(14.3%)**	10 **(43.5%)**	12 **(44.4%)**	28 **(84.8%)**	55 (46.6%)
total	35 (100%)	23 (100%)	27 (100%)	33 (100%)	118 (100%)

Répartitions (pourcentages) en ligne

	âges des enfants interrogés				
croit ▼	3 ans	4 ans	5 ans	6 ans	*total*
oui	30 **(47.6%)**	13 **(20.6%)**	15 **(23.8%)**	5 **(7.9%)**	63 (100%)
non	5 (**9.1%)**	10 **(18.2%)**	12 **(21.8%)**	28 **(50.9%)**	55 (100%)
total	35 (29.7%)	23 (19.5%)	27 (22.9%)	33 (28.0%)	118 (100%)

On pourrait aussi, troisième point de vue, considérer que le nombre total d'enfants est aléatoire (par exemple à l'entrée d'un magasin de jouets avec un Père Noël en vitrine). Les huit réponses du tableau sont aléatoire, sans contrainte sur les effectifs en ligne, en colonne, ou total. On admettrait alors que le nombre qui figure dans chaque case est expliqué (du moins son espérance mathématique) par l'âge et la croyance au Père Noël. On serait amené à postuler que ce nombre suit une loi de Poisson dont le paramètre dépend de l'âge et de la croyance au Père Noël.

Autre point de vue : on voudrait tester comment les croyances se *modifient* avec l'âge, l'individu statistique serait alors l'enfant qu'on suivrait au cours de sa croissance. Evidemment, le sondage précédent ne permet pas d'expliquer *le devenir* de la croyance puisque les résultats sont conditionnés à *un instant donné* et à une population particulière.

On distinguera les *coupes instantanées* (ou études *rétrospectives*) des *suivis* (de cohortes, ou études *prospectives*).

On distinguera aussi les cas où la population sondée est représentative ou non de la population totale.

En somme, cet exemple montre, à partir d'un tableau croisé élémentaire, qu'on ne peut pas lancer un traitement statistique sans savoir ce qu'on cherche à mesurer et comment le tableau a été obtenu.

3.2 Exemple 2 : accueil des étudiants

Cet exemple-ci peut être facilement transposé pour une enquête de qualité sur un produit, un test sur la réaction des patients à un traitement, sur le goût des clients pour un nouveau produit financier qui leur est proposé, etc.

Il s'agit d'un *état des lieux,* de comportements à un instant donné ; la population statistique est décomposée en deux sous-populations caractérisées par l'année d'étude. Le comportement aléatoire des individus porte sur la qualité de l'accueil.

Il n'y a pas à proprement parler de variable *explicative* du choix, même s'il est possible que la connaissance du contexte suggère des explications sur les différences de comportement d'une année à l'autre : ces explications pourraient être reliées au changement d'un individu d'une année à l'autre, ce qui d'ailleurs ne correspond pas à l'expérimentation choisie (coupe instantanée) et donc n'est pas validable par le modèle statistique. Le modèle peut, au mieux, comparer les attitudes de deux sous-populations sans pouvoir distinguer ce qui vient de l'administration (qui ne reçoit pas de la même façon les étudiants d'année

différente), de l'évolution de comportement des étudiants (qui savent mieux où s'adresser, qui changent d'attitude vis à vis de l'administration, qui sont plus ou moins exigeants, qui appartiennent à des générations différentes).

C'est une étude *rétrospective*, comme dans l'exemple précédent.

Une présentation des répartitions en colonnes sera éclairante, soit pour voir les écarts de chaque sous-population, par année, à celle de l'ensemble (répartition marginale), soit pour voir les écarts entre les deux sous-populations.

accueil ▼	1ère année	2ème année	total
mauvais	74 **(34,0%)**	31 **(20.4%)**	105 **(28.4%)**
moyen	122 **(56.0%)**	77 **(50.7%)**	199 **(53.8%)**
bon	18 **(8.3%)**	43 **(28.3%)**	61 **(16.5%)**
très bon	4 **(1.8%)**	1 **(0.7%)**	5 **(1.4%)**
total	218 (100%)	152 (100%)	370 (100%)

Il est assez naturel de comparer des répartitions dans les deux sous-populations.

La réponse d'un individu résulte alors d'un choix dans un ensemble de 4 modalités ordonnées (de la variable accueil). Pour mettre en valeur les différences entre les deux répartitions, plusieurs indicateurs sont intéressants.

Le plus courant est un écart global entre la distribution telle qu'elle est (observée) et ce qu'elle aurait dû être (estimée) si les distributions étaient les mêmes, donc en tenant compte uniquement des données marginales (dernière ligne et dernière colonne).

Deux statistiques sont généralement calculées pour tester cette indépendance entre l'accueil et l'année d'étude : le χ^2 de Pearson et le rapport de vraisemblance G^2.

Plus précisément, la distribution estimée est calculée par le produit des effectifs marginaux divisé par l'effectif total :

$$m_{ij} = \frac{n_{i+} n_{+j}}{n_{++}}$$

m_{ij} est l'effectif estimé en cas d'indépendance
n_{i+} est la somme des effectifs du tableau dans la ligne i
n_{+j} est la somme des effectifs du tableau dans la colonne j
n_{++} est l'effectif total

Deux mesures des écarts entre effectifs estimés en cas d'indépendance m_{ij} et effectifs observés n_{ij} sont proposées pour leur propriétés statistiques intéressantes qui ne se limitent pas d'ailleurs à ce cas particulier :

$$\chi^2 = \sum\sum \frac{(n_{ij} - m_{ij})^2}{m_{ij}} \qquad G^2 = 2\sum\sum n_{ij} \log(\frac{n_{ij}}{m_{ij}})$$

Ces statistiques suivent asymptotiquement des lois du χ^2 à 3 degrés de liberté sous l'hypothèse d'indépendance : $(nb\,colonnes - 1).(nb\,lignes - 1) = (2-1).(4-1) = 3$

Une condition assurant la convergence demande que les effectifs théoriques estimés m_{ij} ne soient pas trop faibles ; une règle classique exige qu'ils soient supérieurs à 5, quite à regrouper éventuellement des classes.

Dans le cas présent, on regroupe les deux classes *bon* et *très bon*.

Table des effectifs, avec entre parenthèses les effectifs théoriques estimés :

accueil ▼	1ère année	2ème année	total
mauvais	74 (61.9)	31 (43.1)	105
moyen	122 (117.2)	77 (81.8)	199
bon	18 (35.9)	43 (25.1)	61
très bon	4 (2.9))	1 (2.1)	5
total	218	152	370

Le calcul des écarts à l'indépendance donne :
$\chi^2 = (74 - 61.9)^2/61.9 + (31 - 43.1)^2/43.1 + ... = 28.98$
et $G^2 = 29.05$ (3 degrés de liberté)

Après regroupement :

accueil ▼	1ère année	2ème année	total
mauvais	74 (61.9)	31 (43.1)	105
moyen	122 (117.2)	77 (81.8)	199
bon, très bon	22 (38.9)	44 (27.1)	66
total	218	152	370

$\chi^2 = 24.11$ $G^2 = 24.04$ (2 degrés de liberté)

La comparaison des deux distributions (colonnes) est souvent faite en comparant les "chances" ou les "cotes" (*odds*) par leur rapport (*odds ratio*) plus ou moins proche de 1 :

Par exemple on peut comparer les accueils très négatifs en faisant le rapport :

$$effectif\_des\_mauvais \,/\, effectif\_des\_moyens$$

en prenant une catégorie de référence "accueil moyen", ce qui donne ici 74/122=0.61 pour la première année et 31/77=0.40 pour la seconde.

Pour la catégorie regroupée "bon, très bon" on trouve 0.18 et 0.57. Le choix de la catégorie de référence n'est pas fondamentale puisque le rapport de deux odds ratios est encore un odds ratio :

$$\frac{Mauvais\_contre\_Moyen}{Bon\_contre\_Moyen} = Mauvais\_contre\_Bon$$

Ce qui s'exprime dans le tableau regroupé par :

$$\frac{74}{122} \times \frac{122}{22} = \frac{74}{22} = 3.36 \;\; et \;\; \frac{31}{122} \times \frac{122}{44} = \frac{31}{44} = 0.70$$

Pour tenir compte de l'ordinalité des catégories, on peut s'intéresser à d'autres rapports, dérivés des précédents comme :

Mauvais contre (moyen,bon et très bon)
(Mauvais et moyen) contre (bon et très bon)
(Mauvais, moyen et bon) contre (très bon)

Traiter des données ordinales en considérant les fréquences cumulées plutôt que les fréquences brutes, présente de nombreux avantages en statistique pour avoir des résultats stables devant les erreurs d'appréciation dans les réponses, dans le découpage en catégories, dans les regroupements éventuels de catégories. On verra en détail comment en tenir compte.

Les deux mesures proposées pour s'assurer de l'indépendance (ou *a contrario* de la dépendance) traitent les lignes et les colonnes de la même façon, ce qui laisse penser que la lecture du tableau aurait pu se faire ligne par ligne.

Si on s'en tient à la seule indépendance, les deux lectures sont équivalentes, mais si on veut savoir comment on s'éloigne de l'indépendance, on doit s'expliquer sur le type de dépendance qui est recherchée, et pour cela il faut donner un sens à la répartition d'une sous-population. Il peut s'agir des étudiants d'une année (ou filière), lecture par colonnes, ou à un type d'accueil (les mauvais accueils, par exemple), lecture par lignes. Dans ce dernier cas, l'aléa porte sur l'appartenance à telle année d'étude, de l'étudiant bien ou mal accueilli.

Une lecture par lignes conduit à des probabilités conditionnelles à l'accueil, la lecture par colonnes, à des probabilités conditionnelles à l'année d'étude.

L'objectif d'un examen ligne à ligne serait lié à la question : **qui sont les mécontents?** Alors que l'étude initiale posait la question : **comment est-on accueilli?**

Les deux points de vue seront repris pour voir à quelles mesures de dépendance ils conduisent, mais les rapports des chances (odds ratios) auront la même signification dans les deux cas.

Cet exemple montre que plusieurs représentations du tableau de données sont possibles ; les rapports de chances fournissent un instrument pratique pour décrire les différences entre deux distributions.

3.3 Exemple 3 : une promotion de diplomés

Cet exemple est très simple, la réponse binaire fait penser à un score ; les deux variables de classement pourraient avoir de nombreuses modalités : ce sont des magasins et des produits, ou des hopitaux et des traitements, ou des types de famille et des types de logement, etc.

Les faibles effectifs figurant dans le tableau ne permettront pas de construire des estimations très précises (variance élevée des estimateurs), mais les effectifs peu élevés facilitent les calculs qu'on peut faire à la main. C'est donc un exemple qui porte plus sur la technique (estimateurs, tests) que sur la signification statistique des résultats.

La population statistique est constituée de 38 étudiants en DESS, dont on ne retient que la note globale en fin d'année.

Les variables de classement sont binaires : région de provenance (Paris, Province) et diplôme à l'entrée (maîtrise Mass ou MST, maîtrise d'économie), correspondant à deux filières. La réponse est la note finale, qui est réduite à deux modalités : moyenne (< 12) ou bonne (≥ 12). L'objectif est de comparer les résultats (moyen, bon) des étudiants, et de savoir si on peut les associer à la région ou à la filière d'origine.

Il s'agit d'une étude rétrospective car on s'en tient à une population fixée qui n'est pas représentative de la population totale des étudiants parisiens, provinciaux, ni de celle des titulaires d'une maîtrise.

Les variables de classement peuvent être considérées comme explicatives *conditionnellement* au fait que l'étudiant appartient à l'échantillon. Si les étudiants s'étaient présentés au hasard, et avaient été sélectionnés au hasard sur la France entière, les variables seraient éventuellement explicatives. En l'absence de connaissance des raisons pour lesquelles les candidatures ont été déposées, puis retenues, tout résultat sera conditionné à l'échantillon donné.

Pour donner des probabilités de réussite, selon la maîtrise par exemple, il faudrait avoir une probabilité a priori pour qu'un étudiant ayant telle maitrise soit retenu dans ce DESS. Par contre, dans des conditions de recrutement identiques d'une année à l'autre (d'une région à l'autre, d'une maîtrise à l'autre) on pourrait comparer les distributions des notes d'une année à l'autre, etc.

Si le recrutement Paris-province n'est pas identique, on ne peut donc pas répondre à la question : **un étudiant parisien a-t-il plus de chances de réussir qu'un étudiant provincial?**

On ne peut pas non plus répondre à la question : **a-t-on plus de chances de trouver des parisiens parmi les bons ?** En effet, les probabilités *portent sur la note*, et non sur la région qui, elle, est déterminée par le recrutement, sur lequel on n'a pas d'information.

Par contre on peut répondre à la question : **un étudiant provincial *inscrit* dans ce DESS réussira-t-il mieux qu'un étudiant parisien?** Il s'agit de probabilités conditionnelles à l'échantillon.

Si on s'en tient au seul critère de classement Paris-province, on remarque que les répartitions (colonnes) sont identiques :

	Paris	province	total
note < 12	**5**	**6**	11
note ≥ 12	**10**	**12**	22
total	15	18	33

Peut-on en déduire que la région "ne compte pas"?

Si on s'en tient au critère Economie/Mass-Mst, les chances sont de 14 contre 8 (odds ratio 1.75) pour Economie et 8 contre 3 (odds ratio 2.67) pour Mass-Mst

	Economie	Mass-Mst	total
note < 12	**8**	**3**	11
note ≥ 12	**14**	**8**	22
total	22	18	33

Ces deux tableaux résumés ne donnent qu'une information partielle parce qu'ils ne tiennent compte que de distributions marginales (une seule variable de classement), qui conduirait par exemple à considérer la variable de classement Paris-province comme superflue.

Il ne faut pas beaucoup d'attention pour détecter une différence fondamentale entre les deux sous-populations Paris, province :

Paris		
	Economie	Mass,Mst
note < 12	3	2
note ≥ 12	6	4
odds	**2**	**2**
odds ratio (economie/Mass,Mst)=1		

province		
	Economie	Mass,Mst
note < 12	5	1
note ≥ 12	8	4
odds	**1.6**	**4**
odds ratio (economie/Mass,Mst)=0.4		

A Paris le classement Economie/Mass-Mst ne modifie pas les chances de succès (2 contre 1). L'effet du classement se traduit alors par un odds ratio de 2/2=1, alors que pour les étudiants de province il est 1.6/4=0.4.

Quand les odds ratios sont égaux on dit qu'il y a *association* du couple note-maitrise, *homogène* entre Paris et province. Cette homogénéité est mesurée avec la statistique de Breslow-Day valable pour les tables simples de structure $(2 \times 2 \times K)$

$$\sum \frac{(n_{ijk} - m_{ijk})}{m_{ijk}}$$

i, j correspondent aux variables associées (notes, filière), k correspond à la variable de classement qui stratifie la population (région)

n_{ijk} est l'effectif observé

m_{ijk} est l'effectif estimé en cas de odds ratios égaux.

La statistique de Breslow-Day suit asymptotiquement un χ^2 à $(nombre\,de\,strates - 1)$ degrés de liberté sous l'hypothèse que les odds ratios sont égaux pour toutes les strates k.

Ici, les strates étant les régions, il s'agit d'un χ^2 à 1 degré de liberté, la statistique vaut 0.30 mais le faible effectif de la table interdit d'en tirer une conclusion.

On pourrait l'utiliser dans un cas plus général que celui-ci où il y aurait plus de deux sous-populations (départements par exemple) et dont on voudrait comparer les résultats (effectifs de reçus au bac) suivant un critère de classement (scientifique-littéraire).

Au lieu de tester l'homogénéïté, on peut tester l'*indépendance conditionnelle* qui doit se traduire par des odds ratios proches de 1 dans toutes les strates (et pas seulement égaux). Il s'agit par exemple des tests de Cochran-Mantel-Haenszel et dérivés qui sont applicables à des tables $(I \times J \times K)$ plus complexes que la précédente.

En somme, dès qu'il y a plusieurs variables de classification, les réponses et les chances, mesurées marginalement et conditionnellement peuvent être différentes, comme ici, et même apparemment contradictoires. Un cas particulier célèbre est le "paradoxe de Simpson".

Les odds ratios sont toujours un bon moyen pour évaluer l'influence respective de chaque variable de classification sur les réponses.

3.4 Exemple 4 : concours de Premier Surveillant

Il s'agit d'examiner les classes d'âge et d'ancienneté des promus (reçus au concours interne de Premier Surveilant dans l'Administration Pénitentiaire). En fait, deux points de vue complètement différents sont envisageables ; selon qu'on s'intéresse au *devenir* d'une classe d'âge où la population statistique est une classe d'âge, ou qu'on s'intéresse à l'âge d'une classe de promus où la population est celle des promus.

Dans le premier cas la question est **où va t-on?** On parle alors de *cohortes* (sous-populations), et d'étude *prospective* (évolution ou suivi, des individus). Les sous-populations sont fixées a priori et leur comportement est aléatoire.

Dans le second cas on s'intéresse à l'état des lieux , la question est **qui sont les reçus?** sans savoir exactement *comment et pourquoi* ils le sont, et notamment sans savoir qui s'est présenté. C'est une vue instantanée, *a posteriori*.

Du point de vue statistique, il est hors de doute que l'incertitude, pour un individu, porte sur le fait d'être promu ou non. Mais dans le premier cas (suivi de cohorte ou étude prospective) la population est définie au départ alors que dans le second (coupe instantanée ou étude rétrospective) elle est définie à l'arrivée : on dit aussi qu'elle est *endogène* car elle est une conséquence (aléatoire) de l'expérimentation.

Dans ce dernier cas on ne parlera pas de probabilité d'être reçu, puisqu'on ne dispose que des reçus, mais de probabilité d'avoir tel âge le jour de la promotion. C'est une probabilité conditionnelle. Le nombre des reçus dépend naturellement non seulement de l'âge, mais aussi du nombre de personnes de cet âge qui étaient susceptibles d'être promues.

Les enquêtes portant sur des cohortes, comparent les réactions de sous-populations d'âge différent face à un défi, un danger, une situation (commerciale, financière, médicale) et supposent qu'on prenne le temps de les suivre. Si on ne dispose que d'une coupe

instantanée, il faudrait pour en tirer une information analogue mettre en regard une population neutre (*échantillon témoin*) ou disposer de probabilités a priori correspondant à une population "normale". On pourrait alors comparer la population des promus (ou des non promus) à un instant donné soit à la population générale, soit à une sous-population, témoin, *représentative* de la population générale.

Le tableau suivant répertorie, pour la cohorte des surveillants entrés dans l'administration pénitentiaire dans les années 80 à 84, ceux qui on réussi le concours de Premier Surveillant et l'âge qu'ils avaient l'année de leur réussite :

	Hommes de la génération 80-84 (date d'entrée)				
âge	30 ans et moins	31-35 ans	36-40 ans	41 ans et plus	total
effectif	112	194	136	84	526
pourcentage	21.1 %	36.9 %	25.8 %	16.0 %	100%

Une autre forme d'étude de cohorte pourrait considérer les tranches d'âges des surveillants susceptibles de passer Premier Surveillant cette année, et de compter ceux qui sont reçus, comme dans le tableau suivant :

âge ▶	moins de 31 ans	31-35 ans	36-40 ans	41 ans et plus	total
non promus	2283 (99,9%)	2413 (98.8%)	935 (98.7%)	635 (99.1%)	6266
promus	2 (0.1%)	28 (1.2%)	12 (1.3%)	6 (0.9%)	48
total	2285 (100%)	2441 (100%)	947 (100%)	641 (100%)	6314

Les deux études ne portent pas sur les mêmes populations (générations d'entrants, contre âge actuel des candidats) et ne concernent pas les mêmes évènements (reçus jusqu'à cette année, contre reçus cette année).

Les différences sont dans le classement en sous-populations, et dans l'évènement que l'on suit.

Dans le premier tableau la population des surveillants de la tranche 31-35 ans sont ceux qui sont entrés entre 80 et 84, qui ont réussi le concours depuis, et qui ont eu ce concours à un âge compris entre 31 et 35 ans.

Dans le second tableau la population des surveillants candidats (de la tranche 31-35 ans) sont ceux qui n'étaient pas encore Premier Surveillant avant ce jour, et quelque soit leur date d'entrée dans l'Administration Pénitentiaire.

En somme, les études de cohortes demandent une définition très précise des sous-populations en cause et permettent de comparer des probabilités dans des sous-populations, alors qu'une étude rétrospective analyse un état des lieux et s'en tient à des probabilités conditionnelles à la population totale de l'échantillon.

3.5 Exemple 5 : gravité des accidents

La population est l'ensemble des accidents ayant réellement eu lieu en 1998, dans le département du Nord et impliquant deux véhicules. Il ne s'agit pas d'un sondage, ni d'une expérimentation, ni d'un essai clinique, où on contrôlerait certaines variables de classement. L'"individu statistique" est l'accident. Mais, comme on associe bijectivement à chaque accident un *conducteur responsable*, on peut aussi considérer que ce conducteur est l'individu statistique et la population est définie comme :

- *l'ensemble des conducteurs ayant provoqué un accident impliquant deux véhicules, en 1998, dans le département du Nord.*

Revenons au tableau des données et supposons qu'on connaisse la probabilité pour qu'un jeune (circulant dans le Nord en 1998) ait un accident : Prob(acc de jeune 18-24 ans) = 0.001

Hommes	**caractère de gravité de l'accident**				
	aucun tué	1 tué	2 tués	*plus de 2 tués*	total
âge < 18 ans	3712	74	4	0	3790
18-24 ans	5990	286	27	5	6308
25-64 ans	7642	517	34	10	8203
âge ≥ 65 ans	155	29	0	0	184
total	17499	906	65	15	18485

Femmes	**caractère de gravité de l'accident**				
	aucun tué	1 tué	2 tués	plus de 2 tués	total
âge < 18 ans	1489	28	1	0	1518
18-24 ans	1687	53	2	0	1742
25-64 ans	2663	95	1	4	2763
âge ≥ 65 ans	89	8	0	0	97
total	5928	184	4	4	6120

L'examen du tableau donne la probabilité pour que, dans un accident provoqué par un jeune (18-24 ans), il y ait au moins un tué :

$$\Pr(jeune + tué/acc\_de\_jeune) = \frac{286 + 27 + 5 + 53 + 2}{6308 + 1742} = 0.046$$

On en déduit la probabilité "pour qu'un jeune provoque un accident mortel"

$$\Pr(jeune + tué) = \Pr(jeune + tué/acc\_de\_jeune) \times \Pr(ac\_de\_jeune) = 0.000046$$

Il faut donc bien distinguer la probabilité pour qu'*un accident provoqué par un jeune soit mortel*, ce qui peut s'expliquer par un manque d'attention, une maladresse, etc. de la probabilité pour un jeune *provoque un accident mortel,* ce qui peut s'expliquer par le nombre de kilomètres parcourus, le type de trajet qu'il prend, l'état de santé physique ou

psychologique, la manière de conduire, etc. Dans le premier cas la population de référence est celle des accidents ayant eu lieu, dans le second cas il s'agit des usagers de la route.

Nos données ne nous permettent de comparer les risques (hommes contre femmes, jeunes contre moins jeunes), que sur la *gravité* des accidents et non sur l'*occurence* des accidents. Tant qu'on ne connait pas le nombre de jeunes qui circulent, on aura la probabilité d'être tué par un jeune, mais pas la probabilité qu'un jeune tue quelqu'un.

Ce type d'étude est *rétrospective*, elle a un caractère conditionnel, les données reflètent un état des lieux, une fois que l'accident est arrivé, par opposition à une étude prospective où on comparerait la situation (âge, genre,...) dans une population construite avant que l'accident n'arrive, et où l'accident n'est que potentiel.

De la même façon, la probabilité pour qu'un client *à risque* d'une assurance automobile ait une voiture puissante, n'est pas identique à la probabilité pour qu'un client ayant une voiture puissante soit *à risque*.

Un exemple analogue pourrait être une analyse des étudiants ayant réussi ou échoué à un examen, une analyse des âges et des formations des chômeurs d'une région, une analyse des défaillances des entreprises, des bons et des mauvais payeurs parmi les clients d'un organisme de crédit, des malades arrivant aux urgences.

D'autre part, comparons les hommes et les femmes dans la catégorie 18-24 ans, avec les tableaux résumés suivants :

effectifs	aucun tué	un tué	plus d'un tué	total
hommes (18-24 ans)	5990	286	32	6308
femmes (18-24 ans)	1687	53	2	1742
total	7677	339	34	8050

répartition globale(%)	aucun tué	un tué	plus d'un tué	total
hommes (18-24 ans)	74.41	3.55	0.40	78.36
femmes (18-24 ans)	20.96	0.66	0.02	21.64
total	95.37	4.21	0.42	100

On remarque que pour les accidents graves, caractérisés par l'existence de tués, les probabilités sont assez différentes pour les hommes et les femmes. Ce qui veut dire en gros qu'on risque plus de se faire tuer par un homme que par une femme.

Le **rapport des risques** est $(3.55+0.40)/(0.66+0.02) = 5.81$.

répartition en ligne	aucun tués	un tué	plus d'un tué	total
hommes (18-24 ans)	94.96 %	4.53 %	0.51 %	100 %
femmes (18-24 ans)	96.84 %	3.04 %	0.12 %	100 %

Les répartitions en ligne des hommes et des femmes se ressemblent avec un moindre risque de gravité pour les femmes. Ce qui veut dire que dans un accident, si c'est une femme qui conduit, on risque moins de se faire tuer, mais pas de façon aussi nette que dans le calcul précédent.

Le nouveau rapport des risques est (4.53+0.51)/(3.04+0.12)=1.59.

Il est simple de voir d'où vient le paradoxe : même si les hommes et les femmes étaient indiscernables par leur répartition en ligne, le fait qu'il y ait plus d'hommes que de femmes dans l'échantillon implique un risque plus grand de se faire tuer par un homme que par une femme. Et si on interdit aux hommes de circuler, les femmes vont devenir franchement dangereuses...

Si on raisonne en **rapport de chances** (odds ratios) il n'y a plus de paradoxe :
Les "chances" pour qu'il y ait au moins un tué contre pas de tué, sont :

avec un conducteur (homme de 18-24 ans) :

$$\frac{286+32}{5990} = 0.053$$

avec une conductrice (femme de 18-24 ans) :

$$\frac{53+2}{1687} = 0.033$$

Le odds ratio mesurant le danger (la gravité) selon que c'est un homme ou une femme qui conduit est alors

$$\frac{0.053}{0.033} = 1.6$$

Il vaudrait donc mieux que ce soit une femme qui conduise (quand il y a un accident).
Mais qui provoque les accidents (nombre d'accidents par kilomètre parcouru)? Plutôt les hommes ou plutôt les femmes? C'est un tout autre problème.

Cet exemple met en évidence l'importance des probabilités conditionnelles et les dangers de l'interprétation des résultats ; si les probabilités sont quelquefois ambiguës, les odds ratios ont toujours un sens. Ce qui explique en partie l'utilisation fréquente du modèle logistique pour les données catégorielles.

3.6 Exemple 6 : aspirine

L'échantillon étudié est composé d'individus à risques, qui ont accepté de participer à l'essai. Ils ne sont pas représentatifs de la population totale. Cet échantillon a été découpé en deux classes de patients : les uns prennent de l'aspirine, les autres prennent un placebo.

Le découpage est controlé par un tirage aléatoire qui a été fait de telle façon que la structure des deux sous-échantillons soit la même pour l'âge, le sexe (et d'autres variables individuelles). L'âge, le sexe, le traitement sont donc exogènes (contrôlés).

La réponse est la valeur du burden (expression de l'étendue des adénomes présents lors de la coloscopie). C'est une variable aléatoire.

On désire savoir si le médicament est efficace, au sens où les probabilités pour que le burden soit faible (non récidive) est différente selon le traitement ; et si cette différence dépend de facteurs propres aux individus comme l'âge, le sexe, les antécédents.

Y a-t-il un facteur commun de récidive dû à l'âge, au sexe, indépendamment du traitement?
Y a-t-il un effet général du traitement indépendamment des autres facteurs?
Le traitement agit-il de façon différente selon l'âge?
Etc.

Cette expérimentation s'appliquerait aussi bien pour tester l'impact d'une mesure concernant la sécurité dans un quartier, l'impact d'une nouvelle présentation d'un produit (carte de fidélité ou emballage), d'une nouvelle organisation du travail dans une entreprise, etc.

Le tableau de contingence est présenté de façon à mettre en valeur l'objectif : l'effet du traitement sur la non récidive (burden faible).
Sur chaque ligne du tableau on trouve les réponses (taille du burden) pour chaque traitement. Une ligne concerne une sous-populations caractérisée par l'âge, puis par le sexe ,etc. Il y a donc quatre petits tableaux de contingence

	Aspirine			Placebo		
burden	$< 6\,mm$	$\geq 6\,mm$	total	$< 6\,mm$	$\geq 6\,mm$	*total*
âge ▼						
≤ 56	50	3	53	34	7	41
56-64	39	6	45	32	7	39
≥ 65	23	4	27	22	10	32
sexe ▼						
femmes	36	1	37	32	4	36
hommes	76	12	88	56	20	76
antécédents ▼						
non	88	9	97	69	12	81
oui	24	4	28	19	12	31
initial ▼						
nombre ≤ 2	87	6	93	72	9	81
nombre > 2	25	7	32	16	15	31
Ensemble	112	13	125	88	24	112

En termes de risques, on pourrait comparer les risques de récidive (burden $\geq 6mm$) selon le traitement, et cela pour chaque sous-population : par exemple pour les patients de la première ligne (âge ≤ 56), les risques sont

$$3/53 = 0.0566 \; et \; 7/41 = 0.1707$$

Le rapport des risques a un sens dans la mesure où le tirage aléatoire a été bien fait (randomisation) et donc que la sous-population est représentative de la population générale, dite à risque

$$0.0566/0.1707 = 0.33$$

Le odds ratio est un peu différent :

$$\frac{3/50}{7/34} = 0.29$$

On n'a pas essayé de représenter les autres croisements âge par sexe, âge par antécédents, etc. qui, dans le cas présent, n'apportent pas d'information sur l'effet du traitement.

