Ricco Rakotomalala

Étude des dépendances - Variables qualitatives

Tableau de contingence et mesures d'association

Version 2.1

Université Lumière Lyon 2

Avant-propos

Ce support décrit quelques mesures statistiques destinées à quantifier et tester la liaison entre 2 variables qualitatives. Elles exploitent le tableau de contingence formé à partir des variables. Le domaine étant très vaste et les mesures innombrables, nous ne pourrons certainement pas prétendre à l'exhaustivité. Nous mettrons l'accent sur l'interprétation, les formules associées et la lecture pratique des résultats.

Nous nous concentrerons essentiellement sur la dépendance entre variables nominales. Le traitement des variables ordinales fera l'objet d'une partie distincte (Partie IV).

Un document ne vient jamais du néant. Pour élaborer ce support, je me suis appuyé sur différentes références, des ouvrages, mais aussi des ressources en ligne qui sont de plus en plus présents aujourd'hui dans la diffusion de la connaissance.

Les seuls bémols par rapport à ces documents en ligne sont le doute que l'on pourrait émettre sur l'exactitude des informations prodiguées, mais la plupart de leurs auteurs sont des enseignants-chercheurs qui font sérieusement leur travail ; une disponibilité plus ou moins aléatoire, au gré des migrations des serveurs et de la volonté de leurs auteurs, auquel il est très difficile de remédier ; les informations sont disparates, avec une absence d'organisation, à la différence des ouvrages qui suivent une ligne pédagogique très structurante.

Néanmoins, ces ressources en ligne renouvellent profondément le panorama des documents disponibles pour les enseignements. La gratuité n'est pas le moindre de leurs atouts.

Concernant ce support de cours, je me suis largement appuyé sur les documents exceptionnels, le mot est faible, de D. Garson ([16, 17], en 2011) qui ont malheureusement disparu aujourd'hui (avril 2020). Dommage. Tant le niveau du discours que la clarté du propos étaient remarquables. Ils m'ont personnellement permis de bien comprendre l'essence des mesures d'association entre variables qualitatives.

J'ai essayé de m'en démarquer en proposant un document en français (tant qu'à faire), en systématisant la présentation des formules, et en illustrant chaque situation à l'aide d'exemples.

De plus, si D. Garson faisait référence à un logiciel commercial pour la présentation des résultats (SPSS pour ne pas le citer), nous mettrons l'accent, pour notre part, sur le travail sur tableur [1] et l'utilisation des logiciels libres accessibles gratuitement sur le Web.

Bien entendu, selon la formule consacrée, ce document n'engage que son auteur. Toutes suggestions ou commentaires qui peuvent en améliorer le contenu sont le bienvenu.

1. Je suis intimement convaincu des vertus pédagogiques du tableur dans l'enseignement des techniques de traitement des données, tant pour les techniques statistiques que pour les méthodes de data mining (je vois déjà s'agiter les ayatollahs du tera-octet dans le fond de la salle, mais si, si, je persiste et signe). Une très bonne référence en France est la page Excel'ense de la revue MODULAD – http://www-rocq.inria.fr/axis/modulad/excel.htm

Table des matières

Partie I Mesures d'association entre variables qualitatives nominales

1 Liaison entre 2 variables qualitatives nominales . . . 3
- 1.1 Problème, données et notations . . . 3
- 1.2 Table de contingence . . . 4
- 1.3 Profils lignes et profils colonnes . . . 6

2 χ^2 d'écart à l'indépendance et mesures dérivées . . . 9
- 2.1 Statistique du χ^2 . . . 9
- 2.2 Test d'indépendance du χ^2 . . . 10
- 2.3 Décomposition du χ^2 - Contribution au χ^2 . . . 12
- 2.4 Mesures normalisées dérivées du χ^2 . . . 14
- 2.5 Autres tests et mesures symétriques . . . 16

3 Indicateurs asymétriques - Mesures PRE . . . 21
- 3.1 Les mesures PRE - Proportional Reduction in Error . . . 22
- 3.2 λ de Goodman et Kruskal . . . 23
- 3.3 τ de Goodman et Kruskal . . . 26
- 3.4 U de Theil . . . 30
- 3.5 Récapitulatif sur les mesures PRE . . . 34
- 3.6 Mesures PRE symétriques . . . 35

Partie II Cas particuliers

4 Tables 2×2 - Cas des variables binaires . . . 41
- 4.1 Coefficient ϕ et coefficient de corrélation . . . 43
- 4.2 χ^2 et correction de continuité pour les petits effectifs . . . 44
- 4.3 Coefficient $\mathcal{Q}$ de Yule basé sur les comparaisons par paires . . . 45
- 4.4 Test de Mc Nemar - Comparaison de proportions pour échantillons appariés . . . 50

5 Risque relatif, Odds et Odds-Ratio 53
5.1 Association Facteur / Maladie - Tableau de contingence 2×2 53
5.2 Probabilités conditionnelles 55
5.3 Risque relatif 56
5.4 Odds 60
5.5 Odds ratio 60
5.6 Récapitulatif pour le croisement "CLASS vs. PURPOSE regroupé" 64
5.7 Odds-ratio et coefficient $\mathcal{Q}$ de Yule 65
5.8 Odds-ratio dans les tableaux $2 \times C$ 66

6 Coefficient de concordance pour les variables nominales 69
6.1 Concordance de jugements - Le coefficient κ 69
6.2 Cas de 2 juges - κ de Cohen 71
6.3 Cas de m juges - κ de Fleiss 76
6.4 Nombre de juges quelconque - Formule généralisée 79

Partie III Association partielle

Partie IV Traitement des variables ordinales

7 Caractériser les associations ordinales - Inadéquation de la corrélation 89
7.1 Caractérisation des associations 89
7.2 Variables ordinales et corrélation 90
7.3 Les comparaisons par paires pour caractériser les associations ordinales 92
7.4 Exemple "Amount vs. Duration" sur le fichier GERMAN-CREDIT 94

8 Les mesures symétriques 97
8.1 γ de Goodman et Kruskal 97
8.2 τ_b de Kendall 100
8.3 τ_c de Kendall 102

9 *d* de Sommers - Une mesure asymétrique 107
9.1 Définition et estimation 107
9.2 Intervalle de confiance 108
9.3 Test de significativité 109
9.4 Une version symétrique du d de Sommers 110

10 Association ordinale pour les variables binaires (Mantel-Haenszel) 111

11 Association partielle pour variables ordinales 113

Partie V Annexes

A Gestion des versions 117

B Description des données GERMAN CREDIT 119

C Description du classeur EXCEL - GERMAN CREDIT 125

D Calculette en ligne (I) pour les tableaux 2×2 127

E Calculette en ligne (II) pour les tableaux 2×2 129

F Les mesures d'association dans le logiciel TANAGRA 131

Littérature 133

Partie I

Mesures d'association entre variables qualitatives nominales

1

Étude de la liaison - Table de contingence et profils

1.1 Problème, données et notations

Très souvent dans les études statistiques, nous sommes emmenés à étudier la liaison entre 2 ou plusieurs variables. La notion de *corrélation* est bien ancrée dans l'inconscient collectif. L'idée est de vérifier l'existence d'une relation entre deux phénomènes observés et, si elle existe, d'en mesurer l'intensité. Mieux encore, on essaie de savoir si l'un des phénomènes influence significativement l'autre. Par exemple, on veut savoir si la réduction de la vitesse permet de diminuer le nombre d'accidents sur la route ; est-ce que l'augmentation des heures d'études permettent d'améliorer la moyenne de l'étudiant, etc.

Lorsque les variables sont qualitatives nominales (discrètes nominales, catégorielles, etc.) c.-à-d. elles prennent un nombre dénombrable de modalités et il n'y a pas de relation d'ordre entre les modalités (ex. l'exemple le souvent cité est le sexe, il y a 2 valeurs possibles $\{homme, femme\}$, et n'allez surtout pas dire que l'une est plus élevée que l'autre, ou que sais-je encore...), la notion de corrélation, au sens large, qui s'appuie sur l'évolution concomitante ou opposée de valeurs n'est plus applicable. Il nous faut donc quantifier de manière différente la relation existant entre les phénomènes étudiés.

Il y a de nombreux exemples :

— En épidémiologie, on cherche à savoir si l'exposition à un facteur de risque (ex. tabagisme) entraîne l'apparition d'une maladie (ex. maladie cardio-vasculaire) ;
— En médecine, on évalue si, pour une maladie donnée, un diagnostic positif implique forcément la présence de la maladie ;
— En sociologie, on chercher à savoir s'il y a un lien entre la profession du père et la filière choisie par un étudiant à l'Université ;
— En marketing, on évalue la relation entre l'exposition à une publicité télévisuelle et l'acte d'achat ;
— En sociologie politique, on essaie de déterminer s'il y a un lien entre les régions et le candidat choisi lors des élections ;
— Etc.

Données GERMAN CREDIT

Dans ce support, nous utiliserons principalement le fichier GERMAN CREDIT qui recense les caractéristiques de 1000 demandeurs de crédits. Il comporte 23 variables avec, entre autres, l'objet de la demande de crédit (achat de voiture, équipement HI-FI, etc.), le statut de la personne (mariée, divorcée, etc.), son emploi (qualifié, non-qualifié, etc.)... La description complète des données est disponible en annexes (Annexe B). Le fichier de données au format XLSX accompagne ce support sur notre site Web (http://eric.univ-lyon2.fr/~ricco/cours/ouvrages.html). Ces données sont très connues, elles ont fait partie du projet STATLOG qui visait à évaluer les mérites comparées de plusieurs algorithmes d'apprentissage supervisé, l'objectif est de prédire au mieux la fiabilité (défaillant ou non c.-à-d. a totalement remboursé le crédit ou pas) du demandeur de crédit (URL des données http://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)).