Comparer les rapports de risques (Relative Risk) n'est pas facile, et reste ambigu quand il s'agit de classements différents (âge, sexe, antécédents). Les chances (cotes, odds) et leur leur rapport (odds ratio) ont une propriété multiplicative qui donne un sens aux comparaisons :

quand on passe d'une classe d'âge à une autre, par combien sont multipliées les chances de non récidive ?

Il est raisonnable de postuler que chaque individu traité a une probabilité de récidive (ou non récidive) qui dépend du traitement et des autres facteurs exogènes (âge,...), que les effets sont indépendants en probabilité d'un individu à l'autre. En somme que l'occurence de récidive est une variable de Bernoulli, que l'effectif des récidives pour chaque sous-population est une variable binomiale.

D'autre part, l'objectif affiché présuppose que le paramètre de la loi de Bernoulli, est fonction des catégories (traitement, âge, etc.). En cherchant à évaluer "comment les chances sont multipliées" on sous-entend que les différents effets sont multiplicatifs.

Explicitons les conséquences de ce sous-entendu.

Par exemple dire que le traitement est efficace s'exprime sous la forme :

$$chances\ de\ non\ récidive\ avec\ aspirine = \frac{Prob(burden faible/Aspirine)}{Prob(burden fort/Aspirine)}$$

$$chances\ de\ non\ récidive\ sans\ aspirine = \frac{Prob(burden faible/Placebo)}{Prob(burden fort/Placebo)}$$

L'efficacité du traitement est mesurée par le rapport (odds ratio) :

$$OR = \frac{Prob(burden faible/Aspirine)}{Prob(burden fort/Aspirine)} / \frac{Prob(burden faible/Placebo)}{Prob(burden fort/Placebo)}$$

Ce nombre n'a d'intérêt (statistiquement) que s'il représente un effet intangible, qu'il apporte une information valable pour toute la population étudiée, qu'il est le même pour chaque individu.

C'est donc qu'il existe un nombre constant E tel que :

$$\frac{\Pr(burden faible/Aspirine)}{\Pr(burden fort/Aspirine)} = E \times \frac{\Pr(burden faible/Placebo)}{\Pr(burden fort/Placebo)}$$

Pour deux individus A (traité à l'asppirine), B traité au placebo, on traduira cette égalité par:

$$\frac{P_A}{1-P_A} = E\,\frac{P_B}{1-P_B}$$

P_A est la probabilité de non récidive de
P_B est la probabilité de non récidive de B
E est une constante indépendante de l'individu

Maintenant supposons que le traitement soit plus efficace pour les femmes que pour les hommes, il faudra introduire deux constantes E_H et E_F au lieu d'une seule constante,

$$\frac{P_{AH}}{1-P_{AH}} = E_H\,\frac{P_{BH}}{1-P_{BH}} \quad et \quad \frac{P_{AF}}{1-P_{AF}} = E_F\,\frac{P_{BF}}{1-P_{BF}}$$

Si l'effet conjugué du traitement et du sexe peut être considéré comme le produit d'un effet traitement valable pour tous et d'un effet sexe valable pour tous, les deux effets E_H et E_F seront exprimés sous la forme :

$$E_H = E_T \quad et \quad E_F = E_T.E_S$$

E_T est l'effet du traitement (aspirine contre placebo)
E_S est l'effet du sexe (femme contre homme)

L' effet du seul traitement égal pour les deux sexes s'écrit alors :

$$\frac{P_{AH}}{1-P_{AH}} = E_T\frac{P_{BH}}{1-P_{BH}} \quad et \quad \frac{P_{AF}}{1-P_{AF}} = E_T\frac{P_{BF}}{1-P_{BF}}$$

L'effet du sexe égal pour les deux traitements s'écrit :

$$\frac{P_{AF}}{1-P_{AF}} = E_S\,\frac{P_{AH}}{1-P_{AH}} \quad et \quad \frac{P_{BF}}{1-P_{BF}} = E_S\,\frac{P_{BH}}{1-P_{BH}}$$

L'effet conjugué sexe et traitement s'écrit:

$$\frac{P_{AF}}{1-P_{AF}} = E_T.E_S \frac{P_{BH}}{1-P_{BH}}$$

dont on déduit :

$$\frac{P_{AH}}{1-P_{AH}} = \frac{E_T}{E_S} \frac{P_{BF}}{1-P_{BF}}$$

En prenant le logarithme de ces expressions on trouve un modèle linéaire par rapport aux paramètres $\alpha = \exp(E_T)$, $\beta = \exp(E_S)$.

En somme, chaque individu suit une loi de Bernoulli de paramètre $p(s,t)$, ces lois sont indépendantes, la fonction qui relie le paramètre de cette loi aux caractéristiques de l'individu (variables exogènes) a la forme :

$$Log\ \frac{p(s,t)}{1-p(s,t)} = \alpha.t + \beta.s$$

ou de façon équivalente :

$$p(s,t) = \frac{\exp(\alpha t + \beta s)}{1+\exp(\alpha t + \beta s)} = \frac{1}{\exp(-\alpha t - \beta s)+1}$$

$p(s,t)$ désigne la probabilité de non récidive
$t \in \{0,1\}$pour le traitement (aspirine, placebo)
$s \in \{0,1\}$pour le sexe (homme, femme)
α, β sont les paramètres à estimer.

L'espérance mathématique de la réponse, $p(s,t)$, est liée à la combinaison linéaire des variables de classement, $\alpha t + \beta s$, par une fonction, dite "de lien"(*link*) qui est dans le cas présent une logistique :
$Logit(p(s,t)) = \alpha t + \beta s$

Cet exemple montre comment l'examen des données et les objectifs fixés ont conduit à l'écriture d'un modèle logistique.

Ce modèle n'est évidemment pas le seul envisageable, il est très utilisé pour combiner des facteurs de risque, il ne fait pas d'hypothèse sur la construction de l'échantillon observé. Dans cet exemple clinique, ce sont les traitements qui sont "controlés", c'est à dire fixés par l'expérimentateur (mais pas n'importe comment), contrairement aux exemples précédents où l'échantillon est donné sans préalable.

Les probabilités qu'on déduit des odds ratios sont conditionnelles à la population et ne sont pas toujours intéressantes. En particulier, quand on parle de risques de récidive, il faut savoir si on fait référence aux odds ratios ou aux probabilités.

3.7 Exemple 7 : les étrangers de Paris

On compare maintenant les deux recensements de 1990 et de 1999. Chaque quartier est une entité dont on veut examiner l'évolution. Cette évolution se traduit par la modification du nombre d'habitants et de leur répartition dans le quartier selon deux critères, le genre et la nationalité. Les réponses des quartiers sont des répartitions internes qui peuvent se réduire par exemple aux pourcentages d'étrangers (hommes et femmes).

Par exemple :

quartier	Goutte-d'Or (ouvriers)			
année	1990		1999	
nationalité	française	étrangère	française	étrangère
hommes	63 %	37 %	70 %	30 %
femmes	72 %	28 %	73 %	27%
quartier	Chaillot (cadres retraités)			
année	1990		1999	
nationalité	française	étrangère	française	étrangère
hommes	80 %	20 %	81 %	19 %
femmes	79 %	21 %	80 %	20%

La lecture "en ligne" a été privilégiée, ce qui revient à comparer des probabilités (baisse des étrangers à la Goutte d'Or, surtout pour les hommes 37% à 30%), ou un rapport des "risques" de 0.30/0.37=0.81 de 1990 à 1999. Ce rapport peut être comparé à celui de Chaillot qui est proche de 1.

Cette examen du tableau considère que la population statistique est l'ensemble des quartiers, et qu'on la compare à deux époques différentes, comme dans une enquête où on aurait posé la même question sur une situation (revenu, diplôme, emploi, état financier), un comportement, une opinion electorale ou un sujet d'actualité, deux fois à trois mois d'intervalle. Ce pourrait être aussi la répartition des ventes dans un magasin, un diagnostic médical avec cette particularité que c'est exactement la même population qui est interrogée. Il s'agit d'*un suivi* avec éventuellement des variables explicatives externes.

Pour être plus complet, on aurait pu, dans la même optique (évolution des quartiers), s'intéresser à la répartition conjointe des catégories genre et nationalité de chaque quartier, on passerait alors d'une comparaison de lois binomiales à celle de lois multinomiales. Les données seraient présentées alors sous la forme :

quartier	Goutte-d'Or (ouvriers)			
année	1990		1999	
nationalité	française	étrangère	française	étrangère
hommes	32.7 %	19.1 %	35.9 %	15.5 %
femmes	34.8 %	13.4 %	35.5 %	13.1%
total	100%		100%	

quartier	Chaillot (cadres retraités)			
année	1990		1999	
nationalité	française	étrangère	française	étrangère
hommes	35.3 %	10.0 %	37.6 %	9.0 %
femmes	43.4 %	11.3 %	42.7 %	10.7 %
total	100%		100%	

Il serait alors naturel de comparer des odds ratios. A la goutte d'or, en 1990, considérons la catégorie "hommes de nationalité étrangère", leur taux de présence dans le quartier est 19.1%. Leur cote (chances d'en rencontrer un contre ne pas en rencontrer) est 19.1/(100 - 19.1). En 1999 la cote devient 15.5/(100 - 15.5). Le rapport des cotes décrit l'évolution des chances de rencontrer un homme de nationalité étrangère de 1990 à 1999 :

$$\frac{19.1}{100-19.1}/\frac{15.5}{100-15.5} = 1.27$$

Le même facteur multiplicatif calculé pour les femmes de nationalité étrangère est :

$$\frac{13.4}{100-13.4}/\frac{13.1}{100-13.1} = 1.03$$

La variation de structure de la population du quartier s'exprime aussi bien par un rapport des proportions (19.1/15.5=1.23) pour les hommes, et (13.4/13.1=1.02) pour les femmes. Mais sans la connaissance des populations totales du quartier, ces proportions ne correspondent pas exactement aux risques de rencontre et un rapport des rapports de proportions n'a pas de sens très clair.

Dans le quartier de Chaillot on retrouve les odds ratios de 1.12 et 1.06 de sorte que l'effectif des hommes étrangers a moins diminué que'à la Goutte-d'Or. Ce qui peut s'interpréter comme une baisse de la main d'oeuvre masculine étrangère dans un quartier à prédominance ouvrière.

On peut s'intéresser à des différences de comportement d'un ou de plusieurs quartiers fixés. On dispose alors de deux photographies de la population parisienne (ou d'une sous-population), les variables de classement sont le genre, la nationalité.

Mais l'analyse d'une répartition complexe ne se réduit pas à quelques odds ratios ; on aimerait savoir si, sur Paris, les variations démographiques sont liées au quartier, sont homogènes d'une période à l'autre, et que ce soit vu aussi en relation avec le genre et la nationalité. Il s'agira alors d'évaluer des probabilités de variables de classement conditionnellement à d'autres variables, de détecter des indépendances conditionnelles.

On considère donc les 80 quartiers comme des individus et on explique leur évolution (en tenant compte du nombre d'habitants) avec des variables de classement telles que le genre, la nationalité, le type de quartier. On *suit l'évolution* des quartiers d'un recensement à l'autre.

Une forme plus générale que la régression logistique peut répondre à ces objectifs : il s'agit du modèle loglinéaire qui consiste à définir l'effectif probable d'une cellule dans un

tableau de contingence à plusieurs dimensions, comme un produit d'effets conditionnellement aux variables de classements, ou à leur combinaisons.

Par exemple on retiendra les trois variables A (pour année), G (pour genre), N (pour nationalité) ; l'effectif d'une cellule sera indexée par (a, g, n, q) ou a, g, n ont deux valeurs possibles et q en a 80.

La loi du nombre d'habitants de la cellule (a, g, n, q) est multinomiale. L'espérance mathématique du nombre d'habitants (quelque fois appelée "nombre attendu") associée à cette cellule H_{agnq} est paramétrée par exemple sous la forme suivante :

$$\log H_{agnq} = \lambda + \lambda_a^A + \lambda_g^G + \lambda_n^N + \lambda_{an}^{AN} + \lambda_{gn}^{GN}$$

λ_a^A est un effet global de 1990 à 1999 (par exemple une diminution de la population globale)

λ_g^G est un effet global de proportion entre hommes et femmes

λ_n^N est un effet global de proportion entre français et étrangers

λ_{an}^{AN} est une interaction entre la nationalité et l'année traduisant une évolution de la proportion d'étrangers de 1990 à 1999.

λ_{gn}^{GN} est une interaction entre le genre et la nationalité traduisant des différences entre la répartition de la population étrangère (hommes, femmes) et la répartition (hommes, femmes) de la population française.

λ est un coefficient de calibrage lié à la population totale.

Cet exemple montre qu'il y a plusieurs façons d'examiner les données selon les informations qu'on y cherche :

des tables partielles où la notion de réponse à un sens global (nombre d'étrangers, nombre de femmes),

des comparaisons de répartitions (évolution de 1990 à 1999),

une analyse globale où toutes les variables interviennent et dont on essaie de mesurer les effets moyens (marginaux) et les interactions.

Plusieurs points de vue ont été proposés selon qu'on considère une population d'individus (habitants) ou des strates d'individus (quartiers). Les différents points de vue sont simultanément pris en compte dans la représentation générale du modèle loglinéaire.

3.8 Exemple 8 : les actifs résidant à Paris

L'objectif est la comparaison des habitants de deux arrondissements de Paris, deux sous-populations. On déterminera ce qui les distingue, l'âge, le niveau d'étude, la catégorie socio-professionnelle.

Contrairement à l'essai clinique de l'exemple 5, la population d'arrondissement n'est pas à expliquer, du moins pas par le niveau d'étude où l'âge des habitants. S'il y a une explication, il faudrait tenir compte de raisons historiques ou de moyens financiers. Avec

les seules données dont on dispose, on se limite à une description des deux sous-populations dont on étudiera la répartition interne selon les trois critères (classe d'âge, niveau d'étude, catégorie socio-professionnelle). Ces répartitions sont aléatoires et dérivent chacune d'une loi multinomiale ; on admet qu'elles sont indépendantes.

Une variable pourrait éventuellement introduire un caractère explicatif au modèle, c'est la distance entre le domicile et le lieu de travail (du chef de famille). On peut tester si cette variable permet d'expliquer des différences d'un arrondissement à l'autre et si elle apporte une information spécifique à côté des critères déjà retenus. Cette variable nous intéresse parce qu'elle peut-être traitée qualitativement ou quantitativement.

Si on avait étudié les migrations, c'est-à-dire la répartition des nouveaux habitants depuis le recensement précédent (dont les données sont d'ailleurs disponibles), on pourrait écrire un modèle explicatif et prendre comme loi de probabilité du nombre d'arrivants, dans une catégorie, une loi de Poisson avec un paramètre qui s'écrirait, par exemple, par le produit de trois facteurs (correspondant aux trois variables retenues). L'étude serait assez différente.

Cet exemple ressemble aux études dites "avec échantillon témoin" (case control) où on compare une population particulière à une autre censée représenter le "tout-venant" ou plus précisément la population totale. Dans notre cas, on ne peut pas vraiment dire qu'un habitant du treizième arrondissement représente le parisien moyen et que celui du cinquième est un cas "intéressant" comme on le ferait dans une étude médicale ou de marketing, mais dans la mesure où les deux répartitions sont indépendantes (ce qui n'était pas le cas dans l'exemple sur le goût du thé), les estimations des odds ratios sont les mêmes.

Une problématique analogue se retrouve quand on construit un score afin d'évaluer une mesure publicitaire par son impact sur un échantillon de clients ; il est souhaitable de disposer d'un échantillon témoin pour vérifier que les variations de répartition sont bien la conséquence de la mesure publicitaire. Sans échantillon témoin, on risque de faire des confusions entre probabilités et probabilités conditionnelles.

Il s'agit maintenant d'aborder la comparaison des deux arrondissements par la nature des classements (*covariates*).

L'arrondissement, nous l'avons dit, met en évidence deux populations différentes, et supposées indépendantes, de sorte qu'une lecture des tableaux doit se faire séparément par arrondissement, même si on y cherche des facteurs communs. On privilégie donc une lecture simple des répartitions (en pourcentage) dans chaque arrondissement.

	classes d'âge			
résidence ▼	20-29	30-39	40-49	50-59
5ème Ardt	**25.04 %**	**29.25 %**	**27.37 %**	**18.33 %**
13ème Ardt	**25.43 %**	**30.53 %**	**26.71 %**	**17.33 %**

	niveau d'études			
résidence ▼	<=brevet	bac	1e cycle	2e cycle
5ème Ardt	**31.56 %**	**13.45 %**	**12.19 %**	**42.80 %**
13ème Ardt	**47.92 %**	**15.71%**	**11.90 %**	**24.77%**

résidence ▼	**catégories socio-professionnelles***				
	ouv	emp	proi	pic	cpis
5ème Ardt	**7.09 %**	**19.67 %**	**20.25 %**	**7.31 %**	**45.69 %**
13ème Ardt	**13.46 %**	**28.19 %**	**23.62 %**	**4.84 %**	**29.88 %**

* *OUVrier, EMPloyé,PROfession Intermédiaire, Patrons Industrie et Commerce, Cadres et Professions Intellectuelles Supérieurs*

Il n'est pas difficile de constater que les répartitions des âges sont proches dans les deux arrondissements alors que des disparités apparaissent pour les niveaux d'étude et les catégories professionnelles.

Pour évaluer les différences des répartitions en cinq classes, on peut comparer les probabilités en ligne, ou les chances en référence à une classe d'âge de référence. La comparaison doit tenir compte de la variance des statistiques comparées ; il faut donc revenir aux distributions initiales.

Par exemple avec le modèle :

$$Log\frac{p_{ij}}{p_{0j}} = a_j + b_i$$

p_{0j} désigne la probabilité associée à la classe d'âge de référence dans l'arrondissement j

p_{ij} désigne la probabilité associée à la classe d'âge i dans l'arrondissement j

a_j $(j = 1, 2)$ et b_i $(i = 1, 2, 3)$ sont des paramètres à estimer.

Les facteurs âge et arrondissement sont multiplicatifs, on peut tester l'égalité des a_j (pas de différence entre les deux arrondissements), ou celle de certains b_i (deux classes d'âges ont le même rapport à la référence et pourraient donc être regroupées), la nullité d'un b_i signifie que la tranche d'âge i est proche de la classe de référence.

Ce modèle n'est pas aussi innocent qu'il le parait puisqu'il est sous-entendu qu'il y a un effet *arrondissement* **égal** pour toutes les tranches d'âges :

$$Log\frac{\Pr(age\, i\, dans\, le\, 5ème)}{\Pr(age\, i'\, dans\, le\, 5ème)} - Log\frac{\Pr(age\, i\, dans\, le\, 13ème)}{\Pr(age\, i'\, dans\, le\, 13ème)} = a_5 - a_{13}$$

et cette différence ne dépend pas (est la même pour) des classes d'âge i et i'.

De façon équivalente :

$$\frac{\Pr(age\, i\, dans\, le\, 5ème)}{\Pr(age\, i'\, dans\, le\, 5ème)} \Big/ \frac{\Pr(age\, i\, dans\, le\, 13ème)}{\Pr(age\, i'\, dans\, le\, 13ème)} = \exp(a_5 - a_{13})$$

Un modèle plus général comporterait des *interactions* et s'écrirait sous la forme :

$$Log\frac{p_{ij}}{p_{0j}} = a_j + b_i + c_{ij}$$

ou plus directement :

$$Log\frac{p_{ij}}{p_{0j}} = c_{ij}$$

Mais dans cet exemple particulier, le grand nombre de modalités des variables de classement exige certaines précautions dans le traitement statistique.

3.9 Exemple 9 : modèle logistique et score

La recherche d'une méthode pour distinguer deux types d'individus, par leurs traits communs ou une association de ces traits, est un cas particulier de classification (réduite à deux classes) et, plus généralement, c'est l'objet de l'analyse discriminante statistique qui est souvent isolée dans les ouvrages de statistique, bien qu'elle fasse appel à de nombreux modèles classiques.

Un aspect fondamental vient de ce qu'elle doit conduire à *une règle de décision*, construite à partir d'une population connue, et qui doit être appliquée à une population incomplètement connue ; la qualité de la règle de décision s'évalue à partir les risques de se tromper. C'est donc une méthode de classification *rétrospective* (ou supervisée) au sens où on construit la règle de décision sur du complètement connu, mais la règle de décision doit être utilisée en *prospective* (ou prévision) pour être opérationnelle.

C'est là que se situe la difficulté essentielle que nous allons regarder de près : comment détecter si le client qui se présente est un bon ou un mauvais payeur, comment évaluer le risque que l'on prend en l'acceptant, en ne l'acceptant pas, éventuellement comment évaluer les coûts afférants? Et tout cela à partir de deux échantillons (bons payeurs, mauvais payeurs) connus.

Dans le cas présent nous partons de l'hypothèse que la régression logistique va nous permettre d'élaborer un bon modèle de représentation des risques. La règle de décision consistera à fixer un seuil pour une estimation du logarithme des chances (cotes) associé au mauvais payeur. Cette estimation du logarithme des chances est appelée *score* de l'individu.

Si on a retenu (estimé) les probabilités d'être un mauvais payeur par la fonction

$$Log\frac{\Pr(Client = mauvais\ payeur)}{\Pr(Client = bon\ payeur)} = f(caractéristiques\ du\ Client)$$

Le score sera la valeur (la note) attribuée au client connu par ses caractéristiques :

$$score = f(caractéristiques\ du\ Client)$$

Le seuil du score ou la valeur plafond (en anglais cut-off) sera la valeur au delà laquelle la cote du client, ou le risque, sont trop élevés pour qu'on veuille lui accorder le prêt qu'il sollicite.

Cet exemple pourrait aussi bien concerner une évaluation des risques de défaillance d'une entreprise, d'une situation de portefeuille d'actifs, d'accidents cardiaques, de récidive (maladie ou délinquance), de catastrophe naturelle. Plus généralement, il s'agit de diagnostiquer une situation à partir d'une information limitée, sur la base de ce qu'on a pu connaître auparavant ; le diagnostic s'exprime sous la forme un indice numérique, le score.

3.10 En conclusion

Ces exemples, destinés à aborder les modèles de données catégorielles par des problèmes concrets, ont montré l'importance du protocole d'expérimentation pour évaluer l'adéquation des données aux objectifs posés.

Quelques remarques rappellent les éléments essentiels qui se dégagent de cette première exploration :

- Les données que nous avons citées sont *qualitatives*, ordonnées ou non. Mais, par exemple, on pourrait préférer que des tranches d'âge soient prises quantitativement. Il existe des modèles proches, où des variables quantitatives peuvent être mélangées avec des variables qualitatives. Les régressions logistiques et poissonniennes sont souvent utilisées avec des variables quantitatives.

- La notion de *réponse* fait référence à une occurence, variable *dichotomique* (oui, non) ou *polytomique* (A,B,C,...), pour un individu. La réponse est une variable aléatoire dans l'expérience (enquête, interrogatoire, essai clinique, diagnostic).

- Les données se présentent sous deux formes principales : soit individu par individu, avec pour chacun sa réponse et ses caractéristiques (table où il y a autant de lignes que d'individus), soit avec des nombres de réponses de chaque type, pour chaque catégorie d'individus (tableau de contingence). L'information contenue dans ce dernier cas est plus concentrée et parfaitement suffisante. Dans une logique de causalité, la réponse s'identifie à la variable expliquée *endogène*, les catégories qui définissent les sous-populations sont les variables explicatives *exogènes* qui peuvent être *contrôlées* (études cliniques, sondages par quotas,...) ou non controlées (n'ayant pas participé à la construction de l'échantillon).

- Les lois de probabilités des exemples présentés sont essentiellement les lois binomiales, multinomiales, et de Poisson. Les modèles linéaires généralisés utilisent d'autres lois de probabilité qui ont bien des points communs avec celles-ci mais l'interprétation des résultats est moins facile.

- Les paramètres utiles pour les objectifs posés peuvent être des *probabilités* d'évènements, *conditionnelles* ou non, des *chances*, ou *cotes,* (*odds*) exprimées sous la forme *p/(1-p)*, des rapports de chances, de cotes (*odds ratios)*, des espérances mathématiques d'effectifs (nombre attendu de réponses).

- Un point fondamental, qui n'apparait pas toujours explicitement dans les exemples présentés, est la "*fonction de lien*" (link function) qui définit l'espérance mathématique de la réponse comme fonction des variables de classement, caractéristiques des individus ou variables exogènes. Dans ces modèles, appelés "linéaires généralisés", la fonction de lien sera souvent une fonction numérique simple d'une combinaison linéaire de variables indicatrices.

- Dans un contexte où la causalité n'est pas essentielle ou n'a pas de sens (études rétrospectives), on s'intéressera non seulement à l'adéquation du modèle (du paramétrage) aux données, mais à des *indépendances conditionnelles* (lignes-colonnes, ou entre variables de classement), à des *associations*, *homogènes* qui simplifient le modèle et fournissent des interprétations intéressantes.

4 Exemple 1 : Père Noël

La population interrogée est un échantillon d'enfants, la réponse à la question sur la croyance au Père Noël est binaire (*oui*, non), les réponses sont indépendantes en probabilité (la réponse d'un enfant n'influence pas celle d'un autre). Un enfant est caractérisé par son âge (c'est la seule variable qu'on retient pour les distinguer), qui comporte 4 modalités :

3 ans, 4 ans, 5 ans, 6 ans.

La constitution de l'échantillon (plan d'expérience) conditionne la suite, mais pas autant qu'on pourrait le craindre.

4.1 Le tableau croisé

Commençons par une description du tableau croisé initial avec une procédure classique qui nous fournit les statistiques du Chi2 (χ^2) et le rapport de vraisemblance G^2.

	âges des enfants interrogés				
Père Noël ▼	3 ans	4 ans	5 ans	6 ans	*total*
oui	30 (25.42 %)	13 (11.02 %)	15 (12.71 %)	5 (4.24 %)	63 (53.39 %)
non	5 (4.24 %)	10 (8.47 %)	12 (10.17 %)	28 (23.73.%)	55 (46.61 %)
total	35 (29.66 %)	23 (19.49 %)	27 (22.88 %)	33 (27.97 %)	118 (100 %)

Les chiffres entre parenthèses sont les pourcentages de la population totale (118 enfants).

$\chi^2 = 34.23$ et $G^2 = 37.67$

Ces statistiques mesurent de deux façons différentes comment la répartition dans le tableau des données observées **s'écarte** de la répartition estimée dans le cas particulier où les lignes seraient proportionnelles, et donc les colonnes aussi, ce qui se comprend comme une indépendance des répartitions en ligne et en colonne.

Cette indépendance peut se tester dans le cadre de la loi multinomiale, ce qui permet de définir les lois asymptotiques des deux statistiques qui sont des lois du χ^2 avec 3 degrés de liberté, $3 = (2-1) \times (4-1)$, sous l'hypothèse d'indépendance.

Les "p-values" correspondantes sont

$$\begin{aligned} \Pr(\chi^2 &\geq 34.23) \leq 0.0001 \\ \Pr(G^2 &\geq 37.67) \leq 0.0001 \end{aligned}$$

Ce qui permet de rejeter l'hypothèse d'indépendance.

Programme SAS associé:

```
data perenoel;
input reponse $ age effectif;
```

```
datalines;
oui    3    30
oui    4    13
oui    5    15
oui    6    5
non    3    5
non    4    10
non    5    12
non    6    28
;
run;
proc freq data=perenoel;
weight effectif;
tables reponse*age    /    chisq;
run;
```

4.2 Analyse des réponses par logistique

Pour quantifier les variations des réponses d'une classe d'âge à une autre, on ne les exprimera pas a priori par des différences ou rapport des probabilités de *non* dans la population totale : parce que les tranches d'âge n'ont pas le même effectif, un rapport 4.24/8.47 pour les classes *3 ans, 4 ans* n'est pas pertinent. On préfère examiner les probabilités *relatives à chaque classe.*

Il est donc naturel de considérer que la loi de probabilité du nombre de réponses *non* dans une classe est binomiale avec une probabilité qui dépend de la classe et que ces lois sont indépendantes.

En introduisant les chances (ou cotes) on écrit le modèle sous la forme :

le nombre n_i de "non" de la classe i suit une loi binomiale $\mathcal{B}(N_i, p_i)$
p_i est la probabilité de la réponse "non" dans la classe d'âge i
N_i est le nombre total de réponse dans la classe i
La fonction de lien qui définit le paramètre p_i peut s'exprimer sous la forme

$$Log\frac{p_i}{1-p_i} = \alpha_i$$

les quatre paramètres à estimer sont les α_i
Ou de façon équivalente :

$$Log\frac{p_i}{1-p_i} = \lambda + \lambda_i^A$$

λ est une constante ("intercept"), qui correspond à la moyenne des α_i, et les quatre coefficients λ_i^A sont contraints : leur somme est nulle.

Le coefficient λ est la moyenne des logarithmes des chances (ici -0.139), mais il ne correspond pas exactement au logarithme des chances *calculé sur la loi marginale* qui donne ici Log(55/63)=-0.136. Ce coefficient n'est donc pas très intéressant, tester sa nullité non plus.

Dans ce cas particulier, les résultats de l'estimation de ces paramètres sont résumés dans le tableau suivant :

paramètres		estimation	écart-type	Test de Wald	p value
constante	λ	-0.139	0.223		
3 ans	λ_1^A	-1.653	0.408	16.422	0.0001
4 ans	λ_2^A	-0.124	0.371	0.111	0.74
5 ans	λ_3^A	-0.085	0.353	0.057	0.81
6-ans	λ_4^A	1.862	$\lambda_4^A = -\lambda_1^A - \lambda_2^A - \lambda_3^A$		

auxquels on ajoute un test global (d'indépendance de l'âge) :
avec pour hypothèse de base

$$p_1 = p_2 = p_3 = p_4$$

ce qui, compte tenu de la contrainte sur les λ_i^A s'écrit aussi :

$$\lambda_1^A = \lambda_2^A = \lambda_3^A = 0$$

La statistique de Wald donne ici 26.54, elle est rejetée avec un risque inférieur à 0.0001 (loi du χ^2 avec 3 degrés de liberté).