Remarque 1 (Particularités du fichier GERMAN CREDIT). Pour avoir manipulé dans tous les sens ce fichier dans différents contextes, je me suis rendu compte qu'il présentait quelques singularités. Nous ne savons pas trop en quelle circonstances et à quel endroit ce fichier a été constitué, ni s'il s'agit de données réelles. En l'absence d'informations supplémentaires, nous resterons assez prudents quant à l'interprétation des résultats.

Dans un premier temps, nous nous intéresserons au croisement entre la variable "Housing" (Logement - Y) qui peut prendre 3 valeurs possibles {*for free* (pas de charge à payer : soit la personne habite dans sa famille, soit il a un logement de fonction, etc.) ; *own* (propriétaire) ; *rent* (locataire) } et la variable "Job", qui elle peut prendre 4 modalités différentes {*high qualif/self emp/mgm* (grosso modo, le management et les professions libérales...) ; *skilled* (travail qualifié) ; *unemp/unskilled non res* (sans emploi, emploi non qualifié et non résident) ; *unskilled resident* (les résidents avec un travail non-qualifié)}.

Données qualitatives... nominales et ordinales

Lorsque les variables sont **qualitatives ordinales** c.-à-d. les modalités sont dénombrables et ordonnées (ex. satisfaction : très satisfait, indifférent, mécontent), nous travaillons toujours sur un tableau de contingence mais les indicateurs utilisés sont différents. Nous y reviendrons **dans la partie IV de cet ouvrage**.

Les variables quantitatives, quant à elles, ne seront citées qu'à titre anecdotique, essentiellement pour marquer les différences ou les rapprochements possibles. Elles font l'objet d'un autre fascicule de cours accessible sur notre site web [8].

1.2 Table de contingence

Une table de contingence (ou un tableau de contingence [1]) est un tableau de comptage croisant les modalités de 2 variables. Il s'agit d'un tri croisé. Les cases du tableau correspondent aux effectifs observés

1. http://en.wikipedia.org/wiki/Contingency_table

n_{lc} associés conjointement à la modalité y_l de la variable Y en ligne ($Y \in \{y_1, \ldots, y_l, \ldots, y_L\}$), et la modalité x_c de la variable X en colonne ($X \in \{x_1, \ldots, x_c, \ldots, x_C\}$) (Tableau 1.1). Nous observons dans les totaux du tableau, en ligne ($n_{l.}$) et en colonne ($n_{.c}$), les effectifs marginaux. Ils correspondent au comptage selon un tri à plat de la variable en ligne et en colonne. Dans la dernière case du tableau, nous avons l'effectif global ($n_{..} = n$).

Y × X	x_1	$\cdots$	x_c	$\cdots$	x_C	Total
y_1	n_{11}	$\cdots$	n_{1c}	$\cdots$	n_{1C}	$n_{1.}$
$\vdots$		$\ldots$				$\vdots$
y_l	n_{l1}	$\cdots$	n_{lc}	$\cdots$	n_{lC}	$n_{l.}$
$\vdots$		$\ldots$				$\vdots$
y_L	n_{L1}	$\cdots$	n_{Lc}	$\cdots$	n_{LC}	$n_{L.}$
Total	$n_{.1}$	$\cdots$	$n_{.c}$	$\cdots$	$n_{.C}$	$n = n_{..}$

Tableau 1.1. Forme générique d'une table de contingence

Remarque 2 (Autre notation des effectifs conjoints). Pour certaines formules, par commodité, nous linéariserons le tableau en associant une modalité fictive k à chaque couple de modalité (l, c). Nous disposons ainsi d'un tableau, lu de gauche à droite, de haut en bas, avec $K = L \times C$ cases. On notera l'effectif conjoint observé $o_k = n_{lc}$; les effectifs marginaux n'existent pas dans cette configuration, l'effectif global reste n.

Dans notre exemple "HOUSING vs. JOB", le tableau permet d'accéder à plusieurs informations (Tableau 1.2) :

Housing × Job	high qualif/selp emp/mgm	skilled	unemp/unskilled non res	unskilled res	Total
for free	33	63	4	8	108
own	94	452	13	154	713
rent	21	115	5	38	179
Total	148	630	22	200	1000

Tableau 1.2. Croisement "Housing" et "Job"

— Il y a 1000 observations dans le fichier.
— 108 personnes occupent librement leur logement (Housing = for free).
— 148 personnes occupent une fonction managériale ou exercent une profession libérale.
— Dans les effectifs conjoints, nous observons notamment que 33 personnes occupant un poste hautement qualifié disposent gratuitement d'un logement.
— Etc.

1.3 Profils lignes et profils colonnes

Les effectifs sont relatifs à une taille d'échantillon. Ils ne sont pas très intéressants si nous voulons les rapporter au niveau de la population. Il semble plus indiqué de passer aux proportions, nous pouvons les extraire de différentes manières.