Le test du score donne exactement la même chose que le test du χ^2 dans le tableau croisé, le test du rapport de vraisemblance aussi.

La conclusion qu'on en tire est le rejet de l'hypothèse d'indépendance des réponses et de l'âge.

D'autres tests sont disponibles (colonne *Test de Wald*), comme par exemple la nullité d'un coefficient λ_i^A ce qui doit se comprendre, pour la classe d'âge correspondante, comme une interrogation sur *la proximité de sa distribution et de la distribution marginale.* Ce n'est pas exactement l'objectif poursuivi, mais on remarque néanmoins que ce sont les classes *3 ans* et *6 ans* qui "font la différence", qui créent la dépendance.

Un inconvénient de la présentation précédente des paramètres est la difficulté de les comparer, à moins de faire intervenir les corrélations des estimateurs. On peut alors préférer présenter les résultats sous la forme suivante qui propose un calcul de odds ratios relatifs à une classe dite de référence (ici la classe *6 ans*).

effets (odds ratios)	estimation	intervalles à 95%	
3 ans / 6 ans	0.030	0.008	0.114
4 ans / 6 ans	0.137	0.039	0.484
5 ans / 6 ans	0.143	0.042	0.483
6 ans / 6 ans	1 par définition		

Par exemple, le nombre 0.137 est l'estimation de la quantité suivante :

$$\frac{p_2/(1-p_2)}{p_4/(1-p_4)}$$

qui est le rapport des chances de répondre *non* si on est dans la deuxième classe (celle des *4 ans*) aux chances de répondre *non* quand on est dans la quatrième (celle des *6 ans*).

L'inverse de 0.137 (7.28) est le rapport des chances de répondre *oui* si on est dans la classe des *4 ans* aux chances de répondre *oui* quand on est dans celle des *6 ans*.

Dans cet exemple très simple on peut calculer les odds ratios directement :

$$0.137 = \frac{10/13}{28/5}$$

Ce rapport des cotes ou rapport des chances (odds ratio) n'est pas égal au "rapport des risques" ou rapport des probabilités de *non* (*relative risk*) qui est : $44.4/84.8 = 0.52$

Programme SAS associé :

```
proc logistic data=perenoel;
weight effectif;
class age (ref='6');
model reponse=age;
run;
```

4.3 Deuxième analyse : sous-modèle

L'examen des coefficients, comme des odds ratios, suggère qu'il n'y a pas de différence notable entre les classes *4 ans* et *5 ans*. Pour tester l'égalité des deux coefficients de ces classes, on peut très simplement reprendre la même démarche en confondant les deux classes en une classe unique puis en comparant les deux modèles par le rapport des vraisemblances ou la *déviance*.

Désignons par M un modèle à étudier et par S le modèle dit *saturé* qui est tel que *le nombre de paramètres est exactement égal au nombre de données indépendantes*. Ce nombre est ici :

(nombre de colonnes - 1)×(nombres de lignes - 1) = 4

La déviance est la différence :

$$-2(LM - LS)$$

où LM et LS sont les valeurs des logarithmes de la vraisemblance maximum de chaque modèle.

Dans de nombreux cas, comme celui où nous nous trouvons, cette déviance suit asymptotiquement une loi du χ^2 dont le degré de liberté est égal à la différence du nombre de paramètres des modèles S et M. Dans le cas présent (lois binomiales), cette statistique est

souvent appelée G^2, elle est une mesure de l'écart du modèle aux observations et ressemble à une somme pondérée de résidus (c'est la même statistique qui a été utilisée pour tester l'indépendance).

Si on dispose de deux modèles emboîtés, comme dans le cas particulier où deux classes sont regroupées, on les compare par la différence de leurs déviances qui suit asymptotiquement une loi du χ^2 dont le degré de liberté est égal au nombre de contraintes imposées pour passer du "sur-modèle" au "sous-modèle".

Si on reprend la régression logistique en confondant les deux classes, on obtient :

effets (odds ratios)	estimation	intervalles à 95%	
3 ans / 4 et 5 ans	0.212	0.071	0.637
6 ans / 4 et 5 ans	7.127	2.365	21.483

Avec $-2Log(L4) = 125.368$ pour le modèle initial avec 4 classes d'âge.

alors que pour le modèle précédent avec seulement 3 classes $-2Log(L3) = 125.373$

La différence des déviances est égale à 0.005 , ce qui est très faible pour un χ^2 à 1 degré de liberté.

La pertinence du regroupement des deux classes est confirmée.

(Le modèle à 4 classes $L4$ étant saturé, sa déviance est nulle ; la différence calculée est donc la déviance du modèle à 3 classes $L3$)

Programme SAS associé :

```
proc format; value age 4,5=45;run;
proc logistic data=perenoel;
weight effectif;
format age age. ;
class age (ref='45');
model reponse=age;
run;
```

4.4 Troisième analyse : extension à une variable continue

La variable âge qui sert à classer les enfants aurait pu être prise comme variable quantitative dans une modélisation analogue, où l'explication de la dépendance doit répondre à la question : **comment une année de plus modifie-t-elle les réponses ?**

l'hypothèse ou l'idée sous-jacente est alors que l'effet de l'âge sur les réponses est homogène et qu'il aurait un sens pour des classes d'âge non présentes dans l'échantillon comme l'âge de 7 ans (extrapolation), ou 5 ans et demie (interpolation).

Introduire une variable quantitative n'implique évidemment pas qu'elle soit explicative, elle peut tout simplement aider à bien décrire une dépendance.

Il arrive souvent qu'on dispose de variables quantitatives qu'on souhaite transformer en variables qualitatives en découpant des tranches. La discrétisation peut alors assouplir

la relation linéaire impliquée par la variables quantitative, mais avec un choix judicieux des tranches.

Le modèle proposé pour une variable quantitative (âge) sera alors :

$$Log\frac{p_i}{1-p_i} = \lambda + \alpha.age(i)$$

α est un paramètre à estimer
$age(i)$ est l'âge de l'individu i

L'indice i ne fait plus référence à la classe d'âge mais à l'âge lui-même. Le coefficient α est souvent présenté sous la forme d'un rapport de chances (odds ratio) .
Lorsque que l'individu i a un an de plus que l'individu j , alors :

$$\frac{p_i}{1-p_i} = \frac{p_j}{1-p_j} \times \exp(\alpha)$$

Dans le cas présent on obtient :

$$\alpha = 1.0286$$
$$exp(\alpha) = 2.797\,.$$
$$\text{intervalle de confiance } 95\% \ (1.889; 4.143)$$
$$-2LogL = 129.100$$

Dans ce cas particulier, où les données se regroupent en 4 classes, on peut considérer que ce modèle est un sous modèle du premier $L4$ (mais pas du second $L3$), avec une contrainte d'alignement des λ_i^A ce qui correspond à deux contraintes linéaires. De sorte qu'une comparaison des déviances a un sens :
La statistique sera :

$$129.100 - 125.368 = 3.632$$

ce qui pour un χ^2 à 2 degrés de liberté correspond à une p-value de 0.16 ; la contrainte est donc acceptable avec un risque de première espèce de 5%.

Programme SAS associé :

```
proc logistic data=perenoel;
weight effectif;
model reponse=age;
run;
```

4.5 Régression de Poisson

Les analyses précédentes procèdent de la même logique : comparer les distributions (*non* contre *oui*) des différentes classes d'âge (lecture du tableau par colonnes), soit quatre lois binomiales.

On peut aussi s'intéresser directement aux effectifs, en considérant l'ensemble de l'échantillon comme une suite d'épreuves indépendantes où le contrôle porte seulement sur la taille n de l'échantillon (cet effectif pouvant d'ailleurs n'avoir pas été fixé a priori, mais résultant de facteurs externes, comme du jour, du lieu, du temps fixé pour l'enquête). Il s'agit alors d'une loi multinomiale.

Il est aussi possible de considérer que la probabilité de trouver un nombre n_{ij} d'enfants de la classe d'âge i et d'opinion j (*oui*, *non*) est donnée par une loi de Poisson de paramètre $n\lambda_{ij}$. Pour exprimer l'effet multiplicatif qui a été accepté jusqu'ici on exprimera ce paramètre sous la forme :

$$n\lambda_{ij} = \exp(c_{ij}) \ \ ou \ \ Log(\lambda_{ij}) = \lambda_0 + \gamma_{ij} \tag{M1}$$

γ_{ij} est un paramètre à estimer, qui caractérise la classe d'âge i, la réponse j
λ_0 est une constante liée à la taille de l'échantillon.
Pour tenir compte de $\sum n_{ij} = n$, il est nécessaire d'ajouter une contrainte :

$$\sum \gamma_{ij} = 0 \ \ ou \ bien \ \gamma_{i_0 j_0} = 0 \ \ où \ i_0, j_0 \ désigne \ une \ classe \ de \ référence$$

La fonction de lien du modèle linéaire généralisé est alors le logarithme.
Le nombre de paramètres est égal au nombre d'observations ($i \times j$), le modèle est saturé.
Il existe d'autres écritures de l'expression du logarithme qui conduisent à des modèles différents comme par exemple :

$$Log(\lambda_{ij}) = \lambda_0 + a_i + b_j \tag{M2}$$

a_i est un paramètre à estimer, qui caractérise la classe d'âge
b_j est un paramètre à estimer, qui caractérise les réponses *oui* et *non*
λ_0 est une constante à estimer, liée à la taille de l'échantillon.
Pour tenir compte de $\sum n_{ij} = n$, il est nécessaire d'ajouter deux contraintes qui sont classiquement :

$$\sum a_i = \sum b_j = 0$$

ou encore, en choisissant des catégories de références :

$$a_0 = 0, \ b_0 = 0$$

Dans ce modèle, les deux facteurs qui déterminent le nombre de réponses sont indépendants (l'additivité dans le logarithme correspond à estimer les probabilités des cellules comme produit des probabilités marginales).

Autre modèle :

$$Log(\lambda_{ij}) = \lambda_0 + b_j * age(i) \tag{M3}$$

Dans le cas présent on obtient pour le **premier modèle M1** (saturé) :

paramètre	estimation	écart-type	test de Wald	p-value
constante	1.609	0.447	12.95	0.0003
non et 3 ans (p_1)	-0.000	0.633	0.00	1.00
non et 4 ans (p_2)	0.693	0.548	1.60	0.21
non et 5 ans (p_3)	0.876	0.532	2.71	0.10
non et 6 ans (p_4)	1.723	0.486	12.59	0.0004
oui et 3 ans ($1-p_1$)	1.792	0.483	13.76	0.0002
oui et 4 ans ($1-p_2$)	0.956	0.526	3.30	0.07
oui et 5 ans ($1-p_3$)	1.099	0.516	4.53	0.03
oui et 6 ans ($1-p_4$)	0.0	*référence*		

logarithme de vraisemblance : $Log(LM1) = 220.2419$

On retrouve les odds ratios de la première logistique, par exemple :

$$OR = \frac{p_2/(1-p_2)}{p_4/(1-p_4)}$$

qui est le rapport des chances de répondre *non* si on est dans la classe des *4 ans* aux chances de répondre *non* quand on est dans celle des *6 ans* ; il se déduit du modèle :

$$\begin{aligned} Log(np_2) &= 1.609 + 0.693 \\ Log(n(1-p_2)) &= 1.609 + 0.956 \\ Log(p_2/(1-p_2)) &= 0.693 - 0.956 \\ Log(p_4/(1-p_4)) &= 1.723 - 0.0 \\ Log\frac{p_2/(1-p_2)}{p_4/(1-p_4)} &= 0.693 - 0.956 - 1.723 = -1.986 \\ OR &= \exp(-1.986) = 0.137 \end{aligned}$$

Le **deuxième modèle M2** dit "d'indépendance des effets" donne :

paramètre	estimation	écart-type	test de Wald	p-value
constante	2.869	0.194	218.32	<0.0001
réponse *non*	-0.136	0.185	0.54	0.46
réponse *oui*	0.0	*référence*		
age 3 ans	0.059	0.243	0.06	0.81
age 4 ans	-0.361	0.272	1.77	0.18
age 5 ans	-0.201	0.160	0.60	0.44
age 6 ans	0.0	*référence*		

logarithme de vraisemblance : $Log(LM2) = 201.4059$

La constante n'apporte pas d'information sur les effets testés, mais elle est indispensable pour calculer les estimations des effectifs, ce qui n'est pas l'objectif principal.

Le coefficient des *non* distingue dans le calcul des effectifs la part des *non* par rapport aux *oui* tels qu'ils se présentent dans l'échantillon, *indépendamment* des classes d'âge. Ce nombre peut se calculer directement à partir du tableau initial en lisant la distribution marginale :

$$Log\frac{59}{63} = -0.136$$

Il en est de même pour les classes d'âge :

$$Log\frac{35}{33} = 0.059$$

Le **troisième modèle M3** est très différent du modèle logistique puisque la linéarité, ou l'homogénéïté introduite par la variable *âge* porte sur le *logarithme des effectifs* ou des probabilités, alors que pour le modèle logistique la linéarité porte sur le logarithme des odds (chances, cotes).

C'est une différence importante entre une modélisation d'un odds ratio et celle d'un rapport des risques ; une linéarité dans l'une ne correspond pas une linéarité dans l'autre.

Les modèles précédents, appelés *loglinéaires,* quand ils sont appliqués à des données exclusivement catégorielles, peuvent être traités indifféremment dans le cadre d'une épreuve où le contenu d'une cellule $i, j, k...$ suit la loi de Poisson ou la loi multinomiale. Les estimations sont les mêmes au maximum de vraisemblance. Cela vient de ce qu'une loi de Poisson conditionnée par le nombre total $\sum n_{ijk...} = n$ est une loi multinomiale.

Programmes SAS associé :

```
proc genmod data=perenoel;
class reponse age;
model effectif=reponse*age/dist=poi link=log type3 obstat;
run;
proc genmod data=perenoel;
class reponse age;
model effectif=reponse age/dist=poi link=log type3 obstats;
run;
proc genmod data=perenoel;
class reponse ;
model effectif=reponse*age/dist=poi link=log type3 obstats;
run;
```

5 Exemple 2 : accueil des étudiants

La population est un ensemble d'étudiants classés par leur appartenance à un cycle d'études (première ou seconde année) et par la qualité de l'accueil qu'ils disent avoir trouvé dans leur démarches administratives. Nous avons finalement retenu trois classes pour la qualité de l'accueil. On évite ainsi les trop petits nombres de la classe *très bon* qui a été regroupée avec la classe *bon*.

Pris globalement, le tableau croisé des effectifs a déjà montré que les variables de classement n'étaient pas indépendantes. On peut mesurer précisément l'écart à l'indépendance dans chaque cellule par la relation :

$$dev_{ij} = \frac{(n_{ij} - m_{ij})^2}{m_{ij}} \quad avec \quad m_{ij} = \frac{n_{i+}n_{+j}}{n_{++}}$$

		données brutes	
accueil ▼	année 1	année 2	total
bon	22	44	66
moyen	122	77	199
mauvais	74	31	105
total	218	153	370

écarts à l'indépendance		
accueil ▼	année 1	année 2
bon	-16.9	16.9
moyen	4.75	-4.75
mauvais	12.1	-12.1

Les écarts à l'indépendance ne sont pas liés de façon simple aux effectifs ; il est difficile de bien comprendre ce qu'ils signifient. Il reste que certains d'entre eux, plus importants, révèlent une hétérogénéité qu'il peut être intéressant de quantifier.

Dans cet exemple, plusieurs points de vue vont se succéder :

- un secrétariat reçoit des étudiants qu'on classe a priori en trois classes (les mécontents, les normaux, les euphoriques) et on se demande si dans chaque classe ils sont plutôt dans l'année 1 ou l'année 2. Rappelons qu'ici les deux années ne sont pas liées, qu'il ne s'agit pas d'évolution des personnes en cause : c'est une étude *rétrospective*. La réponse aléatoire est alors l'*année*, la variable de classification est la qualité de l'entretien, vue par l'administration. On comparera alors les distributions en ligne.

- à l'inverse on peut considérer que les personnels de l'administration sont d'humeur ou d'efficacité différentes, peut-être parce que ce ne sont pas les mêmes qui s'occupent des deux années ; les étudiants appartiennent à deux populations distinctes. La réponse aléatoire est alors la nature même de l'entretien, de l'*accueil* ; et l'année sert simplement à classer les étudiants (ou les types de secrétariat). On comparera les distributions en colonne.

- On peut encore considérer que les 370 personnes interrogées sont simplement classées suivant deux critères (accueil, année), et on s'intéresse alors à la répartition globale pour éventuellement trouver des explications ou du moins une analyse de l'ambiance générale. On traitera la répartition globale des 370 opinions exprimées.

Programme SAS associé :

```
data etudiants;
input accueil $1-8 annee effectif ;
datalines;
mauvais 1 74
moyen 1 122
bon 1 18
très bon 1 4
mauvais 2 31
moyen 2 77
bon 2 43
très bon 2 1
;
run;
proc freq data=etudiants;
weight effectif;
tables accueil*annee
    /norow nopercent expected deviation chisq;
run;
proc format; value $regroup
'bon'='bon, très bon' 'très bon'='bon, très bon';
run;
proc freq data=etudiants;
format accueil $regroup.  ;
weight effectif;
tables accueil*annee
    /norow nopercent expected deviation chisq;
run;
```

5.1 Modèle 1 : logistique

Il s'agit d'examiner, pour chaque qualité d'accueil, d'examiner s'il s'agit plutôt d'étudiants de l'*année 1* que de l'*année 2*. La réponse est l'*année*, elle est binaire. Le modèle logistique s'écrira :

$$Log\frac{p_i}{1-p_i} = \lambda + \lambda_i^{ac}$$

où p_i désigne la probabilité pour qu'il s'agisse d'un étudiant de première année,

$q_i = 1 - p_i$ la probabilité pour qu'il s'agisse d'un étudiant de seconde année.
i désigne la qualité de l'accueil
avec l'une des trois contraintes suivantes :

$$\begin{aligned} \lambda_1^{ac} + \lambda_2^{ac} + \lambda_3^{ac} &= 0 \\ ou\ \lambda_1^{ac} &= -\lambda_2^{ac} - \lambda_3^{ac} \end{aligned} \tag{C1}$$

$$\lambda_1^{ac} = 0 \tag{C2}$$

$$\lambda = 0 \tag{C3}$$

Le modèle s'écrit de façon équivalent :

$$Log\frac{q_i}{1-q_i} = Log\frac{1-p_i}{p_i} = -\lambda - \lambda_i^{ac}$$

Il n'y a qu'un changement de signe des coefficients, les tests et les ajustements sont les mêmes.

Résultats avec la contrainte C1 :

Accueil ▼	coeff	estimation	écart-type	test de Wald	p-value
constante	λ	0.212	0.123	3.00	0.083
mauvais (mécontents)	λ_3^{ac}	0.658	0.174	14.28	0.0002
moyen	λ_2^{ac}	0.248	0.149	2.784	0.005
bon (euphoriques)	λ_1^{ac}	-0.906	*= - 248 - 0.658 (contrainte)*		

La forme des résultats privilégie la mesure de la détérioration de l'accueil de première en deuxième année, mesurée par des odds ratios :

accueil ▼	odds ratio	estimation	intervalle de confiance 95%	
mauvais/bon	$\exp(\lambda_3^{ac} - \lambda_1^{ac})$	4.78	2.46	9.25
moyen/bon	$\exp(\lambda_2^{ac} - \lambda_1^{ac})$	3.17	1.76	5.69

La nullité d'un coefficient λ_i^{ac} signifie que, pour la catégorie i, les chances d'une année sur l'autre sont les mêmes dans cette catégorie que pour la moyenne (la répartition marginale) ; ce n'est donc pas elle qui "fait la différence". L'égalité à 1 d'un odds ratio (égalité de deux coefficients) signifie que les deux sous-populations se comportent de la même façon de la première à la deuxième année.

Les tests de Wald, comme les intervalles de confiance des odds ratios, assurent que cette détérioration est statistiquement significative.

Programme SAS associé :

```
proc logistic data=etudiants;
format accueil $regroup.  ;
weight effectif;
class accueil (ref='bon, très bon');
model annee=accueil;
run;
```

Résultats avec la contrainte C2 :

Accueil▼	estimation	écart-type	test de Wald	p-value
constante	-0.693	0.2611	7.05	0.008
mauvais	1.563	0.338	21.44	0.0001
moyen	1.153	0.299	14.89	0.0001
bon	0.0	*référence*		

Les tests de nullité du coefficient de *mauvais* correspond à l'hypothèse d'égalité des répartitions entre première et deuxième année des deux catégories : *mauvais accueil* et *bon accueil* (référence).Le rejet de cette hypothèse signifie que les étudiants "pénibles" (accueil mauvais) sont plus probablement en première année que ne le sont les étudiants "agréables" (accueil bon). *Il faut garder à l'esprit que c'est l'année de l'étudiant, se présentant dans un secrétariat, qui est aléatoire.*

Les chances de recevoir un étudiant mécontent (plutôt qu'euphorique) de première année est $exp(1.563) = 4.8$ fois celles d'un étudiant de deuxième année. Ce qui inciterait à croire que l'université sélectionne les étudiants les plus aimables ou les moins exigeants.

Les odds ratios sont les mêmes que dans le modèle précédent ainsi que la vraisemblance et le test global de Wald.

Programme SAS associé :

```
proc logistic data=etudiants;
format accueil $regroup.  ;
weight effectif;
class accueil /descending param=glm;
model annee=accueil ;
run;
```

Résultats avec la contrainte C3 :

Accueil▼	estimation	écart-type	test de Wald	p-value
bon	-0.693	0.261	7.05	0.008
moyen	0.460	0.146	1.00	0.0001
mauvais	0.870	0.214	16.54	0.002

Les trois variables indicatrices de l'accueil sont orthogonales, leurs corrélation est nulle. Il est facile de retrouver tous les résultats précédents, coefficients estimés et écarts-types (les estimateurs ne sont pas corrélés). La somme des trois statistiques de Wald (33.58) est égale à la statistique globale de Wald des modèles C1, C2.

Cette forme de modèle n'est pas intéressante pour les tests des coefficients. La nullité du cofficient *moyen,* par exemple, signifie seulement qu'il y a le même nombre d'étudiants en première année qu'en deuxième année dans la catégorie *moyen* $p_2 = 1 - p_2 = 0.5$, ce qui ne suffit pas pour comparer la détérioration ou non de l'accueil

Pour atteindre l'objectif poursuivi (comparer l'accueil des étudiants d'une année sur l'autre) on doit revenir aux odds ratios qui sont les mêmes que dans les formulations C1,C2.

Programme SAS associé :

```
proc logistic data=etudiants;
format accueil $regroup.  ;
weight effectif;
class accueil /param=glm;
model annee=accueil /noint;
run;
```

5.2 Modèle 2 : logistique généralisée aux réponses polytomiques

Le second point de vue est celui de l'étudiant qui trouve un accueil aléatoire et cet accueil ne suit pas nécessairement la même loi suivant qu'il s'agit d'un étudiant en première ou en deuxième année. La réponse est ternaire (qualité de l'accueil) ; on admet qu'elle suit une loi multinomiale de paramètres $(p_{11}, p_{21}, p_{31}, n_1)$ pour la première année et $(p_{12}, p_{22}, p_{32}, n_2)$ pour la seconde année, que les deux lois sont indépendantes.

La fonction de lien qui définit le paramètrage des p_{ij} peut prendre, entre autres, l'une des formes suivantes :

$$Log\frac{p_{ij}}{1-p_{ij}} = \lambda_i^{ac} + \lambda_{ij}^{an}$$

$$Log\frac{p_{ij}}{p_{0j}} = \lambda_i^{ac} + \lambda_{ij}^{an}$$

$$Log\frac{1-\sum_{k=1}^{k=i} p_{kj}}{\sum_{k=1}^{k=i} p_{kj}} = \lambda_i^{ac} + \lambda_{ij}^{an}$$

avec les contraintes :

$$\forall i \ \sum_j \lambda_{ij}^{an} = 0$$

La première forme consiste à analyser chaque éventualité séparément, et donc elle n'assure pas que la somme des probabilités estimées soit égale à 1, ce qui est un peu génant. On peut introduire cette dernière contrainte, mais la maximisation de la vraisemblance n'est plus aussi facile. Généralement cette formulation n'est pas retenue.

La seconde forme est une extension simple du cas binaire, en considérant pour chaque multinomiale une catégorie de référence (accueil *bon*, par exemple, désigné par l'indice 0). On la désigne par le terme "logistique généralisée".

La troisième forme, qui consiste à travailler sur les probabilités *cumulées*, est aussi facile à implémenter, mais elle n'a de sens que si les modalités (le type d'accueil) sont ordonnées.

Nous allons examiner les deux dernières formes et nous reprendrons dans le modèle suivant (loglinéaire) une paramétrisation plus générale qui permettra de voir les contraintes sous-jacentes à chaque modèle.

5.2.1 Modèle logistique généralisé

Deux lois multinomiales de paramètres $(p_{11}, p_{21}, p_{31}, 218)$ pour la classe d'étudiants *année 1* et $(p_{12}, p_{22}, p_{32}, 152)$ pour la classe *année 2*. Comme les sommes des p_{i1} et des p_{i2} sont égales à 1, il reste donc quatre paramètres à estimer. Le modèle retenu est donc saturé.

$$Log\frac{p_{ij}}{p_{0j}} = \lambda_i^{ac} + \lambda_{ij}^{acan}$$

L'indice i désigne le type d'accueil, 0 l'accueil *bon* et j désigne l'année.

désignation de l'accueil	coefficient	estimation	écart-type
mauvais (global))	λ_1^{ac}	0.431	0.169
moyen (global)	λ_2^{ac}	1.136	0.150
mauvais (année 1)	λ_{11}^{acan}	0.782	0.169
moyen (année 1)	λ_{21}^{acan}	0.577	0.150
mauvais (année 2)	λ_{12}^{acan}	-0.782	0.169
mauvais (année 2)	λ_{22}^{acan}	-0.577	0.150

On en déduira par exemple :

$$\begin{aligned} Log\frac{\Pr(mauvais\,/\,année\ 1)}{\Pr(bon\,/\,année\ 1)} &= 0.431 + 0.782 = 1.213 \\ d'où\ les\ chances\ (odds) &= \exp(1.213) = 3.4 \end{aligned}$$

$$\begin{aligned} Log\frac{\Pr(mauvais\,/\,année\ 2)}{\Pr(bon\,/\,année\ 2)} &= 0.431 - 0.782 = -0.351 \\ d'où\ les\ chances\ (odds) &= \exp(-0.351) = 0.7 \end{aligned}$$

Un étudiant de première année a 7 chances contre deux d'avoir un mauvais accueil plutôt qu'un bon accueil (3.4 contre 1), alors qu'un étudiant de deuxième année a près de trois chances contre deux (0.7 contre 1). On peut en déduire le rapport des chances (odds ratio) qui mesure une amélioration de l'accueil de la première à la deuxième année : $3.4/0.7 = 4.8$.

Ce qui inciterait à croire que l'université adoucit les moeurs.

Les odds ratios ne dépendent pas du protocole d'expérimentation, ils ne parlent que de probabilités conditionnelles. Mais pour les expliquer (les interpréter) la façon dont l'échantillon a été construit, et plus précisément le choix des variables de réponse et de classement qui définissent les lois de probabilités du modèle, sont primordiales.

Le modèle étant saturé, ce chiffre peut être immédiatement calculé à partir du tableau d'origine, ainsi d'ailleurs que tous les coefficients du modèle :

$$\frac{74/22}{31/44} = 4.8$$

Remarquons aussi que les deux multinomiales auraient pu être estimées séparément pour chaque année. Mais les traiter simultanément peut servir à faire des comparaisons de coefficients pour deux années différentes.

L'intérêt d'un modèle saturé n'est pas de retrouver les résultats qu'on pouvait extraire du tableau croisé initial mais d'obtenir au niveau global une log-vraisemblance qui servira à comparer d'autres modèles (déviance). De plus des tests sur les coefficients et leurs combinaisons linéaires conduisent à accepter ou rejeter des simplifications dans le modèle.

Par exemple : **le changement d'année est-il significatif d'une amélioration de l'accueil ?**

Il faut tester les contraintes suivantes :

"mauvais accueil équivalent" : $\lambda_{11}^{acan} = \lambda_{12}^{acan}$ (ou $\lambda_{11}^{acan} = 0$)

"moyen accueil équivalent".: $\lambda_{21}^{acan} = \lambda_{22}^{acan}$ (ou $\lambda_{21}^{acan} = 0$)

Les tests généralement proposés par défaut dans les logiciels correspondent à des hypothèses nulles et sont *globaux* au sens où on s'intéresse en premier lieu à la recherche des classifications qui apportent de l'information. Ainsi l'hypothèse "l'année ne joue aucun rôle dans l'amélioration de l'accueil" doit être rejetée car la statistique du χ^2 à 2 degrés de liberté (calculée par le logiciel) vaut 22.4, ce qui correspond à une p-value inférieure à 0.0001.

Programme SAS associé :

```
proc logistic data=etudiants order=data;
weight effectif;
format accueil $regroup.  ;
class annee;
model accueil=annee /link=glogit;
run;
```

(forme équivalente et tests)

```
proc catmod data=etudiants order=data;
response logits;
weight effectif;
format accueil $regroup.  ;
model accueil=annee/pred=prob;
contrast 'annee 1=annee 2' all_parms 0 0 1 0,all_parms 0 0 0 1 ;
contrast 'accueil aussi mauvais en annee 1 et 2' all_parms 0 0 1 -1 ;
contrast 'mauvais et moyen accueil fusionnés' all_parms 1 -1 0 0 ;
contrast 'moyen et bon accueil fusionnés' all_parms 1 2 0 0;
run;
```

5.2.2 Modèle logistique cumulé, avec réponse ordinale

Comparer les accueils *mauvais, moyen, bon*, deux à deux, n'est peut-être pas indiqué si les modalités sont ordonnées. Dans ce cas, il préférable de tenir compte de cet ordre et il est souvent proposé une autre forme de modèle qui ne se déduit pas des précédents et qui ne donne pas les mêmes résultats sur les ajustements si l'échantillon n'est pas très grand :

L'estimation portera sur les probabilités cumulées, ce qui peut s'écrire dans ce cas particulier sous la forme :

$$Log\frac{1-p_{1j}}{p_{1j}} = Log\frac{p_{2j}+p_{3j}}{p_{1j}} = \lambda_1^{ac} + \lambda_{1j}^{acan}$$

$$Log\frac{1-(p_{1j}+p_{2j})}{p_{1j}+p_{2j}} = Log\frac{p_{3j}}{p_{1j}+p_{2j}} = \lambda_2^{ac} + \lambda_{2j}^{acan}$$

Les coefficients λ_1^{ac} et λ_2^{ac} mesurent les effets globaux de l'accueil en décomposant les modalités en deux groupes, de façon différente

L'indice j caractérise l'effet de l'année j avec les contraintes :

$$\lambda_{11}^{acan} + \lambda_{12}^{acan} = 0 \ et \ \lambda_{21}^{acan} + \lambda_{22}^{acan}$$

Le modèle est saturé comme le précédent. Dans le cas général on n'obtient pas les mêmes résultats avec les deux modèles et dans ce dernier modèle les tests associés aux coefficients ne sont pas tout à fait les mêmes puisqu'il portent sur la significativité d'un *cumul de modalités* de réponse. Si on écrit le modèle logistique généralisé sous la forme équivalente loglinéaire, les effets additifs sur le logarithme des effectifs ne correspondent pas à des effets additifs sur les logarithmes des effectifs cumulés.

désignation de l'accueil	coefficient	estimation	écart-type
moyen + bon / mauvais (global))	λ_1^{ac}	1.014	0.124
bon / moyen + mauvais (global)	λ_2^{ac}	-1.543	0.144
moyen + bon / mauvais (année 1)	λ_{11}^{acan}	-0.348	0.124
bon / moyen + mauvais (année 1)	λ_{21}^{acan}	-0.645	0.144
moyen +bon / mauvais (année 2)	λ_{12}^{acan}	0.348	0.124
bon / moyen + mauvais (année 2)	λ_{22}^{acan}	0.645	0.144

On en déduira par exemple :

$$Log\frac{\Pr(bon\ /\ année\ 1)}{\Pr(mauvais\ ou\ moyen\ /\ année\ 1)} = -1.543 - 0.645 = -2.19$$

$$\begin{aligned} Log\frac{\Pr(mauvais\ ou\ moyen/\ année\ 1)}{\Pr(bon\ /\ année\ 1)} &= 2.19 \\ d'où\ les\ chances\ (odds) &= \exp(2.19) = 8.9 \end{aligned}$$

$$Log\frac{\Pr(bon\ /\ année\ 2)}{\Pr(mauvais\ ou\ moyen\ /\ année\ 2)} = -1.543 + 0.645 = -0.90$$

$$\begin{aligned} Log\frac{\Pr(mauvais\ ou\ moyen\ /\ année\ 2)}{\Pr(bon\ /\ année\ 2)} &= 0.90 \\ d'où\ les\ chances\ (odds) &= \exp(0.90) = 7.9 \end{aligned}$$

Un étudiant de première année a neuf chances contre une d'avoir un mauvais (ou moyen) accueil plutôt qu'un bon accueil, alors qu'un étudiant de deuxième année a huit chance contre une. On peut en déduire le rapport des chances (odds ratio) qui mesure une amélioration de l'accueil de la première à la deuxième année : $8.9/7.9 = 1.1$.

Ce nombre se retrouve à partir du tableau croisé initial :

$$\frac{22/196}{44/108} = 1.1$$

Quand on oppose *bon* à tous les autres, l'amélioration est faible. La différence avec le modèle précédent s'explique par un glissement des accueils *mauvais* dans la classe des accueils *moyens* quand on passe dela première année à la deuxième année, sans que la proportion des accueils *bons* ne change beaucoup.

On voit dans cet exemple l'importance de la présentation des résultats, résultats qui dans ce cas particulier de modèle saturé sont identiques.

Il semble donc que le découpage en classes et les ambiguïtés qu'il engendre (que signifie accueil *moyen*?) soit un point sensible de la modélisation. Avec les modèles *à odds ratios proportionnels* ces ambiguïtés seront évitées.

Programme SAS associé :

```
proc catmod data=etudiants order=data;
response clogits;
weight effectif;
format accueil $regroup.  ;
model accueil=annee/pred=prob;
run;
```

Autre forme équivalente où la fonction de lien s'écrit :

$$\lambda_j^{an} + \lambda_{ij}^{acan}$$

au lieu de :

$$\lambda_i^{ac} + \lambda_{ij}^{acan}$$

```
proc catmod data=etudiants order=data;
weight effectif;
response clogits;
format accueil $regroup.  ;
model accueil=annee _response_(annee)/pred=prob;
run;
```

5.2.3 Une variante avec odds ratios proportionnels

Les *modalités sont ordonnées*, la fonction de lien concerne les *probabilités cumulées* ; mais il y a une propriété particulière de *proportionnalité de odds ratios* quand on compare deux réponses i, i', qui s'exprime sous la forme :

$$\forall j, l \text{ le rapport } \frac{1-\sum_{k=1}^{k=i} p_{kj}}{\sum_{k=1}^{k=i} p_{kj}} / \frac{1-\sum_{k=1}^{k=i} p_{kl}}{\sum_{k=1}^{k=i} p_{klj}} \textbf{ ne dépend pas de i}$$

Ou plus simplement par le modèle :

$$Log \frac{1-\sum_{k=1}^{k=i} p_{kj}}{\sum_{k=1}^{k=i} p_{kj}} = \lambda_i^{ac} + \lambda_j^{an}$$

Ce modèle a donc 3 paramètres indépendants et se déduit des précédents en ajoutant *une contrainte d'égalité* des λ_{ij}^{an} de même indice i.

La propriété de proportionnalité reste vraie par regroupement des classes de réponses, de sorte que si la réponse qualitative est issue d'un regroupement de classes, ou d'un découpage en tranches, la proportionnalité est conservée pour les différentes versions, et les résultats (effets des explicatives) sont les mêmes.

Ce modèle est important dans les applications où la réponse est une discrétisation d'une variable quantitative et qu'on souhaite que les résultats ne soient pas trop sensibles au découpage. A tel point que, par défaut, certaines procédures supposent a priori que cette hypothèse est réalisée. Il faudrait au moins vérifier que cette hypothèse est acceptable (il existe plusieurs tests disponibles).

On peut par exemple comparer le dernier modèle estimé avec celui qui admet la proportionnalité des odds ratios et dont les résultats suivent:

désignation de l'accueil (relatif)	coefficient	estimation	écart-type
moyen ou bon / mauvais (global))	λ_1^{ac}	-1.072	0.123
bon / moyen ou mauvais (global)	λ_2^{ac}	1.499	0.137
année 1	λ_1^{an}	-0.475	0.106
année 2	λ_2^{an}	0.475	0.106
test de proportionnalité : $\chi^2 = 3.4$ p-value : 0.065			

La proportionnalité n'est pas rejetée (avec un risque α de première espèce de 5%). Les coefficients sont cohérents avec les précédents.

Programme SAS associé :

```
proc logistic data=etudiants order=data;
format accueil $regroup.  ;
weight effectif;
class annee;
model accueil=annee;/* option par défaut clogit) */
output out=s predprobs=i;
run;
proc print data=s;run;
```

Autre forme avec une écriture équivalente mais des contraintes différentes :

$$Log\frac{1-\sum_{k=1}^{k=i} p_{kj}}{\sum_{k=1}^{k=i} p_{kj}} = \lambda + \lambda_i^{ac} + \lambda_j^{an} \quad \sum_i \lambda_i^{ac} = 0 \quad \sum_j \lambda_j^{an} = 0$$

```
proc catmod data=etudiants order=data;
weight effectif;
response clogits;
format accueil $regroup.  ;
model accueil=_response_ annee/pred=prob;
run;
```

5.3 Modèle 3 : loglinéaire

Le modèle loglinéaire (de Poisson) sur les effectifs du tableau croisé, s'écrit souvent conditionnellement à l'effectif total. Les probabilités de chaque classe (loi multinomiale) ont pour fonction de lien dans la forme la plus générale (modèle saturé) :

$$Log(p_{ij)} = \lambda_i^{ac} + \lambda_j^{an} + \lambda_{ij}^{acan}$$

$$\sum_i \lambda_i^{ac} = 0, \ \sum_j \lambda_j^{an} = 0, \ \forall j \ \sum_i \lambda_{ij}^{acan} = 0, \ \forall i \ \sum_j \lambda_{ij}^{acan} = 0$$

Ce qui laisse 5 paramètres libres (pour 6 réponses dont la somme est 1).

Une autre formulation du modèle saturé avec les logarithmes des effectifs (au lieu des probabilités) comporte 6 paramètres :

$$Log(n_{ij}) = \lambda + \lambda_i^{ac} + \lambda_j^{an} + \lambda_{ij}^{acan}$$

et les mêmes contraintes que le précédent.

Ce modèle, considéré comme le plus général pour décrire un tableau de contingence, permet de retouver la plupart des modèles précédents (logistiques généralisés), que ce soient les odds ratios ou les tests.

Il permet surtout de construire des tests plus facilement, et en particulier quand il y a beaucoup de variables de classement. Dans le cas présent, on vérifiera seulement la cohérence des résultats.

Dans le modèle saturé, le test de nullité des interactions *année*×*accueil* ($\forall i, j\ \lambda_{ij}^{acan} = 0$) est le même que celui des modèles C1, C2 sur l'effet *accueil* (22.42 pour un χ^2 à 2 degrés de libertés). C'est aussi le même que dans le modèle 2 (logistique polytomique) sur l'effet de *année*.

effet	coefficient	estimation
effet principal accueil	λ_1^{ac}	-0.0912
	λ_2^{ac}	0.6137
	λ_3^{ac}	-0.5225
effet principal année	λ_1^{an}	0.1062
	λ_2^{an}	-0.1062
interaction année 1 - accueil	λ_{11}^{acan}	0.3288
	λ_{12}^{acan}	0.1239
	λ_{13}^{acan}	-0.4527
interaction annee 2 - accueil	λ_{21}^{acan}	-0.3288
	λ_{22}^{acan}	-0.1239
	λ_{23}^{acan}	0.4527

A partir de ce modèle saturé on retrouve les résultats de la logistique généralisée du modèle 2 qui porte sur les probabilités des modalités (non cumulées) :

$$Log\frac{\Pr(mauvais\ /\ année\ 1)}{\Pr(bon\ /\ année\ 1)} = (-0.1062 - 0.0912 + 0.3288) - (-0.1062 - 0.5225 - 0.4527) = 1.21$$

$$Log\frac{\Pr(mauvais\ /\ année\ 2)}{\Pr(bon\ /\ année\ 2)} = (0.1062 - 0.0912 - 0.3288) - (0.1062 - 0.5225 + 0.4527) = -0.35$$

Le test de "non-interaction" (appelé aussi d'indépendance des deux effets) correspond à l'annulation des coefficients λ_{ij}^{acan}.

La statistique du rapport de vraisemblance donne 22.42 et celle de Wald 21.73 pour des χ^2 à 2 degrés de liberté avec une p-value inférieure à 0.0001. La statistique 22.42 avait déjà été trouvée dans le modèle 2 comme déviance (écart au modèle saturé).

D'autres tests peuvent être demandés et ils s'expriment bien sous cette forme loglinéaire.

Par exemple la fusion des catégories *moyen* et *mauvais* (29.81 pour un χ^2 à 2 degrés de liberté) et la fusion des catégories *moyen* et *bon* (63.47 pour un χ^2 à 2 degrés de liberté) seront rejetées.

Le modèle linéaire limité aux seuls effets principaux se résume à :

$$Log(p_{ij}) = \lambda_i^{ac} + \lambda_j^{an}$$

avec les contraintes :

$$\sum_i \lambda_i^{ac} = 0 \quad \sum_j \lambda_j^{an} = 0$$

il y a 3 paramètres libres.

Il est formulé de telle façon que les écarts, entre les probabilités observées dans le tableau de contingence et les probabilités estimées, s'expriment globalement par la déviance, et de façon identique par la statistique, $G^2 = 24.04$ calculée directement à partir du tableau initial. Cette statistique se retrouve aussi dans les logistiques élémentaires (accueil en fonction de l'année, et année en fonction de l'accueil).

Programme SAS associé :

```
proc catmod data=etudiants order=data;
weight effectif;
format accueil $regroup.  ;
model accueil*annee=_response_ / pred=freq;
loglin annee|accueil;
contrast 'pas d interactions'
    annee*accueil 1 1;
contrast 'fusion moyen mauvais'
    annee*accueil -1 1, accueil -1 1 ;
contrast 'fusion bon moyen'
     annee*accueil 1 2, accueil 1 2;
run;
proc catmod data=etudiants order=data;
weight effectif;
format accueil $regroup.  ;
model accueil*annee=_response_ / pred=freq;
loglin annee accueil;
run;
```

La symétrie des variables, et donc des associations possibles, rend le modèle loglinéaire intéressant pour étudier des associations multiples. Il est donc plus général que le modèle logistique, sauf en ce qui concerne les particularités dues aux *modalités ordinales* qui ne sont pas prises en compte. D'autre part la proportionalité des odds ratios peut être testée mais elle implique des contraintes sur les coefficients qui ne sont pas nécessairement prévues dans les logiciels.

6 Exemple 3 : Une promotion de diplomés

Dans cet exemple, les variables sont binaires, donc simples à analyser, mais le tableau de contingence a trois dimensions, ce qui permet d'aborder des aspects plus complexes que ceux des exemples précédents. La simplicité des variables facilite les calculs, mais ne restreint pas la portée des notions qui vont être introduites. D'autre part la faiblesse des effectifs de l'échantillon rend les résultats peu significatifs ; on s'intéressera plutôt aux estimations qu'aux intervalles de confiance, et les tests seront présentés pour l'intérêt des hypothèses testées, bien que les conditions d'application soient souvent insuffisantes pour que leurs propriétés asymptotiques soient assurées.

Nous commencerons par signaler l'ambiguïté des résultats bruts obtenus dans le tableau initial, puis nous examinerons successivement un modèle logistique et un modèle loglinéaire.

6.1 Une lecture rapide du tableau initial

Avant tout calcul statistique, nous reprenons la présentation du tableau initial en admettant que la variable de réponse soit la note, réduite à deux catégories. La population est classée de deux façons, par région et par filière d'origine

	Paris		Province		
filière ▶	Economie	Mass-Mst	Economie	Mass-Mst	total
note < 12	**3**	**2**	**5**	**1**	11
note ≥ 12	**6**	**4**	**8**	**4**	22
total	9	6	13	5	33

que nous comparons à un regroupement Paris-Province

	Paris - Province		
filière ▶	Economie	Mass-Mst	total
note < 12	**8**	**3**	11
note ≥ 12	**14**	**8**	22
total	22	11	33

Calculons directement les odds ratios qui répondent à la question :

les chances d'avoir une note élevée contre une note faible sont-elles les mêmes, si les étudiants viennent de la filière Economie ou s'ils viennent de la filière Mass-Mst, et dans quel rapport ?

- 1 - région Paris : 6 contre 3, et 4 contre 2, d'où :

$$OR_1 = \frac{6/3}{4/2} = 1$$

- 2 - région Province : 8 contre 5, et 4 contre 1, d'où :

$$OR_2 = \frac{8/5}{4/1} = 0.4$$

- 1+2 - région Paris-Province : 14 contre 8, et 8 contre 3 d'où :

$$OR_3 = \frac{14/8}{8/3} = 0.66$$

La conjonction des deux régions donne un odds ratio intermédiaire entre les odds ratios des deux régions. Mais *ce n'est pas automatique*.

On sait que la moyenne d'une variable dans une population est nécessairement située entre les moyennes de la variable dans les deux sous-populations qui la composent. Mais *si la structure des sous-populations évolue*, la moyenne sur la population totale ne *varie* pas nécessairement comme celle de chaque sous-population, on parle alors d'effet de structure. C'est vrai aussi pour l'évolution des chances et un odds ratio calculé sur la répartition marginal peut être très différent de ceux qui sont calculés sur des sous-populations, comme le montre dans l'exemple suivant :

	Paris		Province		
filière ▶	Economie	Mass-Mst	Economie	Mass-Mst	total
note < 12	**1**	**4**	**4**	**2**	11
note ≥ 12	**5**	**12**	**4**	**1**	22
total	6	16	8	3	33

	Paris - Province		
filière ▶	Economie	Mass-Mst	total
note < 12	**5**	**6**	11
note ≥ 12	**9**	**13**	22
total	14	19	33

- 1 - région Paris : 5 contre 1 et 12 contre 4, d'où :

$$OR_1 = \frac{5/1}{12/4} = 1.7$$

- 2 - région Province : 4 contre 4 et 1 contre 2, d'où :

$$OR_2 = \frac{4/4}{1/2} = 2$$

- 1+2 - région Paris + Province : 9 contre 5 et 13 contre 6 d'où :

$$OR_3 = \frac{9/5}{13/6} = 0.83$$

Ce paradoxe, appelé *paradoxe de Simpson*, nous rappelle qu'une distribution marginale (ensemble des deux sous-populations) peut être très différente des distributions conditionnelles ou "partielles" (de chaque sous-population). Il sera donc intéressant de tester si les distributions marginales et partielles sont les mêmes : test de fusion de sous-populations ("collapsibility") ; les associations ou indépendances entre variables devront être distinguées selon qu'elles sont conditionnelles ou non.

On dira par exemple qu'il y a *indépendance conditionnelle* si les odds ratios sont égaux à 1 (dans chaque région, Paris comme province) ce qui n'implique pas qu'il y ait *indépendance marginale*. Si les odds ratios (Paris et Province) sont égaux mais pas nécessairement égaux à 1 on dit qu'il y a une *association homogène*.

Nous allons retrouver ces questions dans un cadre plus facile à généraliser (sans limite du nombre de variables) avec les modèles logistique et loglinéaire.

6.2 Logistique avec plusieurs variables de classement

Au lieu de partir de l'expérimentation et des objectifs pour construire un modèle, ce que nous avons déjà développé dans la section "Premiers pas" et dans le paragraphe précédent, nous ne mettrons pas en cause le modèle logistique mais nous analyserons les différents modèles pour voir comment se hiérarchisent ces modèles, ce que chacun apporte, les contraintes, les hypothèses sous-jacentes à ces contraintes. Nous retrouverons ainsi la plupart des tests classiques qui sont utilisés pour d'autres modèles, plus complexes par le nombre des variables et par le nombre de modalités de chacune d'entre elles.

Le modèle le plus général est saturé. La variable note (≥ 12) suit une loi binomiale dont le paramètre p_j dépend de la filière ou diplome à l'entrée F_i (Economie, Mass-Mst) et de l'académie R_j d'où vient l'étudiant (Paris, Province).

$$Logit(p_{ij}) = Log\frac{p_{ij}}{1-p_{ij}} = \lambda + \lambda_i^F + \lambda_j^R + \lambda_{ij}^{FR}$$

avec les contraintes :

$$\sum_i \lambda_i^F = 0 \ , \ \sum_j \lambda_j^R = 0 \ , \forall j \ \sum_i \lambda_{ij}^{FR} = 0 \ , \forall i \ \sum_j \lambda_{ij}^{FR} = 0$$

Il y a 4 probabilités p_{ij} à estimer, à partir de 4 paramètres indépendants.

Ce modèle compare et analyse quatre distributions (colonnes du tableau initial).

λ est un coefficient "technique" qui est simplement la moyenne des logarithmes des chances dans les quatre sous-populations, il sert de référence pour distinguer la part propre à chacun des autres effets :

$$\lambda = \frac{1}{4}\sum_{ij} Log\frac{p_{ij}}{1-p_{ij}} = -0.8106$$

Il ne correspond donc pas à la répartition marginale (absence de tout effet) qui s'exprimerait par :

$$\lambda_0 = Log\frac{\sum n_{ij}p_{ij}}{\sum n_{ij}(1-p_{ij})} = Log\frac{11}{22} = -0.6931$$

Les coefficients λ_i^F et λ_j^R sont appelés effets principaux, les λ_{ij}^{FR} sont les interactions entre filières et régions.

De la vraisemblance pour le modèle avec constante (distribution marginale)

$$-2Log(L_0) = 42.010$$

et de celle du modèle saturé

$$-2Log(L_S) = 41.423$$

on déduit la déviance par la différence 0.587 (loi du χ^2 à 3 degrés de liberté).

Le modèle suivant, sans interaction, est *à associations homogènes* ou *à odds ratios homogènes.*

$$\begin{aligned} Log\frac{p_{ij}}{1-pij} &= \lambda + \lambda_i^F + \lambda_j^R \\ \sum_i \lambda_i^F &= 0\ ,\ \sum_j \lambda_j^R = 0 \end{aligned}$$

Ce qui veut dire que l'association de la filière et de la note est la même pour les régions (homogénéïté dans les régions) et l'association de la note et de la région est la même pour les filières (homogénéïté dans les filières).

La nullité des interactions est testée par le test de Wald dans le modèle saturé (0.297), ou par la déviance (écart au modèle saturé): $41.727 - 41.423 = 0.304$.

Cette statistique se calcule directement sur les tableaux de contingence $2 \times 2 \times K$, elle est appelée *statistique de Breslow-Day.*

Un modèle plus complexe avec trois variables de classement de la forme

$$Log\frac{p_{ijk}}{1-p_{ijk}} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB} + \lambda_k^C$$

est dit *à association homogène dans les sous-populations induites par la variable C.*

Le modèle qui ne comporte qu'une variable de classement

$$\begin{aligned} Log\frac{p_{ij}}{1-p_{ij}} &= \lambda + \lambda_i^F \\ \sum_i \lambda_i^F &= 0 \end{aligned}$$

est dit *conditionnellement indépendant* de la région R.

En termes d'association note-filière, on dira que les associations sont les mêmes dans les régions.

En termes de odds ratios, on dira que le rapport des chances quand on passe d'une région à l'autre est 1 (et pas seulement égales comme dans le cas précédent d'homogénéïté).

En référence au modèle suivant

$$\lambda_1^R = \lambda_2^R = 0 \Longleftrightarrow Log\frac{p_{i1}}{1-p_{i1}} - Log\frac{p_{i2}}{1-p_{i2}} = 0 \Longleftrightarrow \frac{\frac{p_{i1}}{1-p_{i1}}}{\frac{p_{i2}}{1-p_{i2}}} = 1$$

on peut tester l'indépendance conditionnelle, avec un test de Wald sur les coefficients, ou en comparant les déviances : pour les tableaux croisés $2\times2\times K$, la statistique est calculée directement, on la désigne sous le nom de Cochran-Mantel-Haenszel (CMH).

Enfin il existe des tests plus élaborés qui, partant de l'hypothèses des odds ratios tous égaux à 1, prennent pour contre hypothèse les odds ratios égaux et supérieurs à 1, ou égaux et inférieurs à 1. On parle alors de *dépendance conditionnelle* positive, ou négative.

6.3 Modèles loglinéaires associés

La formulation des liaisons entre les variables par le modèle loglinéaire permet de retrouver les résultats de la logistique et d'autres associations, qui s'interprètent bien dans cet exemple excessivement simple où on peut faire la plupart des calculs à la main. La différence essentielle vient de ce que le modèle loglinéaire part d'une loi multinomiale avec 8 paramètres dans le modèle saturé, alors que le modèle logistique considère qu'il compare 4 lois binômiales indépendantes, ce qui implique quatre contraintes correspondant à un conditionnement par les effectifs des quatre sous-populations (effectifs considérés comme n'étant plus aléatoires).

- Modèle logistique

La vraisemblance est le produit de quatre vraisemblances de lois binômiales :

$$Log(L) = Cte + \sum_{ij}\left[n_{ij}^1 Log(\widehat{p}_{ij}^1) + n_{ij}^2 Log(\widehat{p}_{ij}^2)\right]$$

les indices supérieurs 1 et 2 désignent les deux classes de notes.

La constante Cte contient les logarithmes des factorielles qui sont les mêmes pour tous les modèles et qui n'interviennent donc pas dans la maximisation, de sorte que cette partie de la vraisemblance n'est généralement pas reportée dans les logiciels.

Pour le modèle saturé la logvraisemblance est

$$\begin{aligned} Log(L) - Cte &= 3Log\frac{3}{9} + 6Log\frac{6}{9} + 2Log\frac{2}{6} + 4Log\frac{4}{6} \\ &\quad +5Log\frac{5}{13} + 8Log\frac{8}{13} + 1Log\frac{1}{5} + 4Log\frac{4}{5} \\ &= -20.7115 \end{aligned}$$

Pour le modèle où les quatre sous-populations ont la même distribution, où tous les odds ratios sont égaux à 1 d'une population à l'autre :

$$Log\frac{p_{ij}}{1-p_{ij}} = \lambda$$

L'estimation est :

$$Log\frac{p_{ij}^1}{p_{ij}^2} = -0.69$$

La vraisemblance :

$$\begin{aligned} Log(L) - Cte &= 11Log\frac{11}{33} + 22Log\frac{22}{33} \\ &= -21.005 \end{aligned}$$

- **Modèle loglinéaire**

La vraisemblance a la même allure, mais les probabilités estimées ne sont pas celles de chaque loi binomiale, mais celles de la loi multinomiale :

$$Log(L) = Cte + \sum_{ij} n_{ij}^k Log(\hat{p}_{ij}^k)$$

k désigne la classe de notes.

Dans le modèle saturé :

$$\begin{aligned} Log(L) - Cte &= 3Log\frac{3}{33} + 6Log\frac{6}{33} + 2Log\frac{2}{33} + 4Log\frac{4}{33} \\ &\quad +5Log\frac{5}{33} + 8Log\frac{8}{33} + 1Log\frac{1}{33} + 4Log\frac{4}{33} \\ &= -64.179 \end{aligned}$$

Dans le modèle où toutes les probabilités sont égales (quelque soit la note, la filière ou la région) :

$$Log(L) - Cte = 33Log\frac{1}{8} = -68.622$$

Pour décrire (interpréter) les modèles possibles, l'écriture logistique ira généralement du plus simple au plus complexe par introduction des facteurs principaux, puis des interactions simples aux plus complexes, c'est à dire à des associations successives de la variable dépendante, aux variables indépendantes (contrôlées).

L'écriture loglinéaire ira plutôt du modèle saturé par simplifications successives, ce qui par élimination de certaines associations s'exprimera comme des indépendances conditionnelles.

On peut ainsi mettre en face des modèles loglinéaires équivalents aux modèles logistiques (généralisés). On déduira les coefficients en λ de l'expression logistique se déduit des coefficients en μ de l'expression loglinéaire du même modèle.

Logistique
$\text{Logit}(\text{p}_{ij}^1)=0$
$\text{Logit}(\text{p}_{ij}^1)=\lambda$
$\text{Logit}(\text{p}_{ij}^1)=\lambda+\lambda_i^F$
$\text{Logit}(\text{p}_{ij}^1)=\lambda+\lambda_j^R$
$\text{Logit}(\text{p}_{ij}^1)=\lambda+\lambda_i^F+\lambda_j^R$
$\text{Logit}(\text{p}_{ij}^1)=\lambda+\lambda_i^F+\lambda_j^R+\lambda_{ij}^{FR}$

Loglinéaire
$\log(\text{p}_{ijk}) = \mu$
$\log(\text{p}_{ijk}) = \mu+\mu_k^N$
$\log(\text{p}_{ijk}) = \mu+\mu_k^N+\mu_i^F+\mu_{ik}^{FN}$
$\log(\text{p}_{ijk}) = \mu+\mu_k^N+\mu_j^R+\mu_{jk}^{RN}$
$\log(\text{p}_{ijk}) = \mu+\mu_k^N+\mu_j^R+\mu_i^F+\mu_{ik}^{FN}+\mu_{jk}^{RN}+\mu_{ij}^{FR}$
$\log(\text{p}_{ijk}) = \mu+\mu_k^N+\mu_j^R+\mu_i^F+\mu_{ik}^{FN}+\mu_{jk}^{RN}+\mu_{ij}^{FR}+\mu_{ijk}^{FRN}$

D'autres modèles loglinéaires plus complexes n'ont pas de modèle logistique correspondant. Ils fournissent des tests d'indépendance conditionnelle qui concernent les autres facteurs dans le cas où ils ne sont pas contrôlés.

Ainsi considérons que la note n'est pas une conséquence de l'origine des étudiants, mais qu'elle représente un niveau à l'entrée. Les trois facteurs sont sans doute globalement liés (par les critères de "sélection à l'entrée") dans la population inscrite (acceptée) dans ce DESS. On peut alors se demander comment, *dans l'échantillon*, les facteurs sont liés.

On peut alors proposer d'autres modèles :

$$\begin{aligned}
&\log(\text{p}_{ijk}) = \mu+\mu_k^N+\mu_i^F+\mu_j^R \Longrightarrow \textit{indépendance des } N,F,R \textit{ entre eux}\\
&\log(\text{p}_{ijk}) = \mu+\mu_k^N+\mu_i^F+\mu_j^R+\mu_{ij}^{FR} \Longrightarrow N \textit{ indépendant de } F \textit{ et } R\\
&\log(\text{p}_{ijk}) = \mu+\mu_k^N+\mu_j^R+\mu_i^F+\mu_{jk}^{RN}+\mu_{ji}^{RF} \Longrightarrow\\
&\qquad N \textit{ et } F \textit{ indépendants conditionnellement à } R
\end{aligned}$$

Naturellement on doit s'attendre à ce que deux modèles équivalents donnent lieu à des tests équivalents dans la mesure où l'hypothèse (d'indépendance par exemple) nulle est la même, et la *contre-hypothèse aussi*, ce qui n'est pas toujours évident.