La **fréquence conjointe** est l'effectif conjoint rapporté à l'échantillon total, sa formule est $f_{lc} = \frac{n_{lc}}{n}$. Dans notre exemple, nous dirons que $f_{11} = \frac{33}{1000} = 3.3\%$ des individus ont un emploi hautement qualifié **ET** occupent gratuitement leur logement. En vérité, ce type de fréquence n'est pas très interprétable, elle est peu utilisée en pratique.

Plus intéressant est de ramener les effectifs par rapport aux totaux marginaux en ligne ou en colonne. En effet, chaque ligne (colonne) définit un sous ensemble de la population, nous pouvons calculer les proportions pour chaque groupe, les comparer entre elles et les comparer avec les proportions dans la population globale. On parle de **profils** c.-à-d. *fréquences conditionnelles* que l'on oppose aux *fréquences marginales* lues dans la dernière ligne (colonne) du tableau.

1.3.1 Profil colonne

Les proportions sont calculées dans chaque colonne du tableau $f_{l/c} = \frac{n_{lc}}{n_{.c}}$. Nous discernons mieux les informations importantes. Nous pouvons lire par exemple (Tableau 1.3) :

Housing × Job	high qualif/selp emp/mgm	skilled	unemp/unskilled non res	unskilled res	Total
for free	22%	10%	18%	4%	11%
own	64%	72%	59%	77%	71%
rent	14%	18%	23%	19%	18%
Total	100%	100%	100%	100%	100%

Tableau 1.3. Profil colonne des données "Housing vs. Job"

— 11% des personnes occupent gratuitement leur logement.
— Cette proportion passe à 22% chez les personnes occupant un poste hautement qualifié.
— Il est plus élevé également (18%) chez les "sans emploi et les non-résidents avec un travail non-qualifié".
— Ces mêmes personnes sont moins souvent propriétaire de leur logement, 59%, quand cette proportion est de 71% globalement (dans la marge).

Dernier point important, les proportions étant calculées relativement à chaque colonne, il est naturel que la fréquence marginale en colonne soit systématiquement égale à 100%.

1.3.2 Profil ligne

Dans ce cas, les proportions sont calculées pour chaque ligne. Nous observons par exemple (Tableau 1.4) que 15% des demandeurs de crédits occupent un emploi hautement qualifié. Chez les personnes occupant gratuitement leur logement, cette proportion passe à 31%.

Housing × Job	high qualif/selp emp/mgm	skilled	unemp/unskilled non res	unskilled res	Total
for free	31%	58%	4%	7%	100%
own	13%	63%	2%	22%	100%
rent	12%	64%	3%	21%	100%
Total	15%	63%	2%	20%	100%

Tableau 1.4. Profil ligne des données "Housing vs. Job"

La lecture qualitative du tableau de contingence sur la base des profils est très instructive. Mais en tant que statisticien, nous ne pouvons pas nous en contenter. Il faut caractériser la force du lien à l'aide d'indicateurs numériques, et éventuellement tester si elle significative, reflétant une relation existant réellement dans la population étudiée.

1.3.3 Fréquences et probabilités

Les fréquences (calculées sur l'échantillon) extraites du tableau de contingence sont en réalité des estimateurs des probabilités (définies dans la population). En reprenant les profils ci-dessus :

Probabilité conjointe. La probabilité $P(Y = y_l, X = x_c) = \pi_{lc}$ est estimée par la fréquence conjointe $\hat{\pi}_{lc} = f_{lc}$;

Probabilités conditionnelles. La probabilité conditionnelle $P(Y = y_l/X = x_c) = \pi_{l/c}$ (resp. $P(X = x_c/Y = y_l) = \pi_{c/l}$) est estimée par $\hat{\pi}_{l/c} = f_{l/c}$ (resp. $\hat{\pi}_{c/l} = f_{c/l}$) ;

Probabilités marginales. La probabilité marginale $P(Y = y_l) = \pi_{l.}$ (resp. $P(X = x_c) = \pi_{.c}$) est estimée par $\hat{\pi}_{l.} = f_{l.}$ (resp. $\hat{\pi}_{.c} = f_{.c}$).

2

χ^2 d'écart à l'indépendance et mesures dérivées

2.1 Statistique du χ^2

L'idée du χ^2 de Pearson est de comparer les effectifs observés o_k avec une situation de référence : les effectifs théoriques e_k que l'on obtiendrait si les variables Y et X étaient indépendantes.