On retrouve bien le même χ^2 (3.52) pour la significativité du coefficient des modèles :

$$Logit(p_{ij}^1) = \lambda$$

$$Log(p_{ijk}) = \mu+\mu_k^N$$

De même les χ^2 (0.27 et 3.62) sont les mêmes pour les modèles :

$$Logit(p_{ij}^1) = \lambda+\lambda_i^F \ , \ et \ \ Log(p_{ijk}) = \mu+\mu_k^N+\mu_i^F+\mu_{ik}^{FN}$$

C'est moins simple pour les modèles :

$$Logit(p_{ij}^1) = \lambda + \lambda_i^F + \lambda_j^R$$

ou

$$Log(p_{ijk}) = \mu + \mu_k^N + \mu_j^R + \mu_i^F + \mu_{ik}^{FN} + \mu_{jk}^{RN} + \mu_{ij}^{FR}$$

car les coefficients en λ sont liés de façon plus complexe (néanmoins linéairement) aux coefficients en μ.

Programmes SAS associés :

```
data dess;
input region $ filiere $ note $ effectif;
datalines;
paris eco >=12 6
paris eco <12 3
paris mst >=12 4
paris mst <12 2
province eco >=12 8
province eco <12 5
province mst >=12 4
province mst <12 1
;
run;
proc freq data=dess;
weight effectif;
tables region*note*filiere
 /norow nopercent nocol chisq cmh;
run;
proc logistic data=dess;
weight effectif;
model note= ;
run;
proc catmod data=dess;
weight effectif;
model note*region*filiere=_response_;
loglin note ;
run;
proc logistic data=dess;
weight effectif;
class filiere;
model note= filiere;
run;
proc catmod data=dess;
```

```
weight effectif;
model note*region*filiere=_response_;
loglin note filiere note*filiere;
run;
proc logistic data=dess;
weight effectif;
class region;
model note= region;
run;
proc catmod data=dess;
weight effectif;
model note*region*filiere=_response_;
loglin note region note*region;
run;
proc logistic data=dess;
weight effectif;
class region filiere;
model note=region filiere;
run;
proc catmod data=dess;
weight effectif;
model note*region*filiere=_response_;
loglin note|filiere note|region;
run;
proc logistic data=dess;
weight effectif;
class region filiere;
model note=region filiere region*filiere;
run;
proc catmod data=dess;
weight effectif;
model note*region*filiere=_response_;
loglin note|filiere|region ;
run;
proc catmod data=dess;
weight effectif;
model note*region*filiere=_response_;
loglin note filiere region;
run;
proc catmod data=dess;
weight effectif;
model note*region*filiere=_response_;
loglin note|region filiere|region;
run;
quit;
```

7 Exemple 4 : concours de Premier Surveillant

Les résultats du concours interne de Premier Surveillant sont analysés pour comparer l'âge et l'ancienneté des reçus.

Plusieurs points de vue sont envisageables :

1. Les candidats, les syndicats, considèrent que la réussite est aléatoire, mais que pour des raisons de compétence, d'habitude, ou de politique de la direction, le fait d'être un homme ou une femme, l'âge et l'ancienneté sont des facteurs qui interviennent dans la loi de probabilité. La population est l'*ensemble des surveillants*, le nombre de reçus obéit à une loi de Poisson dont l'espérance mathématique est le logarithme d'une combinaison linéaire des variables genre, âge, ancienneté (effets multiplicatifs).

2. Dans les mêmes conditions, on veut modéliser l'ancienneté des reçus, pour la population des surveillants entrés dans les années 80-84 et 85-89. La variable aléatoire est l'ancienneté, c'est-à-dire le temps d'attente avant le changement de statut, il s'agit d'un *modèle de durée*.

3. La direction du personnel chargée d'affecter les postes de Premier Surveillant a besoin de connaître l'âge des reçus pour évaluer les responsabilités qu'on peut leur confier, ou, pour les mêmes raisons, leur ancienneté. La population est l'*ensemble des reçus cette année*, les variables de classement sont l'âge, l'ancienneté, le genre. On peut admettre que l'âge des reçus est associé à l'ancienneté, au genre, et même, que dans les conditions actuelles l'âge pourrait être expliqué par le genre et l'ancienneté. Ce qui conduirait à estimer l'âge moyen, *attendu*, par l'intermédiaire d'une régression logistique (généralisée). Il pourrait en être de même pour l'ancienneté. Il s'agit d'une étude *rétrospective* puisque la population est fixée par le résultat du concours. L'âge estimé par le modèle pour les hommes, les femmes, avec telle ancienneté, est un ajustement, ou un âge attendu, mais ce n'est pas une prévision, puisque ce n'est qu'un âge moyen observé dans chaque sous-population.

4. Les syndicats désirent avoir une idée des conditions d'accès à la fonction de Premier Surveillant et comparent deux générations de surveillants, recrutés dans les années 80-84 et 85-89, en examinant à quel âge et avec quelle ancienneté ils sont devenus Premier Surveillant. La population est l'ensemble des individus qui sont entrés comme surveillants dans l'Administration Pénitentiaire entre 80 et 89. La variable aléatoire est le passage (oui, non) au grade de Premier Surveillant à telle date fixée, ou dans un intervalle de temps donné. Il s'agit d'un *suivi de cohortes*.

5. Considérons maintenant les seuls Premiers Surveillants de la population précédente, et considérons l'âge qu'ils avaient l'année où ils ont été nommés Premier Surveillant. Cet âge suit une loi multinomiale dont les paramètres dépendent de la génération, du genre, éventuellement de l'ancienneté. Il s'agit d'une régression logistique généralisée, qui ne porte pas sur la même population que dans le cas précédent puisque la probabilité de l'âge est conditionnée par leur succès au concours. Ce qui est probabilisé, c'est l'*âge de réussite*, et non la réussite au concours. Il s'agit d'une étude *rétrospective*.

C'est ce dernier point de vue (numéro 5) que nous allons étudier. Bien que, indépendamment du choix de populations différentes, les résultats des différents traitements puissent donner en termes de probabilités conditionnelles ou de odds ratios des résultats analogues, l'interprétation des modèles serait différente car les conditionnements ne sont pas exactement les mêmes. Les conclusions qu'on voudrait en tirer pourraient alors être contradictoires.

7.1 Tableaux de contingence

Quelques commentaires vont accompagner les tableaux de contingence de l'âge avec chacune des variables qui lui sont associées. La représentation la plus utile pour comparer la distribution des âges est un calcul de pourcentages dans chaque sous-population (genre, génération, ancienneté) :

	genre		
âge ▼	F	H	*total*
30 ou moins	11 (10.19 %)	246 (26.71 %)	257
31 à 35	52 (48.15 %)	348 (37.79 %)	400
36 à 40	30 (27.78 %)	198 (21.50 %)	228
41 et plus	15 (14.39 %)	129 (14.01 %)	144
total	108 (100%)	921 (100%)	1029

Les femmes sont majoritairement dans les classes d'âge 31 à 40, alors que les hommes sont répartis plus uniformément.

	génération		
âge ▼	80-84	85-89	*total*
30 ou moins	114 (19.66 %)	143 (31.85 %)	257
31 à 35	218 (37.59 %)	182 (40.53 %)	400
36 à 40	154 (26.55 %)	74 (16.48 %)	228
41 et plus	94 (16.21 %)	50 (11.14 %)	144
total	580 (100%)	449 (100%)	1029

Hommes et femmes confondus de la génération 85-89 sont nommés à un âge plus bas que dans la génération précédente.

	ancienneté			
âge ▼	1-8	9-10	11 et plus	*total*
30 ou moins	219 (39.67 %)	38 (11.91 %)	0	257
31 à 35	177 (32.07 %)	162 (50.78 %)	61 (38.61 %)	400
36 à 40	90 (16.30 %)	71 (22.26 %)	67 (42.41 %)	228
41 et plus	66 (11.96 %)	48 (15.05 %)	30(18.98 %)	144
total	552 (100%)	319 (100%)	158 (100%)	1029

La première colonne montre qu'à ancienneté égale, les plus jeunes sont plus nombreux. La ligne des plus de 40 ans montre que ce sont les moins anciens qui ont été promus. Mais, faute de connaître la structure de l'échantillon des candidats, on se gardera bien de traduire ces remarques en termes de probabilité d'être reçu pour les jeunes et les plus anciens.

Le zéro vient de ce qu'on trouve difficilement des surveillants de 30 ans avec 11 ans d'ancienneté, qui seraient donc nommés surveillants à 19 ans.

Ce zéro peut-être considéré comme "structurel", c'est-à-dire comme une contrainte externe au modèle. Sinon, on considère que le zéro résulte de l'échantillonnage mais que la probabilité d'appartenir à la cellule n'est pas nulle et doit être paramétrée. Il arrive qu'on ne puisse pas l'évaluer. Ce zéro pourrait aussi poser un problème dans l'évaluation du calcul de vraisemblance lors de l'estimation.

Il s'agit donc d'examiner comment l'âge des reçus au concours de Premier Surveillant (variable polytomique) est lié aux trois variables de classement : genre, génération, ancienneté.

Notre démarche va consister à examiner des modèles relativement simples, à vérifier leur validité statistique et leur adéquation à la question posée :

quel âge ont les Premiers Surveillants l'année de leur nomination?

Mais nous n'en saurons pas plus sur *les conditions d'accès* aux fonctions de Premier Surveillant.

7.2 Modèle logistique polytomique (ordinal) simplifié

On commence par un modèle où l'âge, en quatre classes, est considéré comme une variable ordinale. La loi de probabilté est multinomiale, la fonction de lien porte sur des *probabilités cumulées* :

$$Log\frac{1-p_{1jkm}}{p_{1jkm}} = \lambda_1^{age} + \lambda_{1j}^{genre} + \lambda_{1k}^{gener} + \lambda_{1m}^{anc}$$

$$Log\frac{1-(p_{1jkm}+p_{2jkm})}{p_{1jkm}+p_{2jkm}} = \lambda_2^{age} + \lambda_{2j}^{genre} + \lambda_{2k}^{gener} + \lambda_{2m}^{anc}$$

$$Log\frac{1-(p_{1jkm}+p_{2jkm}+p_{3jkm})}{p_{1jkm}+p_{2jkm}+p_{3jkm}} = \lambda_3^{age} + \lambda_{3j}^{genre} + \lambda_{3k}^{gener} + \lambda_{3m}^{anc}$$

p_{ijkm} sont les probabilités des classes d'âge i, de genre j, de génération k, d'ancienneté m.

Avec les contraintes :

$$\forall i \ \sum_j \lambda_{ij}^{genre} = 0, \ \forall i \ \sum_k \lambda_{ik}^{gener} = 0, \ \forall i \ \sum_m \lambda_{im}^{anc} = 0$$

$$i = 1,2,3 \qquad j = 1,2 \qquad k = 1,2 \qquad m = 1,2,3$$

Le zéro signalé dans les tableaux croisés ne sera pas considéré comme une contrainte structurelle supplémentaire. Pour des raisons dues au logiciel SAS, il faut alors ajouter un élément dans chaque cellule du tableau de contingence, ce qui ne modifie guère les résultats compte tenu des effectifs en cause.

Il y a alors $3{\times}2 \times 2 \times 3 = 36$ classes indépendantes, le modèle (sans interactions) n'a que 15 paramètres : $3 + 3 \times (1 + 1 + 2) = 15$

La part apportée par chaque variable dans une estimation (moindres carrés généralisés) figure dans le tableau suivant :

	d° de liberté	$Wald$	p-value
genre	3	7.31	0.063
génération	3	11.54	0.009
ancienneté	6	91.66	<0.0001

Si on impose en plus la proportionnalité des odds ratios, le modèle s'écrit :

$$\log \frac{1 - p_{1jkm}}{p_{1jkm}} = \lambda_1^{age} + \lambda_j^{genre} + \lambda_k^{gener} + \lambda_m^{anc}$$

$$\log \frac{1 - (p_{1jkm} + p_{2jkm})}{p_{1jkm} + p_{2jkm}} = \lambda_2^{age} + \lambda_j^{genre} + \lambda_k^{gener} + \lambda_m^{anc}$$

$$\log \frac{1 - (p_{1jkm} + p_{2jkm} + p_{3jkm})}{p_{1jkm} + p_{2jkm} + p_{3jkm}} = \lambda_3^{age} + \lambda_j^{genre} + \lambda_k^{gener} + \lambda_m^{anc}$$

avec $3 + 1 + 1 + 2 = 7$ paramètres.

La part apportée par chaque variable, indépendamment des autres, est donnée ci-dessous ("type III analysis of effects"):

	d° de liberté	$Wald$	p-value
genre	1	4.97	0.025
génération	3	12.40	0.0004
ancienneté	6	80.42	<0.0001

Ces tests peuvent être peu fiables dans la mesure où on n'a pas encore introduit d'interactions ; le test de proportionnalité des odds ratios ($Wald$) donne :

$\chi^2 = 83.85$ pour 8 d° de liberté, ce qui correspond à une p-value <0.0001, et qui conduit à un rejet de l'hypothèse de proportionnalité.

Programme SAS associé :

```
proc freq data=cours.admpen;
tables age*(anc genre gener);
run;
proc catmod data=cours.admpen;
response clogits;
model age=genre gener anc
/nodesign noprofile noresponse addcell=1;
run;
proc logistic data=cours.admpen;
class genre gener anc;
model age=genre gener anc;
run;
```

7.3 Modèle logistique polytomique (ordinal) avec interactions

En gardant toujours le même modèle de départ avec les probabilités cumulées, on peut examiner toutes les interactions possibles, éliminer successivement celles qui ne sont pas significatives pour arriver finalement à un modèle où ne sont conservées que les effets principaux et l'interaction entre ancienneté et genre (significativement non nulle au seuil de 0.005).

L'introduction de l'interaction confirme la non validité de l'hypothèse des odds ratios proportionnels.

Nous devons reprendre la régression sans cette contrainte.

Le tableau suivant donne la part apportée par chaque variable dans l'analyse. Il s'agit des variances expliquées successivement par chaque variable et puis l'interaction.

	d° de liberté	$Wald$	p-value
genre	3	7.07	0.0698
génération	3	7.08	0.0695
ancienneté	6	95.90	<0.0001
gener×anc	6	18.99	0.0042

On peut alors effectuer un test de proportionnalité des odds ratios (à partir du modèle complet) en testant l'égalité simultanée des coefficients $\lambda_{ij}^{genre}, \lambda_{ik}^{gener}, \lambda_{im}^{anc}$ ($pour\ i = 1, 2, 3$).

La statistique du χ^2 (6 d° de liberté) est 33.79 (p-value < 0.001) , elle confirme la non proportionnalité. Cette fois, le test est fait à partir du modèle non contraint alors que précédemment il était fait à partir du modèle contraint.

Les variables explicatives ou descriptives des différences d'âge sont donc le genre et le couple (ancienneté, génération) ; ce qui signifie qu'il y a une différence entre les hommes et les femmes, différence qui n'est pas liée à la génération, ni à l'ancienneté.

Les odds ratios correspondants sont calculés en prenant l'exponentielle des différences des fonctions "réponses" combinaison linéaires des coefficients estimés :

âges	réponses	odds ratio H/F
≤ 30 *contre* > 30	0.366+0.366 = 0.732	2.08
≤ 35 *contre* > 35	0.141+0.141 = 0.282	1.32
≤ 40 *contre* > 40	0.062+0.062 = 0.124	1.13

Ce qui signifie que les hommes promus sont plus jeunes que les femmes, quelles que soient leur ancienneté et la génération à laquelle ils appartiennent. Cette différence d'âge est surtout sensible pour les moins de 31 ans.

L'effet croisé génération×ancienneté est plus difficile à évaluer, il faut revenir à la définition de la fonction de lien estimée.

L'interaction principale s'exprime dans les rapports :

$$\frac{\Pr(promu\ <31\ /\ 80-84, 1-8)}{\Pr(promu\ \geq 31\ /\ 80-84, 1-8)} \Big/ \frac{\Pr(promu\ <31\ /\ 80-84, 9-10)}{\Pr(promu\ \geq 31\ /\ 80-84, 9-10)}$$

et

$$\frac{prob(promu\ <31\ /\ 85-89, 1-8)}{prob(promu\ \geq 31\ /\ 85-89, 1-8)} \Big/ \frac{prob(promu\ <31\ /\ 85-89, 9-10)}{prob(promu\ \geq 31\ /\ 85-89, 9-10)}$$

qui valent respectivement :

exp(2.2907-1.4437)=2.33

exp(3.1277-0.7962)=10.3

Ce qui veut dire que dans la génération la plus récente 85-89, l'âge des promus est plus lié à l'ancienneté (1-8 contre 9-10), pour les hommes comme pour les femmes (indépendamment de la variable genre) au sens où les plus jeunes promus sont moins anciens. Cette liaison entre jeunes et moins anciens n'est pas étonnante, mais elle est plus forte sur la seconde génération que sur la première.

Programme SAS associé :

```
proc logistic data=cours.admpen;
class genre anc(ref='11 &+') gener;
model age=genre gener|anc ;run;
proc catmod data=cours.admpen ;
response clogits;
model age=genre gener|anc/addcell=1 pred;
contrast 'odds ratios proportionnels'
all_parms 0 0 0 1 -1 0 0 0 0 0 0 0,
all_parms 0 0 0 0 1 -1 0 0 0 0 0 0,
all_parms 0 0 0 0 0 0 1 -1 0 0 0 0,
```

```
all_parms 0 0 0 0 0 0 0 1 -1 0 0 0,
all_parms 0 0 0 0 0 0 0 0 0 1 -1 0 ,
all_parms 0 0 0 0 0 0 0 0 0 0 1 -1 ;
run;
```

7.4 Modèle logistique généralisé (non ordinal)

Quand les modalités de la variable expliquée ne sont pas ordinales ou qu'on ne veut pas tenir compte de l'ordre, il faut calculer des probabilités conditionnelles, en référence à une modalité particulière (la dernière par exemple, *41 ans et plus,* désignée par l'indice $i = 4$), le modèle s'écrit :

$$Log\frac{1-p_{ijkm}}{p_{4jkm}} = \lambda_i^{age} + \lambda_{ij}^{genre} + \lambda_{ik}^{gener} + \lambda_{im}^{anc} + \lambda_{ikm}^{anc\times gener}$$

avec les contraintes :

$$\begin{aligned} \forall i \sum_j \lambda_{ij}^{genre} &= 0,\ \forall i \sum_k \lambda_{ik}^{gener} = 0, \forall i \sum_m \lambda_{im}^{anc} = 0 \\ \forall i,m \sum_k \lambda_{ikm}^{anc\times gener} &= 0,\ \forall ik \sum_m \lambda_{ikm}^{anc\times gener} = 0 \end{aligned}$$

Les résultats sont résumés dans les tableaux suivants où certains paramètres (signalés par *) peuvent n'avoir pas de sens sans pour autant invalider les autres, ni mettre en cause les tests globaux portant sur chaque variable.

Quand le logiciel traite les zéros du tableau croisé comme une contrainte structurelle, certains paramètres seront redondants dans l'estimation et certains odds ratios n'auront pas de sens (infinis). Le couple (*30 ans ou moins*, *11 et plus*) ayant un effectif nul pour les deux générations, ne gènera que l'estimation de l'effet croisé génération×ancienneté.

Constantes (λ_i^{age})

âge ▼	coefficient
30 ans ou moins	-5.1661
31 à 35 ans	1.0495
36 à 40 ans	0.5201

Effet principal **genre** (λ_{ij}^{genre})

âge ▼	femmes	hommes
30 ans ou moins	-.0.4957	0.4957
31 à 35 ans	0.1057	-0.1057
36 à 40 ans	0.1341	-0.1341

Effet principal **génération** (λ_{ik}^{gener}) :

âge ▼	80-84	85-89
30 ans ou moins	-0.0098	0.0098
31 à 35 ans	-0.1244	0.1244
36 à 40 ans	0.1040	-0.1040

Effet principal **ancienneté** (λ_{im}^{anc}) :

âge ▼	1-8	9-10	11 et plus
30 ans ou moins	5.9496	4.3711	-10.3207
31 à 35 ans	0.0575	0.2683	-0.3258
36 à 40 ans	-0.0772	-0.0362	0.1134

Effet croisé **ancienneté**×**génération**

	génération 80-84		
ancienneté ▶	1-8	9-10	11 et plus
30 ans ou moins	-0.5124	0.3730	0.1394
31 à 35 ans	-0.1447	-0.0958	0.2405
36 à 40 ans	-0.2431	-0.0667	0.3098

	génération 85-89		
ancienneté ▶	1-8	9-10	11 et plus
30 ans ou moins	0.5124	-0.3730	-0.1394*
31 à 35 ans	0.1447	0.0958	-0.2405
36 à 40 ans	0.2431	0.0667	-0.3098

Si les coefficients sont classés régulièrement avec les modalités de la variable expliquée, l'ordinalité de la variable âge est plausible, *mais par pour autant la proportionnalité.*

Programme SAS associé :

```
proc catmod data=cours.admpen;
model age=genre gener|anc
/nodesign noprofile noresponse noiter pred;
run;
proc logistic data=cours.admpen;
class genre gener anc;
model age=genre gener anc/link=glogit;
run;
```

8 Exemple 5 : gravité des accidents

La population étudiée est celle des conducteurs ayant eu un accident (dans le département du Nord, durant l'année 1998). La variable aléatoire est la gravité de l'accident. Les variables de classement sont l'âge et le genre.

Dire que variables explicatives sont l'âge et le genre, et que la variable expliquée est la gravité mesurée par le nombre de tués, revient à sous-entendre que les variables âge et genre sont causales, et pas seulement des variables de classement en sous-populations qui décrivent bien la répartition des accidents graves. Ce n'est pas le *modèle statistique* qui est causal, c'est l'usage qu'on en fait. Nous dirons plutôt qu'il s'agit d'une *description* de la population des conducteurs ayant eu un accident, laissant à d'autres le soin de décider si celle-ci peut aider à dégager des causalités.

8.1 Construction d'un modèle

Supposons que, pour simplifier, on réduise la gravité à la variable binaire (*0 tué*, *au moins un tué*).

Il est évident qu'on ne mesure pas le risque d'être tué puisque la population des *personnes susceptibles d'être tuées lors d'un accident* comprend l'ensemble des personnes impliquées dans l'accident et n'est pas connue ; encore moins celle des *personnes susceptibles d'être impliquées* dans un accident.

On ne mesure pas non plus la probabilité *pour qu'un jeune (18-24 ans) ait provoqué un accident grave*, ni la probabilité *pour qu'un accident grave ait été provoqué par un jeune*, mais plutôt la probabilité *pour qu'un accident, provoqué par un jeune, soit grave.* La probabilité porte uniquement sur la gravité, et on veut la comparer, selon que l'accident a été provoqué par une sous-population (jeune) plutôt que par une autre. L'objectif de l'étude est donc une comparaison des chances de gravité (odds) ou des risques de gravité.

Dans ce cas particulier où les probabilités (risque de gravité) sont très faibles, travailler sur les rapport de risques ou sur les odds ratios revient au même, en soulignant que dans les deux cas on ne compare que des sous-populations de la population des *conducteurs ayant provoqué un accident* et non la population des usagers de la route.

Les données disponibles et les objectifs ayant été définis, nous pouvons examiner les méthodes statistiques et tout d'abord les lois de probabilités à utiliser.

Si la gravité est simplifiée sous la forme d'une variable binaire, il est naturel de considérer qu'on attribue à chaque sous-population (classée par âge et genre) une probabilité pour qu'un accident soit grave. Il ne s'agit pas de la probabilité pour qu'un individu particulier *ait provoqué* un accident grave, mais de la probabilité que l'accident, provoqué par un individu de la sous-population, *soit grave.* On pourrait faire dépendre cette probabilité de l'état de la route, du jour de la semaine, etc. Mais avec les données dont nous disposons on admettra qu'à chaque sous-population correspond un paramètre à estimer (probabilité).

Ensuite nous supposerons que la gravité de chaque accident n'est pas liée à celle des autres accidents au sens où un accident grave n'implique pas que le suivant soit plus ou moins grave : il s'agit d'évènements indépendants en probabilité. Cette hypothèse pourrait être mise en cause quand certains accidents ont un grand retentissement médiatique.

En tenant compte de toutes ces simplifications, nous sommes conduits à considérer que le nombre d'accidents graves suit une loi binomiale dans chaque sous-population.

Le paramètrage du modèle consiste à savoir si les paramètres des différentes lois binomiales des sous-populations (ici il y a 8 lois avec 4 tranches d'âge et 2 genres) sont liés ou non. Sont-ils égaux, sont-ils indépendants de l'âge, différent-ils selon l'âge mais indépendamment du genre.

Le modèle logistique est capable de répondre à ces questions en exprimant les probabilités sous forme de odds ratios paramètrés par des effets multiplicatifs :

$$Log\frac{p_{ij}}{1-p_{ij}} = \lambda_i^{age} + \lambda_j^{genre} + \lambda_{ij}^{age\times genre}$$

p_{ij} est la probabilité qu'un accident soit grave quand il est provoqué par une personne de la tranche d'âge i et de genre j.

On peut vouloir être plus précis en définissant la gravité en trois niveaux selon le nombre de tués. La loi de probabilité est alors multinomiale et, avec un paramètrage analogue au précédent, on utilise un modèle logistique généralisé :

$$\begin{aligned} Log\frac{p_{ij1}}{p_{ij3}} &= \lambda_{i1}^{age} + \lambda_{j1}^{genre} + \lambda_{ij1}^{age\times genre} \\ Log\frac{p_{ij2}}{p_{ij3}} &= \lambda_{i2}^{age} + \lambda_{j2}^{genre} + \lambda_{ij2}^{age\times genre} \end{aligned}$$

p_{ij1}, p_{ij2} sont les probabilités que la gravité soit de niveau 1 ou 2 (pour un conducteur d'âge i et de genre j.

p_{ij3} est la probabilité du niveau de gravité de référence (par exemple aucun tué)

On peut aussi tenir compte de l'ordinalité de la gravité et écrire un modèle logistique cumulé :

$$\begin{aligned} Log\frac{p_{ij1}}{p_{ij2}+p_{ij3}} &= \lambda_{i1}^{age} + \lambda_{j1}^{genre} + \lambda_{ij1}^{age\times genre} \\ Log\frac{p_{ij1}+p_{ij2}}{p_{ij3}} &= \lambda_{i2}^{age} + \lambda_{j2}^{genre} + \lambda_{ij2}^{age\times genre} \end{aligned}$$

Enfin si on veut tenir compte de l'information détaillée sur le nombre de tués, on peut envisager une loi de Poisson avec effets multiplicatifs qui n'est d'ailleurs pas facile à justifier : le modèle de Poisson consisterait à *répartir les tués* dans les sous-populations de conducteurs. Le problème est complètement inversé puisque ce sont les usagers de la route qui deviennent la population (des lapins en quelque sorte), ils ont une probabilité d'être tués différente selon qu'ils rencontrent un conducteur d'âge et de genre différent (des chasseurs en quelque sorte). Ce n'est peut-être pas très plausible bien que statistiquement les résultats sont très proches dans notre exemple.

8.2 Logistique binaire

Un premier modèle simplifié sera examiné d'abord pour servir de référence à des modèles plus complexes et mieux voir ce qu'ils apportent.

La gravité sera réduite à l'existence de tués (*gravité forte*) et la non existence de tués (*gravité faible*) ; les accidents répertoriés ont tous donné lieu à un rapport de police ou de gendarmerie, avec présence de dommages corporels. Il y a donc toujours au moins un blessé même s'il n'y a pas de tués.

Les variables de classements sont l'âge divisé en 4 classes choisies pour l'homogénéité des comportements des conducteurs, et le genre (*hommes, femmes*).

Il y a donc 8 sous-populations et 7 paramètres si on tient compte des interactions âge×genre.

Les tests dit de *type III* mesurent la significativité de chaque effet par rapport à l'ensemble des autres effets (apport d'un effet dans la décomposition de la variance ou en termes de déviance, par variation du χ^2 quand on retire cet effet en gardant les autres).

effet	*d° de liberté*	*Wald*	*p-value*
âge	3	87.08	<0.001
genre	1	14.89	<0.001
âge×genre	3	6.18	0.103

Il est alors possible de laisser tomber l'effet croisé si on s'en tient aux traditionnels 5% pour la significativité.

Le modèle réduit aux effets principaux n'a que 4 paramètres :

effet	*d° de liberté*	*Wald*	*p-value*
âge	3	155.7	<0.001
genre	1	46.16	<0.001

Les statistiques du χ^2 sont plus élevées. Elles n'ont pas le même sens que dans le modèle précédent car la séparation entre sous-populations (hommes, femmes).apporte 46.15 de plus qu'avec l'âge seul, alors que dans le modèle précédent 14.89 mesure ce qu'apporte de plus la séparation (hommes, femmes) *à l'âge et au croisement âge×genre*, ce qui n'est pas facile à interpréter. C'est pourquoi on considère que lorsqu'il y a un croisement comme dans le modèle précédent, les statistiques sur les effets principaux correspondants ne sont pas très intéressantes.