Le procédé s'appuie sur un mécanisme de test d'hypothèses. L'hypothèse nulle est l'indépendance. Dans ce cas, *le contenu du tableau est entièrement défini par ses marges*, en effet, sous H_0 :

$$P(Y = y_l \cap X = x_c) = P(Y = y_l) \times P(X = x_c)$$

La statistique du χ^2 quantifie l'écart (la distance) entre les effectifs observés et les effectifs théoriques

$$\chi^2 = \sum_{k=1}^{K} \frac{(o_k - e_k)^2}{e_k} \tag{2.1}$$

où e_k correspondent aux effectifs sous H_0 : $e_k = \frac{n_{l.} \times n_{.c}}{n}$

Exemple numérique

Housing × Job	high qualif/selp emp/mgm	skilled	unemp/unskilled non res	unskilled res	Total
for free	15.98	68.0	2.4	21.6	108
own	105.5	449.2	15.7	142.6	713
rent	26.5	112.8	3.9	35.6	179
Total	148	630	22	200	1000

Tableau 2.1. Tableau "Housing" vs. "Job" sous hypothèse d'indépendance

Reprenons notre tableau ci-dessus (Tableau 1.2). Calculons le tableau des effectifs théoriques sous l'hypothèse d'indépendance (Tableau 2.1), nous pouvons en déduire le tableau de calcul de la statistique du χ^2 (Tableau 2.2), et obtenir, en faisant la somme :

Housing × Job	high qualif/selp emp/mgm	skilled	unemp/unskilled non res	unskilled res
for free	$18.11 = \frac{(33-15.98)^2}{15.98}$	0.37	1.11	8.56
own	1.26	0.02	0.46	0.91
rent	1.14	0.04	0.29	0.14

Tableau 2.2. "Housing" vs. "Job" - Tableau de calcul de la statistique du χ^2

$$\chi^2 = 18.11 + 0.37 + 1.11 + 8.56 + 1.26 + 0.02 + 0.46 + 0.91 + 1.14 + 0.04 + 0.29 + 0.14 = 32.41$$

Le χ^2 calculé varie de 0 à $n \times \min(L-1, C-1)$. Nous savons qu'en situation d'indépendance, il est égal à 0. En revanche, à cause des fluctuations d'échantillonnage, il peut prendre une valeur strictement positive sans que cela ne soit le reflet d'une liaison significative entre Y et X. Il nous faut donc définir une valeur seuil à partir de laquelle nous pouvons conclure, avec un risque quantifié, que le χ^2 calculé sur l'échantillon reflète réellement une liaison dans la population.

2.2 Test d'indépendance du χ^2

Pour mettre en oeuvre le test, nous devons définir plusieurs éléments [1] :

- La statistique du test est l'indicateur χ^2 proposé dans la section précédente.
- La distribution de la statistique sous l'hypothèse nulle, il s'agit d'une loi du χ^2.
- Le nombre de degrés de liberté, il s'agit *grosso modo* du nombre total de cases moins le nombre de cases que nous pouvons déduire des autres lorsque les marges sont fixées. Dans notre exemple, les marges étant fixées, il suffit de connaître le contenu des cases intérieures, les 2 premières lignes et les 3 premières colonnes, pour déduire le contenu de la 3-ème ligne et la 4-ème colonne. Le nombre de degrés de liberté est donc égal à $2 \times 3 = 6$. La formule générique est $ddl = (L-1) \times (C-1)$.
- Pour un risque α, la probabilité de rejeter l'hypothèse nulle alors qu'elle est vraie délimite la région critique du test c.-à-d. la région dans laquelle nous concluons au rejet de l'hypothèse nulle.

$$R.C. : \chi^2 > \chi^2_{1-\alpha}[(L-1)(C-1)]$$

où $\chi^2_{1-\alpha}[(L-1)(C-1)]$ est le quantile d'ordre $(1-\alpha)$ de la loi du χ^2 à $(L-1)(C-1)$ degrés de liberté.

Dans notre exemple, pour un risque $\alpha = 5\%$, la valeur seuil est $\chi^2_{0.95}(6) = 12.59$. Puisque $\chi^2 = 32.41 > 12.59$, nous concluons que le mode d'occupation du logement et le type d'emploi sont liés chez les demandeurs de crédit de l'organisme étudié.

Remarque 3 (p-value du test). Les logiciels statistiques fournissent rarement le seuil critique du test afin que nous lui comparons la statistique calculée. Ils produisent directement la *p-value* (probabilité critique

1. Voir une description pas à pas de la procédure en ligne : http://davidmlane.com/hyperstat/B155367.html

p_c) qui est égale à $P[\chi^2(ddl) > \chi^2_{calc}]$. Dans ce cas, nous comparons cette valeur directement avec le risque du test. Nous rejetons l'hypothèse nulle si $p_c < \alpha$. Dans notre exemple, $p_c = 0.000014$. La conclusion est cohérente avec la procédure habituelle.