Enfin, on peut vérifier que la statistique portant sur le croisement est proche de la déviance :

$$-2LogL_2 - (-2LogL_1) = 9229.420 - 9223.521 = 5.90 \ (\textit{proche de } 6.18)$$

Le modèle avec effets principaux, âge et genre, peut être écrit de multiples façons selon qu'on probabilise la gravité forte contre la faible ou réciproquement, et selon que l'on introduit ou non une constante générale. Par exemple :

$$Log\frac{\Pr(gravité\ faible/\ âge\ i, genre\ j)}{\Pr(gravité\ forte/\ âge\ i, genre\ j)} = \lambda + \lambda_i^{age} + \lambda_j^{genre}$$

$$\sum_i \lambda_i^{age} = 0,\ \sum_j \lambda_j^{genre} = 0$$

Les paramètres estimés sont :

variable	**coefficient**	**écart-type**	**p-value**
constante	3.0326	0.0588	<0.001
age 16-17	0.9957	0.0872	<0.001
age 18-24	0.1684	0.0652	0.010
age 25-64	-0.1269	0.0603	0.035
age 65 et +	1.0372 *déduit de la contrainte*		
femme	0.2741	0.0403	<0.001
homme	-0.2741 *déduit de la contrainte*		

Les logiciels ne donnent généralement que les paramètres libres, ce qui suffit pour effectuer des tests. Le test de nullité du coefficient d'une modalité signifie qu'elle ne se distingue pas de la moyenne générale parce que la contrainte sur les coefficients impose que cette moyenne soit nulle.

On peut préférer que les coefficients soient calculés en fixant une modalité (dite de référence) dont le paramètre est contraint à 0. Les résultats sont évidemment les mêmes mais la présentation diffère : l'exponentielle des coefficients donne alors immédiatement les odds ratios d'une sous-population par rapport à la sous-population de référence.

Par exemple :

variable	**coefficient**	**écart-type**	**p-value**
constante	1.7213	0.1782	<0.001
age 16-17	2.0329	0.2023	<0.001
age 18-24	1.2056	0.1850	<0.010
age 25-64	0.9103	0.1817	<0.001
age 65 et +	0 *contraint*		
femme	0.5482	0.0807	<0.001
homme	0 *contraint*		

La plupart des logiciels donnent les odds (ou odds ratios) des effets principaux (et des intervalles de confiance), pour les modèles avec seulement des effets principaux. On en déduit par multiplication tous les autres odds ratios. Quelle que soit la représentation, on peut déduire les odds ratios par l'exponentielle de la différence de deux coefficients.

Exemple :

effet	odds	intervalle à 5%	
age 16-17 contre 65 et +	7.636	5.137	11.351
age 18-24 contre 65 et +	3.339	2.323	4.798
age 25-64 contre 65 et +	2.485	1.741	3.548
genre femme contre homme	1.730	1.477	2.027

Si on prend pour référence les accidents des 25-64 (les plus nombreux parce que correspondant aux plus nombreux usagers), on obtient :

effet	odds	intervalle à 5%	
age 16-17 contre 25-64	3.073	2.498	3.780
age 18-24 contre 25-64	1.344	1.179	1.531
age 65 et + contre 25-64	0.402	0.282	0.574
genre femme contre homme	1.730	1.477	2.027

Les deux tableaux présentent différemment la même information :

$$\begin{aligned} Odd(age16-17\ contre\ 25-64) &= \frac{Odd(age16-17\ contre\ 65et+)}{Odd(age25-64\ contre\ 65et+)} \\ &= \frac{7.636}{2.485} = 3.073 \end{aligned}$$

Si on s'intéresse aux chances de gravité forte contre gravité faible, les coefficients des paramètres changent de signe et les odds sont les inverses des précédents.

Comme il s'agit des chances de *gravité faible* contre *gravité forte*, ces résultats montrent bien que la gravité des accidents augmente avec l'âge du conducteur, au sens où plus l'accident est grave, plus on a de chances que le conducteur soit agé. Ces résultats (probabilités conditionnelles) ne sont pas suffisantes pour savoir si les jeunes sont moins dangereux que les plus âgés ; il faudrait connaître en plus les probabilités (a priori) pour que les jeunes et les moins jeunes aient un accident.

Programme SAS associé :

```
proc format ;
value gravA 0='faible' 1-high='forte';
value gravB 0='niveau 0' 1-2='niveau 1-2' 3-high='niveau 3';
value gravC 0='niveau 0' 1='niveau 1' 2-high='niveau 2';
run;
proc logistic data=cours.acc;
format grav grava.  ;
class age genre;
model grav=age|genre;
run;
```

```
proc logistic data=cours.acc;
format grav grava.  ;
class age genre;
model grav=age genre;
run;
proc logistic data=cours.acc;
format grav grava.  ;
class age(ref='25-64') genre;
model grav=age genre ;
run;
proc logistic data=cours.acc;
format grav grava.  ;
class age genre/param=glm;
model grav=age genre ;
run;
```

8.3 Logistique polytomique

Revenons aux données originales, avec la gravité mesurée par le nombre de tués dans chaque accident, et considérons ce nombre comme une variable discrète, ordinale. Une logistique généralisée n'est pas viable parce qu'il y a trop de zéros dans la table. Aussi peut-on, comme première contrainte, envisager un modèle avec odds ratios proportionnels, c'est à dire avec une liaison sur les probabilités cumulées.

Il y a désormais 6 réponses (0,1,..5 tués), 4 âges, 2 genres on dispose de $5 \times 4 \times 2 = 40$ paramètres possibles (modèle saturé). La contrainte de proportionnalité réduit à $5 + (3 + 1 + 3) = 12$ paramètres à estimer.

Le test qui vérifie la proportionnalité suit un χ^2 à 28 degrés de liberté. Ici, la proportionnalité est rejetée avec un risque de 0.0001.

Il faut donc essayer de regrouper les modalités ou renoncer à la contrainte de proportionnalité.

Deux essais ont été faits. En trois classes définies par *0 tué, 1 tué, 2 tués et plus.* La p-value du test de proportionalité est alors de 7%. Avec les classes *0 tué, 1 ou 2 tués, 3 tués et plus,* la p-value est 39%.

On peut donc effectuer une régression logistique avec odds ratios proprtionnels dans ces deux cas qui d'ailleurs conduisent à des résultats très proches. Ce qui correspond bien à ce qu'on attend de ce type de modèle où de petites différences dans le découpage en classes ne doit pas modifier les résultats.

Les résultats de ce dernier modèle (*0 tué, 1 ou 2 tués, 3 tués et plus)* comporte des constantes qui servent au calcul de probabilités cumulées, mais les odds ratios communs sont très proches du modèle (*0 tué, 1 tué, 2 tués et plus),* et du modèle logistique binaire.

variable	**coefficient**	**écart-type**	**p-value**
constante 0	3.0333	0.0588	<0.001
constante 1-2	7.2184	0.2350	<0.001
age 16-17	0.9953	0.0872	<0.001
age 18-24	0.1677	0.0653	0.010
age 25-64	-0.1279	0.0604	0.034
age 65 et +	*1.0451 déduit de la contrainte*		
femme	0.2739	0.0403	<0.001
homme	*-0.2739 déduit de la contrainte*		

effet	odds ratio	intervalle à 5%	
age 16-17 contre 65 et +	7.617	5.122	11.325
age 18-24 contre 65 et +	3.329	2.316	4.786
age 25-64 contre 65 et +	2.477	1.735	3.538
genre femme contre homme	1.729	1.477	2.026

Si on avait rejeté l'hypothèse des odds ratios proportionnels, on aurait eu (probabilités cumulées ou non) quatre paramètres supplémentaires, mais les zéros du tableau croisé correspondant auraient introduit des contraintes.

De plus, l'absence d'accidents avec plus de 1 tué avec les conducteurs de plus de 64 ans, et ceci pour les hommes comme pour les femmes rend le modèle *non identifiable ;* ce qui signifie qu'il y a des relations linéaires entre les paramètres estimés, d'où des valeurs infinies possibles et des redondances qui empèchent d'estimer certains groupes de paramètres.

Compte tenu de l'importance des groupes *0 tué* et *1 tué* par rapport aux autres, un modèle avec une réponse binaire (*0 tué* et *1 tué)* ou (*0 tué* et *1 tué et plus)* est parfaitement acceptable et il évite de comparer les hommes et les femmes pour la sous-population *2 tués et plus* qui est très déséquilibrée, 8 femmes pour 80 hommes, alors que dans la population totale il y a 5928 femmes pour 17499 hommes. D'après ces derniers nombres la tranche *2 tués et plus* présente tout de même un certain intérêt.

Une toute autre façon d'estimer les effets de l'âge et du genre est de concevoir le même modèle en considérant que ce sont les tués qui sont répartis aléatoirement dans les sous-populations (*les lapins ne choisissent pas leur chasseur*).

On peut alors étudier la répartition des tués selon les types de conducteurs, ce qui dans l'échantillon donné revient à travailler avec une loi multinomiale à 7 paramètres (ou de façon équivalente avec une loi de Poisson conditionnée par le nombre total de tués).

Si on ne retient que les effets de l'âge et du genre dans l'expression de l'espérance mathématique de la loi de Poisson, il n'y a que $1 + 3 + 1 = 5$ paramètres à estimer.

Si Λ_{ij} désigne l'espérance du nombre de tués par conducteur d'âge i et de genre j,

$$Log(\Lambda_{ij}) = \lambda + \lambda_i^{age} + \lambda_j^{genre}$$

$$\sum_i \lambda_i^{age} = 0, \sum_j \lambda_j^{genre} = 0$$

En fait les résultats des deux modèles sont proches dans le cas présent ; la différence vient de ce qu'avec la loi de Poisson on n'a pas fusionné les cas 1 tué et 2 tués dans une réponse commune *gravité forte*. Le modèle suppose la proportionnalité des odds ratios (constante unique λ).

Logistique		
variable	**coefficient**	**p-value**
age 16-17	2.0303	<0.001
age 18-24	1.2028	<0.001
age 25-64	0.9072	<0.001
age 65 et +	*0 contraint*	
femme	0.5478	<0.001
homme	*0 contraint*	

Poisson		
variable	**coefficient**	**p-value**
age 16-17	1.8592	<0.001
age 18-24	1.0005	<0.001
age 25-64	0.7261	<0.001
age 65 et +	*0 contraint*	
femme	0.5346	<0.001
homme	*0 contraint*	

(les coefficients du modèle de gauche ont été déduits du dernier modèle logistique présenté plus haut en changeant seulement les contraintes sur les coefficients).

Programme SAS associé

```
proc logistic data=cours.acc;
class age genre;
model grav=age genre ;
run;
proc logistic data=cours.acc;
format grav gravb.  ;
class age genre;
model grav=age genre ;
run;
proc logistic data=cours.acc;
format grav gravc.  ;
class age genre;
model grav=age genre ;
run;
proc genmod data=cours.acc;
class age genre;
model grav=age genre /dist=poisson link=log type3;
run;
proc logistic data=cours.acc;
format grav gravb.  ;
class age genre/param=glm;
model grav=age genre ;
run;
```

9 Exemple 6 : aspirine

Pour mesurer l'efficacité du traitement par l'aspirine de patients à risques, la variable à expliquer doit rendre compte de la récidive (polypes adénomateux).

L'examen va porter sur le "burden", indicateur de la masse totale des adénomes. Cette variable discrétisée sera expliquée, ou du moins conditionnée, par le traitement et par des variables caractérisant le patient, susceptibles d'intervenir sur l'efficacité du traitement.

D'autres variables comme la présence ou non d'adénomes à la seconde coloscopie (après une année de traitement) sont aussi intéressantes. Les variables associées à la récidive peuvent être nombreuses ce qui exige une stratégie de sélection pour rechercher les associations, quand le nombre d'interactions possibles exclut une étude systématique de tous les modèles.

9.1 Une logistique élémentaire pour le *burden*

Le burden, variable quantitative, est discrétisé en deux classes choisies pour des raisons médicales et comme représentant bien la gravité de la récidive :

- $burden < 6\ mm$
- $burden \geq 6\ mm$

Les facteurs de récidives sont limités *a priori* au traitement, à l'âge (en trois classes), au sexe, aux antécédents (présence ou non de polypes observés dans les coloscopies précédentes), à l'état du patient à la coloscopie initiale, facteurs qui sont tous considérés par les médecins comme susceptibles d'être associés au traitement.

Un premier traitement statistique consiste à analyser l'effet marginal, brut, de chaque facteur indépendemment des autres, à essayer un modèle simple où les facteurs sont considérés comme indépendants (effet multiplicatif des odds ratios), puis à essayer des interactions entre les facteurs. Cette démarche est rapide et justifiée quand on a des raisons solides (biologiques en l'occurence) pour penser que les effets des facteurs sont indépendants en grande partie ; ce qui évite d'avoir à traiter toutes les interactions possibles.

On sait aussi que si une interaction existe entre deux variables, l'ignorer peut conduire à négliger des variables importantes, à biaiser les estimations des effets réels (paradoxe de Simpson). Aussi cette méthode de sélection ascendante en partant des effets principaux n'a de sens que si les facteurs explicatifs ne sont pas associés dans l'échantillon pour s'assurer que les estimations ne sont pas biaisées.

Comme la variable la plus importante pour l'étude est l'effet du traitement sur la récidive, il est naturel de commencer par examiner les régressions logistiques qui prennent en compte les interactions entre le traitement et chaque facteur de récidive.

Si le traitement a été distribué au hasard dans l'échantillon (*randomisation* correcte), le traitement et chaque facteur sont indépendants au sens où la distribution de chaque facteur est la même pour les deux sous-populations (aspirine, placebo). Dans ce cas l'estimation de

l'effet marginal du traitement ne sera pas biaisée, mais on ne saura rien sur les interactions possibles.

En particulier le traitement et l'âge, par exemple, sont-ils des facteurs de récidives sans interaction?

(*Il est bien entendu que s'ils ont des effets multiplicatifs ou non, ce n'est aucunement lié à la construction de l'échantillon).*

Les résultats des cinq régressions logistiques qui figurent dans le tableau suivant sont réduits aux odds ratios, la significativité (p-value), pour le facteur traitement (contrôlé), et le facteur lié au patient.

facteurs	odds		p-value	traitement	odds	p-value
âge	<55 vs 65 et +	0.41	0.146	asp vs placebo	0.45	0.032
	56-64 vs 65 et +	0.62				
sexe	femme vs homme	0.29	0.014	asp vs placebo	0.41	0.017
antécédents	non vs oui	0.37	0.09	asp vs placebo	0.44	0.030
adénomes*	1-2 vs 3 et +	0.17	0.0001	asp vs placebo	0.40	0.022
effet marginal réduit au seul traitement				asp vs placebo	0.43	0.022

* *Il s'agit du nombre d'adénomes à la coloscopie initiale avant tout traitement.*

Ce qu'on a fait pour le traitement, n'est pas applicable aux autres variables qui caractérisent la population et qu'on ne contrôle pas. Aussi, tester séparément tous les couples

$traitement \times sexe, traitement \times âge, etc.$

n'est pas justifié (biais des estimateurs et des tests) à moins que toutes les facteurs significatifs ne soient orthogonaux entre eux, ce qui est rarement le cas.

Il faut donc se résoudre à les traiter simultanément. Une procédure de sélection consiste alors à introduire tous les facteurs et à abandonner un par un ceux qui ne sont pas significatifs. On s'aperçoit alors qu'à chaque étape les coefficients et les degrés de significativité bougent, quand le facteur éliminé est associé (corrélé) aux autres. C'est une sélection descendante ("backward"), pas à pas ("stepwise").

Programme SAS associé :

```
proc logistic data=cours.aspi descending ;
class agecolinit sexe atcdpaden nbadeninit traitmt;
model aburden1an=traitmt;
run;
proc logistic data=cours.aspi descending;
class agecolinit sexe atcdpaden nbadeninit traitmt;
model aburden1an=traitmt agecolinit;
run;
proc logistic data=cours.aspi descending;
class agecolinit sexe atcdpaden nbadeninit traitmt;
model aburden1an=traitmt sexe;
run;
proc logistic data=cours.aspi descending;
```

```
class agecolinit sexe atcdpaden nbadeninit traitmt;
model aburden1an=traitmt atcdpaden;
run;
proc logistic data=cours.aspi descending;
class agecolinit sexe atcdpaden nbadeninit traitmt;
model aburden1an=traitmt nbadeninit;
run;
proc logistic data=cours.aspi descending;
class agecolinit sexe atcdpaden nbadeninit traitmt;
model aburden1an=age|traitmt|sexe|atcdpaden|nbadeninit
/selection=backward;
run;
```

9.2 Etude des associations (multiples variables)

Pour avoir une vue complète de l'ensemble des variables, sans privilégier la variable expliquée et la variable traitement, on peut faire une recherche des associations de toutes les variables par un modèle loglinéaire. Il n'apportera pas d'information supplémentaire en fin de compte. Néanmoins, si on se limite à des associations d'ordre pas trop grand les traits essentiels du modèle logistique se retrouvent sans calcul trop long avec le risque que cette limite de l'ordre conduise à des contradictions (souvent dénommées *erreurs de spécification*).

Par exemple, nous allons maintenant reprendre la récidive dans un cas moins simple que le précédent où la récidive est caractérisée par la présence ou non d'adénomes. Les variables en jeu sont les suivantes :

- nombre d'adémomes (récidive)au bout d'un an de traitement (2 classes)
- traitement (2 classes)
- sexe (2 classes)
- âge (3 classes)
- antécédents personnels (adénomes)
- nombre d'adénomes à la coloscopie initiale, avant le traitement (2 classes)
- indice de masse corporelle (4 classes)
- exposition au tabac (2 classes)

Pourquoi ne pas lancer un modèle logistique avec sélection des variables?

On dispose ici de 232 observations complétement renseignées, et la première régression comprend a priori 384 paramètres (tous croisements possibles). Le modèle est saturé mais les tests qui vont permettre d'éliminer les variables ne sont pas appliqués dans les

conditions normales (effectifs trop petits), et ne vont pas nécessairement conduire à un modèle intéressant. Il faut savoir choisir les seuils de sigificativité.

On préfère alors étudier les associations par grandes classes, pour ne retenir que les variables et les associations qui semblent les plus capables d'apporter une information sur la récidive.

Un modèle loglinéaire avec toutes les associations d'ordre trois est déjà saturé, un modèle loglinéaire avec les associations d'ordre 2 ne peut pas donner plus qu'une régression logistique avec des effets principaux sans interaction.

On doit donc se placer entre les deux, en partant du modèle loglinéaire avec toutes les interactions d'ordre 3 et en éliminant les associations d'ordre le plus élevé et non significatives ou bien en choisissant un mélange d'associations d'ordre 3 qui corresponde aux résultats attendus.

En somme, on doit effectuer un compromis entre une étude exhaustive des associations qui compliquent les calculs aussi bien en logistique qu'en loglinéaire. La sélection des variables, si elle repose sur la déviance, est une solution qui peut demander du temps (en calcul) mais qui évite des erreurs de spécification. Cependant on peut limiter l'exploration des associations quand on a de bonnes raisons d'exhiber une variable endogène, d'être assuré de l'orthogonalité d'un facteur (indépendant de tous les autres *dans l'échantillon*).

Programme SAS associé :

```
proc catmod data=cours.aspi;
model
traitmt*nbaden1an*nbadeninit*sexe*atcdpaden*agecolinit*bmi*tabagisme
= _response_/ nodesign noiter noparm noprofile noresponse;
loglin
traitmt|nbaden1an|nbadeninit|sexe|atcdpaden|bmi|
agecolinit|tabagisme @3;
run;
```

Exemple de choix d'associations a priori intéressantes

```
proc catmod data=cours.aspi;
model
traitmt*nbaden1an*nbadeninit*sexe*atcdpaden*agecolinit*bmi*tabagisme
= _response_/ nodesign noiter noparm noprofile noresponse;
loglin nbaden1an|traitmt|nbadeninit
nbaden1an|traitmt|sexe
nbaden1an|traitmt|atcdpaden
nbaden1an|traitmt|bmi
nbaden1an|traitmt|agecolinit
nbaden1an|traitmt|tabagisme;
run;
```

10 Exemple 7 : les étrangers de Paris

Dans cet exemple, l'échantillon est limité à quatre quartiers de Paris qui ont une population relativement homogène, et caractérisée par la dominance d'une catégorie socio-professionnelle. Celle-ci fait leur spécificité et elle peut suggérer une explication de l'évolution du nombre d'étrangers du recensement de 1990 au recensement 1999.

Une étude plus complète consisterait à comparer les évolutions des 80 quartiers et à mettre cette évolution en rapport avec leur structure socio-professionnelle, ne serait-ce que par une classification. Dans le cadre de cet exemple, les quatre quartiers choisis arbitrairement constituent un échantillon (exploratoire) de la population parisienne. Dans l'analyse, les poids des quartiers (nombre d'habitants) n'interviendront que dans le calcul des intervalles de confiance, mais pas dans celui des odds ratios du modèle logistique ou loglinéaire.

nom du quartier	code	nb habitants 90,99	catég. socio-professionnelle
Saint-Ambroise	1102	33476 - 32168	professions intermédiaires
Croulebarbe	1304	19967 - 19526	cadres actifs
Chaillot	1604	21757 - 21213	cadres retraités
Goutte-d'Or	1803	28226 - 28524	ouvriers et retraités

Le suivi des quartiers se fait sur la base des effectifs des habitants classés par leur nationalité (française, étrangère), leur genre (homme, femme) l'année du recensement. Dans cet exemple aucune causalité n'a de sens a priori, il s'agit donc d'étudier comment les variables sont associées.

10.1 Le modèle loglinéaire

Pour savoir comment s'est modifiée la répartition des habitants, de façon homogène ou non selon le quartier, la nationalité, le genre, on peut considérer qu'il y a une population classée par quatre critères (année, quartier, nationalité, genre), ce qui correspond à $2\times4\times2\times2 = 32$ sous-populations (ou 32 réponses possibles) pour les 204857 individus (habitants 90 et 99).

L'utilisation de la loi multinomiale sur l'ensemble, pour mesurer les homogénéités, les indépendances conditionnelles ou les associations éventuelles présente des inconvénients si on veut donner un sens (interpréter) les résultats, comme nous allons le voir.

L'étude de la répartition des habitants dans chaque quartier et sur deux années (soit 8 tableaux de contingences) semble assez raisonnable en faisant abstraction des mouvements migratoires dont on n'a pas une idée assez précise. Le regroupement des quatre quartiers pourrait constituer une sorte d'échantillon représentatif de la situation générale en 1990 et 1999 ; c'est une hypothèse de travail qui limite l'étude mais dont nous nous contenterons. Dans ces conditions, nous disposons d'un tableau de contingence pour chaque recensement (1990, 1999) et de 16 réponses avec les critères quartier, nationalité, genre.

Regrouper les deux échantillons (1990,1999) en un seul comme on le fait quand deux échantillons sont indépendants et qu'on désire les comparer, n'est pas possible car une grande partie des habitants d'un quartier en 1990 le sont encore dans ce quartier en 1999.

Et pourtant on obtient des résultats identiques, que l'on considère une population unique ou plusieurs sous-populations, dans l'analyse des dépendances entre les critères de classement. Cela vient de ce que les odds ratios ne dépendent pas des distributions marginales, mais cela vient aussi de ce qu'on utilise le modèle loglinéaire dans ses fonctionnalités de tester, voire mesurer, des dépendances et non dans celles d'estimer des probabilités. *Seules les probabilités conditionnelles et les odds ratios auront un sens.*

Le modèle loglinéaire aide à repérer les différences significatives quand on passe d'une sous-population à l'autre (tests des effets principaux ou des effets croisés) mais certains coefficients (paramètres) n'auront pas de sens ni de rapport avec notre objectif. Autrement dit, le modèle comportera des paramètres superflus.

Il arrive souvent dans un modèle loglinéaire que les résultats quantifiables intéressants (répondant aux objectifs de l'étude) soient des combinaisons de coefficients et non les coefficients individuellement.

Le modèle s'écrit :

$$Log(\widehat{n}_{aqng}/204857) = \lambda_a^A + \lambda_q^Q + \lambda_n^N + \lambda_g^G + \lambda_{aq}^{AQ} + ...\lambda_{aqn}^{AQN} + ... + \lambda_{aqng}^{AQNG}$$

avec des effets principaux :

a pour année de recensement
q pour quartier
n pour nationalité
g pour genre

puis des effets croisés d'ordre 2 comme AQ , des effets croisés d'ordre 3 comme AQN, et l'effet croisé des 4 variables.

Si tous les croisements figurent, le modèle est saturé.

Pour évaluer les dépendances, on part du modèle saturé et on élimine successivement les croisements qui ne sont pas significatifs (tests portant sur la déviance).

10.2 Analyse des associations

Commençons par le modèle saturé (31 paramètres libres). Les tests conduisent à garder des modèles assez complexes parce que la taille de l'échantillon est élevée.

On peut tout de même s'intéresser à des modèles plus simples. D'habitude une exploration rapide des croisements indispensables est faite en limitant successivement l'ordre des croisements. La déviance mesure l'écart entre les observations et les différents ajustements.

On obtient ici :

croisements (ordre max)	déviance par rapport au modèle saturé	nb de paramètres
4	0	31
3	5.19	28
2	204.87	18
1	8866.19	6

Dans cette phase exploratoire un modèle raisonnable semble être le modèle avec les croisements d'ordre 3, car l'écart au modèle saturé est très faible (p-value = 0.16), mais

un modèle très simple comme celui des croisements d'ordre 2 donne déjà une information non négligeable.

Exprimons en langage courant ce que signifie les trois modèles avec croisements d'ordre 1, 2 et 3.

10.2.1 modèle avec croisements d'ordre 1

La formulation du modèle est réduite à :

$$Log(\widehat{n}_{aqng}/204857) = \lambda_a^A + \lambda_q^Q + \lambda_n^N + \lambda_g^G$$

L'estimation d'un effectif se calcule comme un produit des distributions marginales, ce qui implique que les 4 facteurs interviennent de façon indépendante (il n'y a aucune association) :

quelque soit le quartier la population a évolué de la même façon
d'un recensement à l'autre la répartition n'a pas changé

Ce modèle peu intéressant est d'ailleurs rejeté par le test des déviances.

Si l'un des effets était non significatif, par exemple λ^G qui caractérise le genre, cela signifierait que la distribution est identique pour les hommes et pour les femmes.

10.2.2 modèle avec croisements d'ordre 2

Les 6 croisements possibles des 4 variables sont présents :

$$\begin{aligned} Log(\widehat{n}_{aqng}/204857) &= \lambda_a^A + \lambda_q^Q + \lambda_n^N + \lambda_g^G + \\ &\quad \lambda_{aq}^{AQ} + \lambda_{an}^{AN} + \lambda_{ag}^{AG} + \lambda_{qn}^{QN} + \lambda_{qg}^{QG} + \lambda_{ng}^{NG} \end{aligned}$$

La forme du modèle qui comporte les effets principaux et les effets croisés, avec les contraintes indispensables sur les coefficients, a l'avantage de mettre l'accent sur *ce qu'apporte chaque effet croisé,* comme information supplémentaire par rapport à tous les autres effets retenus dans le modèle.

Les seuls tests intéressants portent sur les effets croisés et non sur les les effets simples qui ne reflètent que des propriétés de distribution marginale. Ces apports de chaque croisement sont mesurés en termes de χ^2 (rapport de vraisemblance ou déviance).

croisement	d° de liberté	χ^2	p-value
année,quartier	3	41.50	<0.0001
année,nationalité	1	169.86	<0.0001
année, genre	1	5.46	0.0194
quartier, nationalité	3	6835.33	<0.0001
quartier, genre	3	415.72	<0.0001
nationalité, genre	1	425.38	<0.0001

Les chiffres de la troisième colonne parlent d'eux-mêmes : la présence d'étrangers varie beaucoup d'un quartier à l'autre, la proportion hommes/femmes (le genre) n'est pas la même pour les étrangers et les français, la proportion d'étrangers a varié entre 1990 et 1999. Par contre De 1990 à 1999 la proportion globale hommes/femmes n'a pas changé très significativement (p-value de 2%).

Les contraintes propres au modèle par la présence des seuls croisements d'ordre 2 se traduisent par une égalité de certaines distributions conditionnelles, et de certains odds ratios :

Ainsi l'association (quartier, nationalité) ne dépend pas des deux autres variables, année et genre ; la distribution du couple est identique quelque soit l'année et le genre. On dit encore que cette association est *homogène* pour l'année et le genre.

Ce qui se traduit aussi par des égalités d'odds ratios et qui veut à peu près dire que le rapport

$$\frac{chances\ de\ rencontrer\ un\ étranger\ à\ la\ Goutte\ d'Or\ plutôt\ qu'un\ français}{chances\ de\ rencontrer\ un\ étranger\ à\ Chaillot\ plutôt\ qu'un\ français}$$

était le même en 1990 et 1999, et que ce soit un homme ou une femme que l'on cherchait.

D'autre part accepter la nullité de l'association partielle (année, genre) comme le suggère le test s'exprime par une indépendance de l'année et du genre *conditionnellement* aux autres variables. Ainsi, dans chaque quartier et pour chaque nationalité, la proportion hommes/femmes n'a pas significativement changé entre 1990 et 1999.

Naturellement les conclusions tirées des tests sur l'influence des couple de variables sont eux-mêmes *conditionnels aux contraintes imposées au modèle.* L'acceptation d'un modèle plus complet, avec par exemple des croisements d'ordre 3 remet en cause les conclusions d'un modèle limité aux croisements d'ordre 2

10.2.3 modèles avec croisements d'ordre 3

On s'en tient aux indications données par le χ^2 du test de rapport de vraisemblance pour évaluer l'importance des croisements d'ordre 3

croisement	d° de liberté	χ^2	p-value
année, quartier, nationalité	3	10.62	0.0140
année, quartier, genre	3	3.60	0.3074
année, nationalité, genre	1	75.48	<0.0001
quartier, nationalité, genre	3	102.74	<0.0001

Il semble bien que les croisements (année, nationalité, genre) et (quartier, nationalité, genre) apportent une information non négligeable (qui met en cause le modèle précédent). Il faudra voir comment varie la proportion hommes/femmes, français et étrangers, d'un quartier à l'autre et d'une année à l'autre. Le changement de structure vient-il du vieillissement de la population traditionnelle ou du type d'emploi qui a attiré les immigrants?

Le modèle peut encore être simplifié en se passant des deux premiers croisements.

10.2.4 Modèle simplifié

Ce modèle comporte 22 coefficients libres. Les croisements d'ordre 3 (année, nationalité, genre) et (quartier, nationalité, genre) impliquent naturellement les 5 croisements d'ordre 2 (année, nationalité), (année, genre), (nationalité, genre), (quartier, nationalité) et (quartier, genre), mais pas le croisement (année, quartier). Il est donc naturel de résumer le modèle aux croisements de base :

croisements	d° de liberté	χ^2	p-value
année, quartier	3	42.48	<0.0001
année, nationalité, genre	1	80.91	<0.0001
quartier, nationalité, genre	3	104.72	<0.0001

Dans ce modèle l'association entre année,nationalité et genre est indépendante du quartier, comme l'association entre quartier, nationalité, genre est indépendante de l'année.

Une autre simplification doit être envisagée en regardant de plus près les modalités (en particulier les quartiers) qui se distinguent.

Il faut d'abord établir le tableau complet des coefficients en tenant compte des contraintes.

10.3 Evaluation des effets du dernier modèle

Coefficients des effets principaux :

variable	**modalité**	**coefficient**	**écart-type**
année	1990	0.0306	0.0029
	1999	(-0.0306)	
quartier	1102	0.2450	0.0049
	1304	-0.4890	0.0069
	1604	-0.0824	0.0053
	1803	(0.3264)	
nationalité	Etranger	-0.7625	0.0028
	Français	(0.7625)	
genre	femme	0.0187	0.0031
	homme	(-0.0187)	

Les coefficients entre parenthèses sont déduits des contraintes.

Les effets principaux rendent compte des effets marginaux :

la différence entre les années 1990 et 1999 (0.0612) correspond à une baisse générale du nombre d'habitants."en moyenne". Comme il s'agit en fait d'une moyenne géométrique, il n'est pas facile la lier aux variations totales ou moyennes (arithmétiques) des tris croisés. Elle est évaluée différemment selon les variables et les croisements figurant dans le modèle. Ce n'est donc pas une référence à une distribution marginale.

Le commentaire qualitatif des coefficients a tout de même un sens, mais pour une quantification des effets (calcul de odds ratio), il faut revenir à la formule du modèle.

Les effets les plus intéressants et qui justifient éventuellement des regroupements de modalités sont à rechercher sur les croisements de rang les plus élevés.

Coefficients du croisement (année, quartier)

année	**quartier**	**coefficient**	**écart-type**
90	1102	0.0124	0.0036
90	1304	0.0081	0.0043
90	1604	0.0029	0.0041
90	1803	(-0.0234)	
99	1102	(-0.0124)	
99	1304	(-0.0081)	
99	1604	(-0.0029)	
99	1803	(0.0234)	

Si on s'intéresse aux évolutions entre 90 et 99, ce sont les différences entre les coefficients 90 et 99 de chaque quartier qui montrera la spécificité de ces quartiers (effet croisé). Les tests associés portent sur la nullité d'un coefficient, ce qui signifie que la différence observée n'est pas significative d'une spécificité du quartier par rapport au comportement moyen sur la période. Il est ainsi facile à voir que ce sont les quartiers 1102 (Saint-Ambroise) et 1803 (Goutte-d'Or) qui se distinguent du comportement moyen, le second en particulier a le mieux résisté à la baisse générale de population alors que les deux autres sont dans la moyenne. Là encore la moyenne tient compte de tous les croisements et n'a pas un sens très facile à exprimer.

Coefficients du croisement (année, nationalité, genre)

annee	**nationalité**	**genre**	**coefficient**
90	Etranger	femme	-0.0251
90	Etranger	homme	(0.0251)
90	Français	femme	(0.0251)
90	Français	homme	(-0.0251)
99	Etranger	femme	(0.0251)
99	Etranger	homme	(-0.0251)
99	Français	femme	(-0.0251)
99	Français	homme	(0.0251)

Il n'y a en fait qu'un paramètre libre (qui est d'ailleurs significatif). Il s'interprète comme une évolution positive de 1990 à 1999 de la proportion d'étrangères (femmes) par rapport à la proportion d'étrangers (hommes), et ceci est indépendant du quartier choisi. Donc c'est vrai pour les quatre quartiers (on peut penser, par exemple, qu'un changement dans les métiers exercés par les étrangers ont éloigné les hommes de Paris).

Coefficients du croisement (quartier, nationalité, genre)

quartier	**nationalité**	**genre**	**coefficient**
1102	Etranger	femme	-0.0232
1304	Etranger	femme	-0.0090
1604	Etranger	femme	0.0503
1803	Etranger	femme	-0.0181
1102	Etranger	homme	(0.0232)
1304	Etranger	homme	(0.0090)
1604	Etranger	homme	(-0.0503)
...	...	...	...

La comparaison des coefficients *étranger-femme* et *étranger-homme*, dans le quartier 1102 (Saint-Ambroise) montre que la proportion de femmes étrangères est plus faible que celle des hommes, alors que c'est le contraire dans le quartier 1604 (Chaillot). Cette comparaison qui associe le genre et la nationalité doit être pensée comme un écart à la moyenne ou plus précisément comme une différence de répartition des étrangers par rapport aux français.

En bref, en 1999 *comme* en 1990 (indépendance de l'année et du croisement nationalité, genre *pour un quartier fixé*), dans certains quartiers la répartition des hommes et des femmes n'est pas la même pour les étrangers et les français. Si dans tous les quartiers on avait le même différence de répartition (association nationalité-genre homogène par rapport aux quartiers) l'association d'ordre 3 serait non significative.

On se rend bien compte qu'il n'est pas facile d'interpréter une association d'ordre 3.

10.4 Modèles logistiques associés

Pour y voir un peu plus clair et quantifier les effets, nous allons abandonner le modèle loglinéaire qui nous a signalé les associations intéressantes, c'est à dire les variables susceptibles d'interférer, pour en prendre une version particulière dont tous les résultats peuvent d'ailleurs se déduire du modèle loglinéaire. La particularité consiste à choisir une variable privilégiée, celle dont les chances (odds) nous intéressent le plus, la nationalité par exemple, qui est binaire.

Si nous calculons le rapport de la probabilité de trouver un étranger à la probabilité de trouver un français dans les différentes sous-populations, il suffit de faire la différence entre deux expressions du modèle loglinéaire avec :

$$n = E \; et \; n = F$$

$$\begin{aligned}
Log(\widehat{n}_{aqEg}/204857) &= \lambda^A_a + \lambda^Q_q + \lambda^N_E + \lambda^G_g + \lambda^{AQ}_{aq} + \lambda^{AN}_{aE} + \lambda^{AG}_{ag} + \lambda^{QN}_{qE} + \lambda^{QG}_{qg} + \lambda^{NG}_{Eg} \\
Log(\widehat{n}_{aqFg}/204857) &= \lambda^A_a + \lambda^Q_q + \lambda^N_F + \lambda^G_g + \lambda^{AQ}_{aq} + \lambda^{AN}_{aF} + \lambda^{AG}_{ag} + \lambda^{QN}_{qF} + \lambda^{QG}_{qg} + \lambda^{NG}_{Fg} \\
Log\frac{\widehat{n}_{aqEg}}{\widehat{n}_{aqFg}} &= (\lambda^N_E - \lambda^N_F) + (\lambda^{AN}_{aE} - \lambda^{AN}_{aF}) + (\lambda^{QN}_{qE} - \lambda^{QN}_{qF}) + (\lambda^{NG}_{Eg} - \lambda^{NG}_{Fg})
\end{aligned}$$

On reconnait un modèle logistique :

$$Log\frac{\widehat{n}_{aqEg}}{\widehat{n}_{aqFg}} = \alpha + \alpha_a^A + \alpha_q^Q + \alpha_g^G$$

Si on était parti du modèle loglinéaire avec tous les croisements d'ordre 3, on aurait obtenu :

$$Log\frac{\widehat{n}_{aqEg}}{\widehat{n}_{aqFg}} = \alpha + \alpha_a^A + \alpha_q^Q + \alpha_g^G + \alpha_{aq}^{AQ} + a_{ag}^{AG} + \alpha_{qg}^{QG}$$

Et le modèle simplifié devient :

$$Log\frac{\widehat{n}_{aqEg}}{\widehat{n}_{aqFg}} = \alpha + \alpha_a^A + \alpha_q^Q + \alpha_g^G + a_{ag}^{AG} + \alpha_{qg}^{QG}$$

Naturellement les effets et les tests sont les mêmes, mais le modèle se comprend mieux.

Tous les modèles logistiques sont des cas particuliers de modèles loglinéaires, avec des méthodes de calcul qui peuvent un peu différer d'un logiciel à l'autre.

Programme SAS associé :

```
data w;
set cours.paris(where=(quartier in ('1803' '1604' '1102' '1304')));
proc freq data=w;
weight effectif;
tables quartier*annee quartier*annee*nationalite*genre;run;
proc catmod data=w;
weight effectif;
model annee*quartier*nationalite*genre=_response_
    / noiter noprofile noresponse;
loglin annee|quartier|nationalite|genre ;
run;
proc catmod data=w;
weight effectif;
model annee*quartier*nationalite*genre=_response_
     / noiter noprofile noresponse;
loglin annee|quartier|nationalite|genre @3;
run;
proc catmod data=w;
weight effectif;
model annee*quartier*nationalite*genre=_response_
     / noiter noprofile noresponse;
loglin annee|quartier|nationalite|genre @2;
run;
proc catmod data=w;
weight effectif;
model annee*quartier*nationalite*genre=_response_
```

```
     / noiter noprofile noresponse;
loglin annee|quartier|nationalite|genre @1;
run;
proc catmod data=w;
weight effectif;
model annee*quartier*nationalite*genre=_response_
     / noiter noprofile noresponse pred=freq;
loglin annee|quartier|nationalite|genre @2
 annee*nationalite*genre quartier*nationalite*genre;
run;
proc logistic data=w;
weight effectif;
class annee quartier genre;
model nationalite=annee|quartier|genre @1;
run;
proc logistic data=w;
weight effectif;
class annee quartier genre;
model nationalite=annee|quartier|genre @2;
run;
proc logistic data=w;
weight effectif;
class annee quartier genre;
model nationalite= annee|genre quartier|genre;
run;
```

11 Exemple 8 : les actifs résidant à Paris

La table que nous allons étudier est de grande taille par le nombre d'individus (26066), par le nombre de variables de classification et leurs modalités, d'ailleurs très inférieur au nombre réel dans le fichier original ; mais cela suffit pour montrer les difficultés qui en résultent :

- arrondissements : 2
- âge : 4 classes limitées
- diplôme : 8 classes
- catégories professionnelles : 5 classes
- distances entre lieu de résidence et lieu de travail : 10 classes

Pour l'âge et les catégories professionnelles on dispose déjà de deux tableaux de contingences avec 20 cellules dont on peut analyser les distributions marginales, ou ligne par ligne, ou colonne par colonne.

5ème Arrdt	cpis	emp	ouv	pic	proi	ensemble
20-29 ans	649	441	130	42	410	1672
30-39 ans	974	347	112	114	406	1953
40-49 ans	892	292	125	193	325	1827
50-59ans	535	233	106	139	211	1224
ensemble	3030	1313	473	488	1352	6676

13ème Arrdt	cpis	emp	ouv	pic	proi	ensemble
20-29 ans	1102	1835	627	95	1271	4930
30-39 ans	1865	1621	761	260	1412	5919
40-49 ans	1794	1198	677	330	1181	5180
50-59ans	1033	813	545	254	716	3361
ensemble	5794	5467	2610	939	4580	19390

Pour repérer complètement les associations, il faudra comparer pour les deux arrondissements sélectionnés deux tableaux de contingences comprenant chacun $4 \times 8 \times 5 \times 10 = 1600$ cellules, soit 3200 cellules où se répartissent 26066 individus, ce qui fait en moyenne 8 individus par case. Il est donc probable que de nombreuses cellules vont être vides, ce qui empêchera d'estimer certains rapports de chances, en essayant, par exemple, de construire un modèle avec toutes les variables. Ces cellules vides constituent un obstacle à certains calculs, des $Log(n_{ijkl..})$ dans l'expression du modèle loglinéaire, de la vraisemblance, etc.

Lorsqu'il y a des cellules vides, on se demande si l'absence d'individus dans ces cellules est le résultat d'un aléa, ou s'il s'agit d'une contrainte structurelle (un zéro théorique) qui modifie alors le nombre de degrés de liberté du modèle et qui impose un traitement

un peu différent de la modélisation. D'autre part, même s'il ne s'agit que d'un aléa, il peut s'introduire dans le modèle un lien entre les variables de classification, qui deviennent colinéaires et qui, comme dans les modèles de régression linéaire, rendent le modèle non identifiable (infinité de solutions pour certains coefficients au maximum de vraisemblance).

En bref, sauf en cas de nullité structurelle des effectifs de certaines classes, il vaut mieux éviter les cellules vides.

Les logiciels s'en débarrassent chacun à leur façon, mais la méthode la plus simple et conforme à un objectif d'analyse de la population globale sera d'opérer un regroupement de classes.

11.1 Les types d'associations

La taille du tableau de contingence nous interdit pratiquement d'évaluer le modèle saturé (3200 paramètres), ce qui suggère de travailler en laissant de côté une variable de classification : par exemple, la distance domicile-travail. Le nombre de classes est alors réduit à 320 classes. Dans ces conditions les estimations portent sur des classes agrégées. Une association entre classes agrégées est parfois appelée *marginale* (répartition marginale des effectifs des classes actives).

D'une façon générale, les associations qui ne comportent pas toutes les variables de classements sont marginales.

Une association *complète* correspond au modèle saturé, ce qui s'exprime par un modèle qu'on peut décrire avec quatre variables de classement A, B, C, D par

$$Log(n_{abcd}) = \lambda_{abcd}^{ABCD}$$

a,b,c,d désignent les modalités respectives des variables ABCD

Un modèle comportant des associations *partielles* sera par exemple

$$Log(n_{abcd}) = \lambda_{abc}^{ABC} + \lambda_{abd}^{ABD}$$

où les quatre variables figurent, mais pas associées simultanément.

Un modèle avec associations marginales (absence de la variable D) sera

$$Log(n_{abcd}) = \lambda_{abc}^{ABC}$$

ou bien

$$Log(n_{abcd}) = \lambda_{ab}^{AB} + \lambda_{ac}^{AC}$$

On peut aussi exprimer la non association des variables en termes d'*indépendance* (à ne pas confondre avec la notion de paramètre "indépendant" ou *libre*, ni avec l'opposition entre variable dépendante, expliquée, réponse, et variable indépendante, explicative, de classement).

Si les associations de variables apparaissent seulement à l'ordre 1 dans le modèle, il s'agira d'indépendance de ces variables. On dit aussi qu'une variable ou un groupe de variable est indépendant d'une autre variable ou d'un groupe comme dans le modèle

$$Log(n_{abcd}) = \lambda_{ab}^{AB} + \lambda_{cd}^{CD}$$

où le groupe A,B est indépendant du groupe C,D ; alors que dans le modèle

$$Log(n_{abcd}) = \lambda_{ab}^{AB} + \lambda_{ac}^{AC}$$

on dit que, *conditionnellement* à A (ou pour chaque valeur de A, ou à A fixé), B et C sont indépendants.

Les indépendances conditionnelles, par groupes ou individuelles, se traduisent par des égalités de odds ratios. Une indépendance non conditionnelle se traduit par l'égalité à 1 de certains odds ratios (effectifs proportionnels).

L'absence d'une variable dans un modèle signifie que cette variable n'intervient pas dans le tableau de contingence et donc que les classes correspondantes ont été agrégées (collapsed).

Si la variable C est absente, le regroupement des effectifs pour chaque couple de modalités de A,B ne modifie pas les résultats. C'est vrai aussi en cas d'indépendance conditionnelle. Ainsi dans l'exemple précédent une agrégation des classes selon C ne modifie pas les odds ratios de l'association A,B (ce qui se vérifie immédiatement par un calcul direct des odds ratios en fonction des paramètres du modèle loglinéaire).

De nombreux modèles peuvent être testés pour mettre en évidence une propriété d'indépendance ou d'association, conditionnelle ou non.

Une remarque suffira pour montrer sur un exemple simple que certains résultats peuvent se présenter différemment selon que l'on prend ou non en compte une variable absente dans le modèle loglinéaire.

Considérons simplement les trois variables de classement LR (lieu de résidence), AGE, DT (distance domicile-travail, soit $2 \times 4 \times 10 = 80$ classes. Le modèle loglinéaire est réduit aux seules variables LR et AGE

$$Log(n_{lad}) = \lambda_{l}^{LR} + \lambda_{a}^{AGE} + \lambda_{la}^{LR \times AGE}$$

qui comporte 8 paramètres indépendants à estimer.

Comme il y a 80 réponses possibles, le modèle n'est pas saturé, les estimations des effectifs n_{lad} seront égales pour les classes (l,a,d) qui ont (l,a) en commun (il y en a 10). L'association entre LR et AGE est marginale relativement à DT.

Reprenons le modèle en ne considérant que les huit classes agrégées. Le modèle s'écrit

$$Log(n_{la}) = \lambda_{l}^{LR} + \lambda_{a}^{AGE} + \lambda_{la}^{LR \times AGE}$$

(l'indice d a disparu). Le modèle est saturé, mais les résultats (estimations, tests) sont identiques, bien que les vraisemblances, ou déviances, soient différentes.

Programme SAS associé

```
proc freq data=cours.actifsred;
tables lr*age*csp;
run;
proc catmod data=cours.actifsred;
model lr*age*dt=_response_ /noprofile nodesign noresponse pred=freq;
loglin lr|age;
run;
proc catmod data=cours.actifsred;
model lr*age=_response_ /noprofile nodesign noresponse pred=freq;
loglin lr|age;
run;
proc catmod data=cours.actifsred;
model lr=age / noprofile nodesign noresponse pred=freq;
run;
```

11.2 Recherche des associations significatives

Même lorsque l'objectif final n'est pas la recherche de toutes les associations, et qu'on veut aboutir à une forme logistique avec beaucoup de variables de classement, il peut être intéressant d'examiner les associations de façon générale.

C'est ce qui va être fait d'abord, pour apprécier des associations ne faisant pas intervenir la variable LR mais qui apportent une information sur des liens analogues à la multicollinéarité des variables explicatives d'une régression linéaire.

La difficulté essentielle vient de ce que malgré l'effectif important de la population, la multiplicité des classes conduit à des classes vides qui sont prises pour la plupart des logiciels pour des contraintes structurelles. Si la méthode d'estimation est celle des moindres carrés généralisés ces cellules doivent être remplies artificiellement (par une option particulière dans les logiciels, plus ou moins automatique).

Le modèle saturé avec les cinq variables a 3200 paramètres. Si on se limite aux associations d'ordre 4 il y a 2144 paramètres et dans notre exemple il y a seulement 1282 classes non vides, ce qui implique des contraintes sur les paramètres.

Avec les seules associations d'ordre 3, il y a 1055 paramètres (même si le programme de maximisation est bien fait, la fonction à maximiser a 1055 variables, ce qui peut donner des doutes sur la validité des calculs numériques).

En regroupant en une seule classe les distances dépassant 25 km il reste 691 paramètres indépendants et les résultats peuvent commencer à être exploités : on ne retiendra que les tests portant sur les croisements d'ordre le plus élevé, 3.

Le test d'ajustement global (comparaison au modèle saturé) dont la statistique G^2 (qui suit un χ^2 à 536 degrés de liberté), est égale à *417,* montre le peu d'écart entre le modèle

saturé et le modèle contraint ; ce qui signifie qu'il n'est pas nécessaire de chercher des associations d'ordre supérieur à 3.

Si on se limite aux associations d'ordre 3, on obtient :

associations	**d° de liberté**	χ^2	**p-value**
LR*AGE*DIPL	21	19.16	0.58
LR*AGE*CSP	12	17.16	0.14
LR*AGE*DT	15	16.38	0.36
LR*DIPL*CSP	28	114.32	<0.001
LR*DIPL*DT	33	34.19	0.41
LR*CSP*DT	20	43.22	0.002
AGE*DIPL*CSP	84	366.15	<0.001
AGE*DIPL*DT	104	122.32	0.11
AGE*CSP*DT	58	53.28	0.65
DIPL*CSP*DT	117	*non calculé*	

(les degrés de libertés tiennent compte des comtraintes imposées par les cellules vides)

Les associations d'ordre 3 les plus significativement différentes de 0 sont :

LR*AGE*CSP
LR*CSP*DT
AGE*DIPL*CSP

On peut alors exclure les autres et engager une élimination successive des croisements non significatifs.

Une autre démarche consiste à se limiter à quatre variables mais en regroupant les données, de façon à supprimer le découpage par la variable DT. On travaille alors sur des répartitions marginales, ce qui peut modifier la significativité des autres associations.

Dans le modèle saturé à 4 variables, l'association d'ordre 4 n'est pas significativement différente de 0. On passe alors au modèle limité aux associations d'ordre 3 (236 paramètres).

Toutes les associations d'ordre 3 sont significatives (p-value < 0.001), sauf l'association LR*AGE*DIPL dont la statistique mesurant son apport suit un χ^2 à 21 degrés de liberté et vaut 22.43 (p-value = 0.38).

On peut donc éliminer cette association pour garder finalement :

associations	**d° de liberté**	**chi2**	**p-value**
LR	1	1874.10	<0.001
AGE	3	292.49	<0.001
DIPL	7	1506.92	<0.001
CSP	4	1894.80	<0.001
LR*AGE	3	23.19	<0.001
LR*DIPL	7	182.10	<0.001
LR*CSP	4	86.65	<0.001
AGE*DIPL	21	1548.69	<0.001
AGE*CSP	12	723.62	<0.001
DIPL*CSP	28	7626.98	<0.001
LR*AGE*CSP	12	20.85	0.053
LR*DIPL*CSP	28	99.55	<0.001
écart au modèle saturé G^2	186	590.92	<0.001

La somme des degrés de liberté est 316, soit un de moins que le nombre total de cellules non vides, car on n'a pas reporté la constante du modèle, ce qui revient à s'intéresser aux répartitions plutôt qu'aux effectifs mais qui ne modifie pas les tests.

La qualité de l'ajustement est mesurée par la statistique du test du rapport de vraisemblance G^2 (dernière ligne) qui compare le modèle courant au modèle saturé ; c'est donc la déviance.

Les associations les plus intéressantes sont celles d'ordre le plus élevé c'est-à-dire :

LR*AGE*CSP et LR*DIPL*CSP à l'ordre 3, et AGE*DIPL qui n'est pas impliqué par les deux précédentes. Les autres associations, qui pourraient d'ailleurs ne pas être significatives sans mettre en cause le modèle, sont des associations marginales.

La mesure d'une association d'ordre élevé est modélisée comme "écart aux marginales". C'est à partir d'elles qu'on peut décrire (interpréter) certains aspects du modèle.

Si on introduit un objectif plus particulier, comme de considérer que c'est la distance LR qui constitue une réponse (ou variable expliquée), les odds ratios porteront toujours sur LR et seules les associations de LR restent utiles. Dès lors on peut travailler avec des régressions logistiques sans expliciter (ni quantifier), toutes les associations secondaires non intéressantes pour décrire ou expliquer LR.

En bref, nous avons exploré les associations entre les variables de classement avec les tests de significativité des associations afin de préparer le travail sur le modèle logistique de la section suivante. Nous n'avons pas cherché à expliquer comment les variables étaient associées, aucune quantification (odds ratios) n'a été explicitée.

Programmes associés

```
proc format;
value $dist '025-029 kms'='>24 kms'
 '030-034 kms'='>24 kms'
 '035-039 kms'='>24 kms'
```

```
 '040-044 kms'='>24 kms'
 '045-049 kms'='>24 kms';
run;
proc catmod data=cours.actifsred;
format dt $dist.  ;
model lr*age*dipl*csp*dt=_response_
/noprofile noiter nodesign noresponse noparm;
loglin lr|age|dipl|csp|dt @3;
run;
proc catmod data=cours.actifsred;
model lr*age*dipl*csp=_response_
   /noprofile noiter nodesign noresponse noparm;
loglin lr|age|dipl|csp ;
run;
proc catmod data=cours.actifsred;
model lr*age*dipl*csp=_response_
   /noprofile noiter nodesign noresponse noparm;
loglin lr|age|dipl|csp @3;
run;
proc catmod data=cours.actifsred;
model lr*age*dipl*csp=_response_
   /noprofile noiter nodesign noresponse noparm ;
loglin lr|age|dipl|csp @2 lr*dipl*csp lr*age*csp;
run;
```

11.3 Régressions logistiques

L'étude des associations nous a conduit à ne retenir que les variables AGE, DIPL, CSP susceptibles de décrire la distribution de LR, avec les interactions AGE×DIPL et DIPL×CSP. Le modèle logistique (binaire) retenu est donc :

$$Log\left(\frac{p_{adc}(13\grave{e}me)}{p_{adc}(5\grave{e}me)}\right) = \lambda + \lambda_a^{AGE} + \lambda_d^{DIPL} + \lambda_c^{CSP} + \lambda_{ac}^{AGE\times CSP} + \lambda_{dc}^{DIPL\times CSP}$$

a, d, s sont les modalités respectives de $AGE, DIPL, CSP$

Le nombre de paramètres indépendants est 54. Les coefficients (et les tests) pourraient se déduire du modèle loglinéaire écrit sous forme développée (avec contraintes)

$$\begin{aligned} Log(n_{ladc}) \;=\; & \lambda + \lambda_l^{LR} + \lambda_a^{AGE} + \lambda_d^{DIPL} + \lambda_c^{CSP} \\ & + \lambda_{la}^{LR\times AGE} + \lambda_{ld}^{LR\times DIPL} + \lambda_{lc}^{LR\times CSP} + \lambda_{ad}^{AGE\times DIPL} + \lambda_{ac}^{AGE\times CSP} + \lambda_{dc}^{DIPL\times CSP} \\ & + \lambda_{lac}^{LR\times AGE\times CSP} + \lambda_{ldc}^{LR\times DIPL\times CSP} \end{aligned}$$

Les calculs numériques fournis par les deux modèles peuvent contenir de petites différences qui proviennent du calcul itératif du *maximum* de vraisemblance, bien qu'ils doivent donner "théoriquement" les mêmes résultats.

Avec moins de paramètres (27) les deux modéles simplifiés suivants présenteront des résultats numériques parfaitement identiques :

$$Log\left(\frac{p_{ac}(13\grave{e}me)}{p_{ac}(5\grave{e}me)}\right) = \lambda + \lambda_a^{AGE} + \lambda_c^{CSP}$$

$$Log(n_{lac}) = \lambda + \lambda_l^{LR} + \lambda_a^{AGE} + \lambda_c^{CSP} + \lambda_{la}^{LR\times AGE} + {}_{lc}^{LR\times CSP} + \lambda_{ac}^{AGE\times CSP}$$

Le modèle logistique n'est pas équivalent à un modèle loglinéaire où on aurait supprimé le croisement $AGE \times CSP$, ce qui aurait introduit la contrainte $\lambda_{ac}^{AGE\times CSP} = 0$.

Ce qui montre que les modèles loglinéaires présentent des singularités qui ne peuvent pas être traduites dans un modèle logistique. Même si on ne s'intéresse qu'à certaines associations (ici, LR décrite par AGE et CSP), il faut considérer les associations dans leur ensemble (le lien AGE, CSP intervient dans les relations entre LR et AGE, LR et CSP). Autrement dit, le modèle logistique précédent, bien qu'il ne fasse intervenir que deux variables sans interaction, ne se réduit pas à des associations marginales. Cependant si l'association $AGE \times CSP$ était nulle, la régression logistique sur AGE et CSP donnerait les mêmes résultats que les régressions logistiques sur AGE seule, et sur CSP seule (comme dans les régressions linéaires où les exogènes sont orthogonales).

Programme associé :

```
proc logistic data=cours.actifsred;
class age dipl csp;
model lr=age dipl csp age*csp dipl*csp ;
run;
proc catmod data=cours.actifsred;
model lr=age dipl csp age*csp dipl*csp
  /noiter nodesign noresponse noprofile;
run;
proc catmod data=cours.actifsred;
model lr*age*csp*dipl=_response_
  /noiter nodesign noparm noresponse noprofile;
loglin lr|csp|age lr|dipl|csp csp|dipl|age;
run;
proc catmod data=cours.actifsred;
model lr*age*csp*dipl=_response_
  /noiter nodesign noparm noresponse noprofile;
loglin lr|csp|age lr|dipl|csp ;
run;
proc catmod data=cours.actifsred;

model lr*csp*age=_response_
  /pred=prob noiter nodesign noresponse noprofile;
loglin lr|csp lr|age csp|age;
```

```
run;
proc catmod data=cours.actifsred;
model lr*csp*age=_response_
  /pred=prob noiter nodesign noresponse noprofile;
loglin lr|csp lr|age ;
run;
proc catmod data=cours.actifsred;
model lr=csp age /pred=prob noiter noprofile;
run;
```

Pour éviter l'étude détaillée des associations et quand on désire aboutir à une régression logistique, certains logiciels proposent des procédures de choix de modèle pas à pas (stepwise) par sélection des variables simples ou croisées. En particulier s'il y a beaucoup de variables on peut essayer de partir du modèle le plus simple en ajoutant les interactions successives (méthode dite *forward*). Si on part du modèle le plus complexe, on supprimera les interactions non significatives (méthode dite *backward*). Dans ce dernier cas les calculs peuvent être assez lourds, mais on reste dans une cohérence théorique rassurante (tests non biaisés quelque soit l'étape).