Remarque 4 (Quelques commentaires sur les effectifs minimum dans les cases). Beaucoup de choses ont été dites concernant la piètre qualité de l'approximation à l'aide de la loi du χ^2 lorsque les effectifs sont faibles. Certains affirment que le test d'indépendance n'est pas valide s'il existe au moins une case où l'effectif théorique $e_k < 5$. D'autres assouplissent cette condition en indiquant que l'approximation est acceptable dès lors que 80% des cases ont $e_k \geq 5$. Enfin, lorsque nous manipulons un tableau 2×2 ($ddl = 1$), l'approximation est invalidée s'il existe au moins une case avec $e_k < 10$; on peut néanmoins s'en sortir en introduisant une modification de la statistique du χ^2, dite correction de Yates, que nous étudierons en détail plus loin. Il faut relativiser tout cela. L'information importante qu'il faut retenir est que les faibles valeurs de e_k ont tendance à "gonfler" la valeur de la statistique, indiquant à tort une liaison significative... Dans notre exemple, 2 cases sur 12 ($\approx 16\%$) ont un effectif théorique inférieur à 5 (Tableau 2.1), nous pouvons considérer que le test est valable.

Remarque 5 (Problème des grands effectifs). Lorsque les effectifs sont très élevés, le test du χ^2 aboutit quasi-systématiquement au rejet de l'hypothèse d'indépendance. La raison est simple. Un petit écart, aussi infime soit-il, se répercute fortement sur la statistique. Aucune normalisation tendant à relativiser le rôle de n n'est présent dans le processus. Le degré de liberté ne dépend que de la taille du tableau. Encore une fois, il faut relativiser les résultats du test. Les mesures normalisées permettent de corriger cet effet *grands effectifs* qui laisse souvent perplexe le praticien. Le t de Tschuprow et le V de Cramer en font partie (section 2.4). De manière plus large, la notion de taille d'effet (*effect size*) mis en avant par Cohen ([3]) peut être une réponse à cet écueil.

Remarque 6 (Autre formulation). Il existe une autre formulation moins connue de la statistique du χ^2 ([5], page 155). Elle a le mérite de la simplicité si nous devons la calculer à la main.

$$\chi^2 = \sum_{k=1}^{K} \frac{o_k^2}{e_k} - n \tag{2.2}$$

Pour notre exemple, nous aurons :

$$\begin{aligned}
\chi^2 &= (\frac{33^2}{15.98} + \frac{63^2}{68.04} + \cdots) - 1000 \\
&= (68.13 + 58.33 + \cdots) - 1000 \\
&= 32.41
\end{aligned}$$

2.3 Décomposition du χ^2 - Contribution au χ^2

2.3.1 Résidus

Détecter une liaison significative, c'est bien. Comprendre la nature de la liaison, c'est mieux. La différence entre le tableau observé et le tableau théorique permet de construire un indicateur, le résidu (Tableau 2.3)

$$res_k = o_k - e_k$$

Housing × Job	high qualif/selp emp/mgm	skilled	unemp/unskilled non res	unskilled res
for free	$17.02 = 33 - 15.98$	-5.0	1.6	-13.6
own	-11.5	2.8	-2.7	11.4
rent	-5.5	2.2	1.1	2.2

Tableau 2.3. "Housing" vs. "Job" - Tableau des résidus

Par construction, $\sum_k res_k = 0$. Le plus intéressant est sans aucun doute le signe du résidu $\text{sgn}(res_k)$ qui indique le sens de l'association : attraction entre les caractères (> 0) ou répulsion (< 0).

2.3.2 Résidus standardisés

Les valeurs des résidus ne sont pas comparables. Il faudrait les normaliser par les effectifs théoriques pour les mettre sur un pied d'égalité : c'est le résidu standardisé. Sa formule est la suivante :

$$rstd_{lc} = \frac{o_{lc} - e_{lc}}{\sqrt{e_{lc}}} \tag{2.3}$$

Le carré du résidu standardisé entre dans la composition de la contribution au χ^2 que nous verrons plus loin.

Le résidu standardisé est souvent utilisé pour détecter le caractère significatif de l'écart à l'indépendance d'une case. On s'appuie sur la distribution de $rstd_{lc}$ qui suit, **très approximativement**, une loi normale centrée et réduite. Nous décidons que la liaison entre 2 modalités de Y et X s'écarte significativement de l'indépendance lorsque $|rstd_{lc}| > 1.96$ pour un test à 5%.

Nous appliquons cette nouvelle formule sur notre tableau croisé (Tableau 2.4). Il semble que 2 cases s'éloignent assez sensiblement de la situation d'indépendance ("logement gratuit vs. emploi hautement qualifié" et "logement gratuit vs. emploi non qualifié et non résident").

2.3.3 Résidus ajustés

La règle de détection fondée sur le $rstd_{lc}$ est très approximative. Elle doit être utilisée à titre indicatif. Pour évaluer réellement le caractère significatif de l'écart, nous lui préférerons une autre normalisation.

Housing × Job	high qualif/selp emp/mgm	skilled	unemp/unskilled non res	unskilled res
for free	$4.26 = \frac{17.02}{\sqrt{15.98}}$ (*)	−0.61	1.05	−2.93 (*)
own	−1.12	0.13	−0.68	0.95
rent	−1.07	0.21	0.54	0.37

Tableau 2.4. "Housing" vs. "Job" - Tableau des résidus standardisés (* significatif à 5%)

Le résidu ajusté divise le résidu par l'estimation de son écart-type. Plus sa valeur est élevée, en valeur absolue, plus l'écart à l'indépendance est marquée. L'approximation à l'aide de la loi normale centrée et réduite est de meilleure qualité. **Nous opterons toujours pour cet indicateur pour détecter les cases s'écartant significativement de la situation d'indépendance**. Elle est plus puissante c.-à-d. elle est plus apte à détecter l'écart à l'indépendance lorsqu'elle existe. En effet, par construction, $radj_{lc} \geq rstd_{lc}$ puisque $(1 - f_{l.})(1 - f_{.c}) \leq 1$.

$$radj_{lc} = \frac{o_{lc} - e_{lc}}{\sqrt{e_{lc}(1 - f_{l.})(1 - f_{.c})}} \tag{2.4}$$

Nous appliquons cette nouvelle formule sur notre tableau croisé (Tableau 2.5). Nous effectuons un test bilatéral à 5%, le seuil critique est le quantile d'ordre 0.975 de la loi normale centrée et réduite : $u_{0.975} = 1.96$. Nous retrouvons les mêmes cases suspectes que précédemment, le test met en lumière deux autres nouvelles cases $radj_{21} = -2.27$ (*répulsion* entre propriétaire de son logement et emploi hautement qualifié) et $radj_{24} = 1.99$ (*attraction*, assez ténue néanmoins, entre propriétaire de son logement et résident avec travail peu qualifié).

Housing × Job	high qualif/selp emp/mgm	skilled	unemp/unskilled non res	unskilled res
for free	$4.88 = \frac{17.02}{\sqrt{15.98(1-148/1000)(1-108/1000)}}$ (*)	−1.06	1.13	−3.46 (*)
own	−2.27 (*)	0.41	−1.28	1.99 (*)
rent	−1.28	0.38	0.60	0.45

Tableau 2.5. "Housing" vs. "Job" - Tableau des résidus ajustés (* significatif à 5%)

2.3.4 Contribution au χ^2

Nous remarquons que la statistique du χ^2 est additive. Pour mesurer l'importance relative d'une case du tableau dans la caractérisation de la liaison, nous pouvons lui associer une valeur, dite *contribution au* χ^2, égale à

$$ctr_k = \frac{\frac{(o_k - e_k)^2}{e_k}}{\chi^2}$$

La contribution est le rapport entre le carré du résidu standardisé et le χ^2 global du tableau. Elle indique la fraction d'information qu'apporte la case dans la caractérisation de la liaison entre les variables. Plus forte sera la contribution, plus la case apporte de l'information.

Pour détecter les cases les plus importantes, il suffit de comparer la contribution avec la contribution moyenne des cases $\frac{1}{L \times C}$. Cela n'a pas valeur de test bien entendu. Mais au moins, avec ce procédé, on peut inspecter très rapidement un grand tableau de contingence.

Pour compléter l'analyse des écarts, il est d'usage d'associer à la contribution le signe du résidu afin que l'on identifie s'il s'agit d'une attraction ou une répulsion des modalités. Dans certains logiciels, il est possible de mettre en surbrillance les cases qui ont une contribution 2 fois (ou x-fois) plus élevée que la moyenne.

Dans notre exemple "Housing vs. Job", en analysant le tableau des contributions, nous constatons que l'information "utile" est quand même dominée par l'attraction "logement gratuit vs. emploi hautement qualifié" et de la répulsion "logement gratuit vs. emploi non qualifié et non résident" (Figure 2.1). Même si, par ailleurs, d'autres associations sont statistiquement significatives (cf. résidus ajustés).

Tableau des contributions

	high qualif/self emp/mgm	skilled	unemp/unskilled non res	unskilled resident
for free	55.9%	-1.2%	3.4%	-26.4%
own	-3.9%	0.1%	-1.4%	2.8%
rent	-3.5%	0.1%	0.9%	0.4%

Fig. 2.1. Tableau des contributions au χ^2 - Attractions et répulsions

Remarque 7 (Analyse des contributions). L'analyse des contributions n'a de sens que si le tableau est porteur d'une information significative (χ^2 significatif). Dans ce cas, elle complète à merveille l'interprétation des résultats en mettant en évidence la sur ou sous-représentation dans les cases. Certains auteurs préconisent son utilisation systématique ([12], page 151).

Remarque 8 (Contributions signées ou non-signées). Le tableau des contributions peut-être signé, comme nous avons choisi de le faire ici. Il peut être non-signé également. Dans ce cas, il est possible de calculer les sommes en ligne et colonne pour identifier les modalités les plus informatives des variables.

Remarque 9 (Analyse des grands tableaux). Lorsque le tableau comporte de nombreuses lignes et de colonnes, l'étude détaillée des contributions devient fastidieuse, on passe à une représentation graphique des écarts à l'indépendance pour mieux situer les attractions et répulsions : c'est le propos de *l'analyse factorielle des correspondances* ([12], pages 201 à 217).