Programme associé :

```
proc logistic data=cours.actifsred;
class age dipl csp ;
model lr=age|dipl|csp /selection=backward;
run;
proc logistic data=cours.actifsred;
class age dipl csp ;
model lr=age|dipl|csp /selection=forward;
run;
proc logistic data=cours.actifsred;
format dt $dist.  ;
class age dipl csp dt;
model lr=age|dipl|csp|dt /selection=backward;
run;
proc logistic data=cours.actifsred;
format dt $dist.  ;
class age dipl csp dt;
model lr=age|dipl|csp|dt /selection=forward;
run;
proc logistic data=cours.actifsred;
format dt $dist.  ;
class age dipl csp dt;
model lr=age|dipl|csp|dt @2/selection=backward;
run;
```

11.4 Introduction d'une variable quantitative

Lorsque des variables descriptives ou explicatives d'une régression logistique ont de nombreuses modalités, il est quelquefois possible de les regrouper sans changer les résultats, ce qu'on peut tester directement (égalité de paramètres). Mais on peut aussi, *recoder* une variable ordinale, par des nombres traduisant les écarts entre les classes. Ce recodage va simplifier les résultats puisque les coefficients à estimer pour chaque modalité sont remplacés par un coefficient unique. Ce qui introduit une contrainte puisque les valeurs correspondant à chaque modalité sont proportionnelles aux valeurs de la variable recodée. A fortiori une variable quantitative peut être introduite dans le modèle, directement, mais il faut bien saisir comment comment la contrainte s'exprime.

Considérons la variable D *distance domicile-travail* qui aurait pu être prise directement (c'est d'ailleurs cette distance qui est calculée dans le fichier original, puis découpée en classes par la suite).

Essayons d'expliquer ce que signifie la distance comme variable quantitative, dans un modèle élémentaire comme le suivant

$$Log\left(\frac{p_{acD}(13ème)}{p_{acD}(5ème)}\right) = \lambda + \lambda_a^{AGE} + \lambda_c^{CSP} + \mu D$$

μ est le coefficient attribué à la variable D représentant la distance domicile-travail

D'abord, quelle que soit la sous-population caractérisée par un âge et une CSP, le facteur *distance* joue de la même façon sur le rapport des probabilités ou les chances "13ème contre 5èeme". Et cet effet est *entièrement défini* par le coefficient μ qui s'exprime alors comme une variation de chances (de cotes) quand la distance varie d'un km.

Quels que soient l'âge et la catégorie socio-professionnelle,une variation de D de 1 km multiplie par μ le rapport

$$\frac{p_{acD}(13ème)}{p_{acD}(5ème)}$$

soit en termes de odds ratios :

$$\mu = \frac{p_{ac(D+1)}(13ème)}{p_{ac(D+1)}(5ème)} \Big/ \frac{p_{acD}(13ème)}{p_{acD}(5ème)}$$

Ce qui implique une identité des odds ratios quelque soit l'âge et la CSP, mais surtout une *homogénéïté* des odds ratios relativement à la distance.

Si on ajoute une interaction âge×distance (ce qui voudrait dire que selon l'âge les effets de la distance sont différents), l'homogénéïté est toujours là, mais avec un coefficient qui dépend de la classe d'âge.

Dans le cas particulier étudié, les classes de distances ont été construites en tenant compte des implications possibles en temps de trajet, et l'effet n'est pas nécessairement

proportionnel. Notamment à partir de 25 km il y a probablement des problèmes spécifiques impliquant le mode de transport. Nous nous en sommes tenus à un modèle très grossier en recodant simplement une classe de distance par le milieu de classe.

Ce qui conduit au modèle

$$Log\left(\frac{p_{adcD}(13ème)}{p_{adcD}(5ème)}\right) = \lambda + \lambda_a^{AGE} + \lambda_d^{DIPL} + \lambda_c^{CSP} + \lambda_{dc}^{DIPL\times CSP} + \mu D + \mu_d D + \mu_c D$$

avec les effets suivants :

facteurs	**d° de liberté**	**chi2**	**p-value**
AGE	3	21.13	<0.001
DIPL	7	85.91	<0.001
CSP	4	83.44	<0.001
DIPL*CSP	28	99.21	<0.001
D	1	90.67	<0.001
D*DIPL	7	20.84	0.004
D*CSP	4	36.27	<0.001

Le modèle ne comporte pas d'effet croisé $D \times DIPL \times CSP$. La distance a un effet homogène à l'intérieur de chaque sous-population caractérisée par un diplôme et une catégorie socio-professionnelle, mais le facteur est différent d'une sous-population à l'autre, ce qui se traduit par la multiplicité des coefficients en μ.

Le coefficient μ représente une pente moyenne, les coefficients μ_d et μ_c sont les écarts de pente associés à chaque diplôme d et chaque catégorie socio-professionnelle c. Les effets de D sur le rapport des effectifs (13ème contre 5ème) sont donc exponentiels avec un taux dépendant additivement de CSP et de $DIPL$.

Si on complète le modèle avec une interaction $D \times DIPL \times CSP$ pour chaque couple d, c le taux de variation par la distance est différent.

facteurs	**d° de liberté**	**chi2**	**p-value**
AGE	3	21.25	<0.001
DIPL	7	44.70	<0.001
CSP	4	54.97	<0.001
DIPL*CSP	28	96.97	<0.001
D	1	0.040	<0.84
D*DIPL	7	13.55	0.006
D*CSP	4	32.17	<0.001
D*DIPL*CSP	28	61.73	<0.001

Les tests portant sur les croisements de niveau inférieur ne sont pas simples à comprendre parce qu'ils représentent un effet marginal qui tient compte des contraintes du modèle.

Ainsi, supprimer l'effet marginal exprimé par la variable D isolée (ce qui ne modifie pas notablement les résultats de l'estimation) modifie les statistiques mesurant l'effet de chaque interaction :

facteurs	**d° de liberté**	**chi2**	**p-value**
AGE	3	20.73	<0.001
DIPL	7	30.08	<0.001
CSP	4	60.31	<0.001
DIPL*CSP	28	103.37	<0.001
D*DIPL	7	27.36	<0.001
D*CSP	4	65.37	<0.001
D*DIPL*CSP	28	82.38	<0.001

Les seuls tests qui sont faciles à poser et utiles pour la sélection du modèle sont ceux qui portent sur les interactions de niveau le plus élevé, l'effet d'une interaction se mesurant comme un écart à la présence de toutes les interactions inférieures (hypothèse nulle). C'est pourquoi les programmes de sélection pas à pas garderont systématiquement toutes les interactions d'ordre inférieur à l'interaction d'ordre le plus élevé

Pour le modèle précédent, la significativité de l'interaction $D \times DIPL \times CSP$ impliquera donc la présence des interactions inférieures : $D, DIPL, CSP, D \times DIPL, D \times CSP, DIPL \times CSP$

Programmes associés :

```
data w ;set cours.actifsred;
distance=input(substr(dt,5,3),3.0)-1.5;
run;
proc freq data=w;tables distance;run;
proc logistic data=w;
class age dipl csp ;
model lr=age csp distance;
run;
proc logistic data=w;
class age dipl csp ;
model lr=age|dipl|csp|distance @3/selection=backward;
run;
```

12 Exemple 9 : modèle logistique et score

Un score est un nombre résumant l'information dont on dispose sur un individu, de façon simple, voire simpliste. Il est construit généralement à partir d'une analyse multivariée (variables continues ou discrètes) dans le but de classer des individus. Dans le cas présent, il y a deux classes d'individus déterminées a priori, et le score doit donner une indication sur la proximité d'un individu quelconque à l'une ou l'autre classe. Par exemple :

- le BMI (Body Mass Index) résume à partir du poids et de la taille d'un individu son "état général" (en fait le degré d'obésité), il sert à détecter les individus "à risques" pour certaines maladies, mais il est aussi utilisé comme variable explicative du risque.
- le taux de cholestérol (en particulier le rapport L.D.L./H.D.L) sert à évaluer les risques d'accidents cardiovasculaires.
- le patrimoine,la situation familiale, les revenus, etc, sont résumés par un score qui est évalué par un organisme de crédit pour accorder ou non un prêt.
- La vente par correspondance (crédits accordés), l'assurance auto (type de contrat proposé), le secteur bancaire (type de placement proposé) se servent de scores pour évaluer leurs clients.

D'une façon générale le score est une formule qui sert à distinguer (discriminer) les individus dans le but de les classer relativement au risque qu'ils prennent (ou que le "scoreur" prend).

Les formules peuvent être calculées de nombreuses façons (analyse discriminante) et l'utilisation du modèle logistique en est un exemple, avec les particularités suivantes :

- Des variables discrètes ou continues, *signalétiques*, sont connues pour chaque individu qui se présente (client).
- L'objectif est de classer les individus dans deux catégories (bon, mauvais).
- La formule qui note l'individu est souvent une fonction croissante d'une combinaison linéaire de variables continues et de fonctions indicatrices des modalités pour les variables discrètes.
- Les individus se caractériseront (seront discriminés) par un rapport qui exprimera leurs chances (odds) d'appartenir aux bons contre celles d'appartenir aux mauvais. On en déduira, pour une population donnée, la probabilité qu'un individu soit bon ou mauvais, compte tenu des valeurs connues de ses variables signalétiques.
- On choisira une valeur seuil du score (cut-off) qui séparera les bons potentiels des mauvais potentiels.

  On calculera pour chaque seuil (scénario) les risques de se tromper et les coûts entraînés par la décision de rejeter (ou considérer comme mauvais) tous les individus dont le score est inférieur au seuil et d'accepter les autres.

Le coût d'une décision peut s'exprimer en terme de probabilité des erreurs (bons considérés comme mauvais, mauvais considérés comme bons), avec éventuellement des coûts propres à chaque type d'erreur et des gains associés à des choix corrects.

Le coût global dépend du score, du seuil choisi mais aussi de la distribution des futurs clients.

D'où la démarche :

1. La régression logistique va servir à construire le score à partir d'un échantillon d'individus parfaitement connu (on sait s'ils sont bons ou mauvais).
2. Sur cet échantillon, on cherchera pour chaque seuil les probabilités correspondantes, puis éventuellement les coûts associés. Mais pour aboutir à une règle de décision conforme aux intérêts du décideur il faudra évaluer les coûts pour la population totale des clients.

12.1 Construction du score

Dans cette étape on se limite à construire la meilleure régression logistique possible, et donc le score le plus discriminant. C'est sur les coûts associés, que dans la deuxième étape on choisira le seuil.

L'exemple sur lequel nous travaillons a été considérablement simplifié par rapport au problème original pour mieux insister sur les éléments d'évaluation du score.

L'échantillon de 7680 individus à qui un prêt a été déjà accordé, se décompose en deux sous-populations : une sous-population de 3680 individus fortement endettés (incapables de faire face à leurs mensualités) et une autre de 4000 individus faiblement endettés (qui sont solvables).

La réponse est binaire (non solvable, solvable) ou (mauvais, bon client), les variables de classement sont la situation familiale $F \times E$ réduite à quatre classes : F, personne vivant seule ou en couple, et E, ayant ou non des enfants.

Le modèle retenu s'écrit :

$$Log\frac{p_{ij}}{1-p_{ij}} = \lambda + \lambda_i^F + \lambda_j^E + \lambda_{ij}^{F\times E}$$

p_{ij} est la probabilité de non remboursement, celle qui intéresse (inquiète) le décideur.

La fonction score est :

$$SC(i,j) = \lambda + \lambda_i^F + \lambda_j^E + \lambda_{ij}^{F\times E}$$

Les valeurs élevées de cette fonction correspondent à une cote élevée (non remboursement contre remboursement normal) et donc à un risque élevé de non remboursement. Ce sont les valeurs élevées du score qui serviront à détecter les mauvais payeurs.

L'estimation par le maximum de vraisemblance donne :

variable	modalité	estimation	écart-type	p-value
constante λ		-0.0814	0.0275	0.0031
effet principal λ^F	couple	-0.0814	0.0275	0.0031
	seul	0.0814	0.0275	0.0031
effet principal λ^E	non	-0.2167	0.0275	<0.001
	oui	0.2167	0.0275	<0.001
effets croisés $\lambda^{F\times E}$	couple,non	0.0832	0.0275	0.0025
	couple,oui	-0.0832	0.0275	0.0025
	seul,non	-0.0832	0.0275	0.0025
	seul,oui	0.0832	0.0275	0.0025

12.1.1 Score déduit de la logistique

Nous retenons que la régression logistique est convenable et en déduisons le score qui donne pour les quatre types d'individus repérables par leur situation familiale :

F	E	score	probabilité associée
couple	non	-0.463	0.386
seul	non	-0.134	0.467
seul	oui	0.134	0.533
couple	oui	0.300	0.574

la relation liant le score S et la probabilité p est :

$$S = Log\frac{p}{1-p} \Leftrightarrow p = \frac{\exp(S)}{1+\exp(S)} = \frac{1}{1+\exp(-S)}$$

Une mesure naïve de la qualité du score consiste à essayer la règle de décision suivante :

si la probabilité (mauvais payeur) est inférieure à 0.5 on refuse le client,
si elle est supérieure on l'accepte.

Dès lors on peut dresser un tableau des choix corrects ou faux :

	résultat du score	
réalité▼	classé mauvais	classé bon
(vrais) mauvais	1880	1800
(vrais) bons	1640	2360

Ce qui peut s'exprimer aussi sous la forme :

choix corrects		choix incorrects	
mauvais	bons	faux mauvais	faux bons
1880	2360	1640	1800

On en déduit en particulier deux indicateurs :

La **sensibilité** (*sensitivity*) du score est le pourcentage de bien classés par les valeurs élevées du score, c'est-à-dire dans la catégorie qui nous intéresse, et qui est ici celle des mauvais payeurs. C'est le pourcentage de **mauvais payeurs bien repérés comme tels** par la règle de décision : 1880 pour 3680, soit 51%.

La **spécificité** (*specificity*) du score est le pourcentage de **bons payeurs bien repérés comme tels** : 2360 pour 4000, soit 59%.

La sensibilité et la spécificité dépendent naturellement de la règle de décision fixée par le choix du découpage entre bons et mauvais potentiels (probabilité estimée inférieure ou supérieure à 0.5).

Un autre point de vue est aussi adopté, qui consiste à vérifier pour tous les couples de l'échantillon si leurs probabilités estimées sont placées dans le même ordre que leur réponse réelle (bons, mauvais). C'est une mesure de cohérence entre les observations et les scores, indépendante du seuil choisi pour le score ou la probabilité limite.

On dispose ainsi du pourcentage de **concordances**, de **discordances**, des ex-aequo (*tie* en anglais) et d'indices divers (Somers' D, Gamma, Tau-a, etc.).

12.1.2 Influence de la taille de l'échantillon

Changeons la taille de l'échantillon en gardant la distribution interne : on divise par 10 tous les effectifs.

La régression logistique fournit exactement les mêmes estimateurs des coefficients. Les écarts-types des coefficients changent et les tests aussi, mais le score, la spécificité, la sensibilité et les indices de cohérence ne changent pas.

Changeons la taille de l'échantillon en gardant la distribution interne de chaque sous-population (bons, mauvais), et, par exemple, en divisant par deux les effectifs des bons seulement :

Echantillon initial

Les bons payeurs		
famille	sans enfants	enfants
personne seule	1280	280
couple	1080	1360
Les mauvais payeurs		
famille	sans enfants	enfants
personne seule	1120	320
couple	680	1560

Echantillon réduit

Les bons payeurs		
famille	sans enfants	enfants
personne seule	640	140
couple	540	680
Les mauvais payeurs		
famille	sans enfants	enfants
personne seule	1120	320
couple	680	1560

Les coefficients du score sont identiques en ce qui concerne les effets principaux et croisés, seule la constante a changé. Le score a été décalé de $0.693 = Log(2)$.

Ce décalage s'explique facilement : quand on compare deux cases correspondantes (par exemple couple sans enfants) on passe du couple(1080,680) au couple (540,680), le rapport mauvais/bons a été multiplié par 2.

Plus généralement comparons deux échantillons avec des sous-populations de même structure, mais d'effectifs totaux différents :

Echantillon A	total bons	NbA
	total mauvais	NmA
Echantillon B	total bons	NbB
	total mauvais	NmB

La conservation des structures des deux populations s'exprime par la proportionalité des effectifs (i,j) des bons et des mauvais, mais pas dans le même rapport :

$$\frac{n_{ijb(A)}}{NbA} = \frac{n_{ijb(B)}}{NbB} \; et \; \frac{n_{ijm(A)}}{NmA} = \frac{n_{ijm(B)}}{NmB}$$

dont on déduit les cotes :

$$\begin{aligned} \frac{n_{ijm(A)}}{n_{ijb(A)}} &= \frac{NmA}{NmB}\frac{NbB}{NbA}\frac{n_{ijm(B)}}{n_{ijb(B)}} \\ \frac{p_{ij(A)}}{1-p_{ij(A)}} &= \frac{NmA}{NmB}\frac{NbB}{NbA}\frac{p_{ij(B)}}{1-p_{ij(B)}} \end{aligned}$$

et les scores :

$$S_{ijA} = Log\left[\frac{NmA}{NmB}\frac{NbB}{NbA}\right] + S_{ijB}$$

Les indices de qualité courants (sensibilité, spécificité, cohérence), ne sont pas modifiés.

On a bien retrouvé la propriété d'indépendance de la régression logistique par rapport à la distribution marginale des deux sous-populations, bons et mauvais. Ce qui laisse beaucoup de liberté pour le choix des échantillons, mais la qualité statistique du score dépend à la fois de la structure de l'échantillon et de sa taille.

Programme SAS associé.

```
data w1;input sitfam $ enf $ dette $ effectif;
datalines;
seul non faible 1280
seul non forte 1120
seul oui faible 280
seul oui forte 320
couple non faible 1080
couple non forte 680
couple oui faible 1360
couple oui forte 1560
;
run;
proc logistic data=w1 descending;
class sitfam enf;
freq effectif;
model dette=sitfam|enf/ctable pprob=0.5 ;
output out=sc xbeta=score;
run;
proc print data=sc;
run;
data w2;input sitfam $ enf $ dette $ effectif;
datalines;
seul non faible 128
seul non forte 112
seul oui faible 28
seul oui forte 32
couple non faible 108
couple non forte 68
couple oui faible 136
couple oui forte 156
;
run;
proc logistic data=w2 descending;
class sitfam enf;
freq effectif;
model dette=sitfam|enf/ctable pprob=0.5 ;
run;
data w3;input sitfam $ enf $ dette $ effectif;
datalines;
```

```
seul non faible 640
seul non forte 1120
seul oui faible 140
seul oui forte 320
couple non faible 540
couple non forte 680
couple oui faible 680
couple oui forte 1560
;
run;
proc logistic data=w3 descending;
class sitfam enf;
freq effectif;
model dette=sitfam|enf/ctable pprob=0.5
output out=sc xbeta=score;
run;
proc print data=sc;run;
```

12.2 Détermination du seuil

Pour que le score devienne opérationnel, il ne faut pas se limiter à la qualité de l'analyse statistique du modèle logistique. Il faut aussi évaluer les conséquences du choix du seuil. On essaie alors plusieurs seuils : l'analyse des erreurs correspondantes éclaire le choix du seuil définitif et donne une idée sur le rendement du score, plus complète que la seule classification consistant à comparer la probabilité estimée à 0.5.

Ceci doit être fait indépendamment de la taille de l'échantillon de façon à pouvoir l'appliquer à la population totale connue ou à la population des nouveaux clients.

Souvent on calcule la sensibilité et la spécificité qui ne dépendent que des seuils exprimés en terme de probabilité. Dans un cas aussi simple que le précédent (quatre classes d'individus), ce n'est pas très utile, mais quand on utilise de nombreuses modalités ou des variables quantitatives, il est intéressant de construire cette table et de représenter la sensibilité en fonction de la spécificité (qui sont liées, rappelons-le, au modèle et non aux effectifs de l'échantillon).

On va reprendre le calcul de la sensibilité et de la spécificité, en examinant les conséquences de différentes règles de décision basées sur un seuil caractérisé par la probabilité à partir de laquelle on rejette le client : ce qui veut dire qu'on considère que sa probabilité d'être mauvais est insupportable quand elle dépasse le seuil fixé.

La sensibilité (mauvais rejetés) va donc décroître de 100% à 0% et la spécificité (bons acceptés) va croître de 0% à 100% quand la probabilité, seuil d'acceptation, passera de 0 à 1.

Le cas "moyen" consiste à prendre 0.5 comme seuil. Le refus de prendre le moindre risque consiste à prendre un seuil très petit ; la crainte de rater des bons clients au risque d'en garder des mauvais conduira à prendre un seuil élevé.

On représente graphiquement soit la sensibilité et la spécificité en fonction du seuil, soit le complément à 100 de la sensibilité et la spécificité en fonction du seuil, soit la sensibilité en fonction de la spécificité (autant de points que de seuils). Dans ce dernier cas, par exemple, plus la courbe est loin de la diagonale, plus le score est efficace. Cette courbe est souvent appelée ROC (Receiver Operating Characteristic).

Programme SAS associé :

```
data w1;input sitfam $ enf $ dette $ effectif;
datalines;
seul non faible 1280
seul non forte 1120
seul oui faible 280
seul oui forte 320
couple non faible 1080
couple non forte 680
couple oui faible 1360
couple oui forte 1560
;
run;
proc logistic data=w1 descending;
class sitfam enf;
freq effectif;
model dette=sitfam|enf/ctable pprob=(0.1 to 0.9 by 0.1) outroc=roc;
output out=sc xbeta=score;
run;
proc print data=sc;
proc print data=roc;
run;
```

Il va de soi qu'au lieu de chercher un *seuil de probabilité* on peut chercher un *seuil sur le score*. Dans ce cas on compte souvent les effectifs correspondants aux quatre types de décision (acceptation ou rejet d'un bon ou un mauvais) par tranches de valeurs du score.

Exemple : un cas analogue un peu moins simple

Pour avoir un exemple où les scores varient plus librement avec des variables continues, nous quittons l'exemple précédent pour revenir à une problématique proche de l'exemple 6 (page 83) sur les effets de l'aspirine.

Il s'agit d'un examen de prévention destiné à définir des patients "à risques" en fonction de leurs antécédents. Le diagnostic porte sur la présence ou non d'adénomes. Il s'appuie sur les antécédents personnels d'adénomes, les antécédents familiaux de cancer, le sexe, l'âge, le BMI et le tabagisme cumulé (en paquets par jour, multiplié par le nombre d'années).

La courbe représentant la sensibilité en fonction de la spécificité, et à partir de laquelle on choisit le seuil convenable a la forme habituelle. Plus elle s'éloigne de la diagonale, meilleur est le score.

Programme SAS associé :

```
proc logistic data=cours.patients;
class atcdpaden atcdfcancol sexe;
model nbaden=atcdpaden sexe bmi agecolinit atcdfcancol*bmi/
ctable pprob=(0.1 to 0.9 by 0.1) outroc=roc;
run;
proc print data=roc;run;
data rocplus;set roc;
seuil=_prob_;atteint=1-_sensit_;pas_atteint=_1mspec_;run;
proc gplot data=rocplus;
symbol1 v=none i=join line=1 c='blue';
symbol2 v=none i=join line=33 c='black';
title ''courbe ROC'';
plot _sensit_ *_1mspec_=1 /overlay ;run;
title ''patients dont le diagnostic est alarmant'';
title2 ''atteints en trait plein, non atteints en pointille '';
plot pas_atteint*seuil=2 atteint*seuil=1/overlay;run;
quit;
```

12.3 Réflexions sur le choix du seuil

Le choix d'un seuil opérationnel pose plusieurs types de problèmes qui conduisent aux quatre remarques suivantes :

1. Si la qualité du score et le choix du seuil sont liés à l'échantillon avec lequel on estime les probabilités, il n'est pas sûr que cette qualité soit la même sur la population totale à moins que l'échantillon soit "parfaitement représentatif". Aussi sera t-il prudent de vérifier la qualité du score sur la population totale (ou une large partie de la population) en calculant les coûts réels des erreurs. Les utilisateurs traduisent cette différence en appelant l'échantillon sur lequel les estimations ont été faites, *ensemble d'apprentissage,* et la population totale ou une partie représentative de la population, *ensemble de test.*

   Cette distinction entre petite population pour l'estimation et grande population pour l'évaluation du score, est nécessaire quand des estimations menées sur une grande population sont impraticables à cause de sa taille. De plus, il peut être préférable de travailler sur un échantillon plus petit, mais dont les données sont parfaitement renseignées en éliminant les individus aberrants ou non intéressants pour l'objectif poursuivi, ce que les praticiens appellent "nettoyer" les données.

2. La qualité statistique du score est liée à la forme de l'échantillon. Un échantillon représentatif de la population globale peut être très déséquilibré au sens où les tailles des deux populations sont très différentes (c'est la taille la moins élevée qui influence principalement la variance des coefficients). Aussi cherche t-on à construire un échantillon où les deux sous-populations sont de même taille. Dans les essais cliniques

on prend un échantillon d'individus à risques le plus grand possible, et on ajoute un échantillon témoin de la même taille constitué d'individus "normaux" disponibles, ce qui exclue toute représentativité de la population globale.

3. Le score est une formule basée sur une estimation qui minimise un certain type d'erreur (maximum de vraisemblance, rapport de vraisemblance, déviance...). Or en fin de compte le décideur minimise un coût qui n'est pas lié de façon simple à ces erreurs. Il peut considérer, pour un essai clinique, qu'une erreur de diagnostic qui considère comme malade un patient réellement sain conduit à des investigations (coûteuses) inutiles. Mais ne pas repérer un vrai malade peut être très grave pour le patient (et pour les coûts sociaux).

   Pour un organisme de crédit, une banque, une assurance, toute erreur se traduit par une perte d'un bon client, ou la prise en charge d'un mauvais client avec des manques à gagner ou des coûts.

   Ces coûts peuvent se déduire facilement des résultats du score, pour chaque seuil, mais alors le décideur n'a pas tout à fait la même fonction objectif que le statisticien. Pour la bonne coopération des parties, il est souhaitable de vérifier que les objectifs sont assez cohérents afin que la formule obtenue par le statisticien convienne au décideur. Si les coûts du décideur se déduisent *linéairement* des nombres de classements corrects et incorrects, la méthode d'estimation n'est pas mise en cause, seul le seuil doit être bien choisi.

4. Une fois le score accepté, il va être appliqué à une population nouvelle dans laquelle les bons et mauvais ne sont pas connus. Pour évaluer le coût probable des erreurs, il faudrait avoir des idées sur la répartition des variables explicatives dans cette population. Si ce sont les mêmes que dans l'échantillon initial, le coût doit être le même que dans cet échantillon "représentatif de l'avenir". Si ce sont les mêmes que dans la population (test) connue, ce seront les mêmes que dans cette population *tant qu'elle n'a pas changé de structure.*

   Sinon on peut être amené à calculer des coûts directement à partir des probabilités a priori sur la répartition des nouveaux individus, le modèle logistique fournissant les probabilités conditionnelles (odds ratio).

Pour répondre à ces quatre difficultés, on peut poser quelques jalons qui invitent à consulter les nombreux ouvrages qui portent sur l'analyse discriminante et qui mettent l'accent sur les problèmes théoriques et pratiques de l'utilisation des scores.

La première remarque ne faisant intervenir que la taille des sous-populations, elle est résolue en modifiant le score, comme on l'a vu, par un décalage lié aux effectifs réels et aux effectifs de l'échantillon.

Il en est de même pour la quatrième remarque dans le cas particulier où on peut prévoir (probabilités à priori) les effectifs des populations de nouveaux clients.

La deuxième remarque est d'ordre pratique ; le choix des populations utilisées pour la construction du score est quelquefois imposé par les données disponibles, mais souvent ce choix est accompagné de nombreuses vérifications (nettoyage des données).

La troisième remarque est complétement dépendante de l'objectif recherché, l'utilité concrète du score.

Finalement il y aura deux conclusions majeures dans le choix du seuil et une mise en garde :

1. La spécificité et la sensibilité calculées à partir de l'échantillon sont uniquement liées au modèle logistique et on peut en déduire les performances théoriques du score pour n'importe quelle sous-population de bons et de mauvais, pourvu que ces sous-populations aient la même structure que celles de l'échantillon avec lequel le score a été construit. Si les coûts sont définis pour chaque type d'erreur (et éventuellement de réussite), le calcul du coût pour chaque seuil s'en déduit facilement.

2. Le choix du seuil se déduit donc de la spécificité et de la sensibilité, mais il peut s'exprimer de plusieurs façons : par la valeur du score attaché à chaque individu, par la valeur de son exponentielle (p/1-p) ou par la probabilité de l'évènement intéressant (par exemple, rejet d'un crédit, patient à risque), et c'est complètement équivalent.

Mise en garde fondamentale

Identifier les erreurs commises dans l'échantillon à partir duquel on a construit le score aux erreurs dans la population totale ou aux erreurs attendues dans une population future est risqué à deux titres :

1. Le score résulte d'une estimation, donc la mesure des erreurs pour un seuil donné est aléatoire, cette mesure peut avoir une loi de probabilité pas très simple, ce qui conduit parfois à vérifier que les spécificité et sensibilité, ne changent pas trop quand on modifie l'échantillon sur lequel elles sont estimées. C'est ce qu'on appelle la *validation* qui, dans le cas le plus simple, consiste à appliquer le score sur un échantillon du même type pour vérifier que les coûts ne différent pas trop de ceux de l'échantillon d'estimation.

2. Supposons que le score soit fiable (spécificité et sensibilité) sur l'échantillon, les erreurs proviennent alors de la nature aléatoire des comportements (réponses) et sont indépendantes ; mais les erreurs *calculées* avec le score estimé sont corrélées, ce qui se traduit par un biais sur l'estimation des erreurs. Il y a des moyens pour corriger ce type de biais comme le rééchantillonnage (bootstrap) ; souvent on effectue un calcul direct résultant de l'application du score, estimé sur le premier échantillon (apprentissage), à un échantillon indépendant de celui qui a servi à l'estimation, appelé souvent *échantillon test*.

   Et s'il s'agit d'une population nouvelle, il faudra suivre les résultats du score pour vérifier qu'il fonctionne bien comme prévu (pas de changement de comportement des clients).

La faiblesse de tous ces tests vient de ce qu'une variation dans les performances d'un score peut provenir du bruit inhérent aux comportements (variance) mais aussi des changements des comportements (biais). De sorte que le choix du seuil doit tenir compte de tous les éléments utiles à l'évaluation des risques que le décideur prend.

Et puis, un score n'est qu'un élément parmi d'autres dans un processus de décision.