2.4 Mesures normalisées dérivées du χ^2

Le test du χ^2 permet de déterminer si une liaison est *significative* ou pas. Lorsqu'elle l'est, l'étude des contributions permet d'identifier pourquoi. En revanche, rien jusqu'à présent ne nous permet de quantifier **l'intensité** de la liaison. La statistique χ^2 varie de 0 à $+\infty$. Elle ne nous est d'aucun secours. Il nous

faudrait une mesure normalisée dont on connaît la valeur maximale lorsque la liaison est parfaite c.-à-d. lorsque la connaissance de Y permet de déterminer avec certitude la valeur de X et/ou inversement.

Dans cette section, nous présentons quelques mesures normalisées dérivées du χ^2. Presque toutes ont un intervalle de définition connu à l'avance, nous permettant ainsi d'évaluer la force de la liaison.

Le coefficient ϕ

Le coefficient ϕ permet d'éliminer l'effet taille en normalisant le χ^2 par n. Il est défini par

$$\phi = \sqrt{\frac{\chi^2}{n}} \tag{2.5}$$

Il peut être calculé pour des tableaux de contingence de taille quelconque, mais il surtout avantageux pour les tables 2×2. En effet, dans ce cas, il s'interprète comme un coefficient de corrélation. Nous y reviendrons en détail dans la deuxième partie de ce support.

Dans notre exemple des données "Housing vs. Job", $\phi = \sqrt{\frac{32.41}{1000}} = 0.1800$; à comparer avec la valeur maximale

$$\phi_{max} = \sqrt{\min(L-1; C-1)}$$

.

Dans notre exemple, $\phi_{max} = \sqrt{\min(3-1; 4-1)} = 1.4142$. Si la liaison est significative, elle est loin d'être intense.

V de Cramer

Le V de Cramer peut être considéré comme une extension de ϕ aux tables de taille quelconque. Il varie entre 0 et 1, quelle que soit la taille du tableau. Il est identique à ϕ pour les tableaux 2×2. C'est **la mesure favorite des statisticiens** car il ne dépend ni de la taille de la table, ni de la taille de l'échantillon. Il propose de plus une interprétation intéressante ([14], pages 225 à 232). Sa formule est la suivante

$$V = \sqrt{\frac{\chi^2}{n \times \min(L-1; C-1)}} \tag{2.6}$$

Dans notre exemple, $V = \sqrt{\frac{32.41}{1000 \times \min(3-1;4-1)}} = 0.1273$.

T de Tschuprow

Le T de Tschuprow est une autre normalisation du χ^2 à l'aide de l'effectif total n et des degrés de liberté. Sa formulation est la suivante :

$$T = \sqrt{\frac{\chi^2}{n \times \sqrt{(L-1)(C-1)}}} \tag{2.7}$$

Dans le cas général, tableau $L \times C$, son interprétation est moins évidente que pour le V de Cramer. Pour un tableau 2×2, tout comme le V de Cramer, il est identique à ϕ.

L'intervalle de définition du T est 0 à

$$T_{max} = \sqrt[4]{\frac{\min(L-1; C-1)}{\max(L-1; C-1)}}$$

Dans certains cas particuliers, table 2×2 avec des marges identiques, le T de Tschuprow peut atteindre la valeur 1.

Reprenons notre exemple, nous obtenons

$$T = \sqrt{\frac{32.41}{1000 \times \sqrt{(3-1) \times (4-1)}}} = 0.1150$$

Que l'on comparera à 0 et $T_{max} = \sqrt[4]{\frac{\min(3-1;4-1)}{\max(3-1;4-1)}} = 0.9036$.

Coefficient de contingence - C de Pearson

Très utilisé naguère, cette mesure est supplantée maintenant par les autres coefficients, essentiellement parce que son interprétation est difficile. Elle aurait été conçue au départ comme une approximation de la corrélation entre deux variables dichotomisées artificiellement (Howell, page 181). Elle est définie de la manière suivante :

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \tag{2.8}$$

Comme $n > 0$, forcément, C ne peut jamais atteindre la valeur 1. Elle varie entre 0 et

$$C_{max} = \sqrt{\frac{\min(L; C) - 1}{\min(L; C)}}$$

Dans notre exemple, $C = \sqrt{\frac{32.41}{32.41+1000}} = 0.1772$; à comparer avec 0 et $C_{max} = \sqrt{\frac{\min(3;4)-1}{\min(3;4)}} = 0.8165$.

Pour obtenir un coefficient variant entre 0 et 1, certains auteurs préconisent l'utilisation du ratio $C^* = \frac{C}{C_{max}}$. Pourquoi pas... L'interprétation de l'indicateur n'est pas plus évidente pour autant.

Conclusion concernant la liaison "Housing vs. Job". Définitivement, on se rend compte que la liaison est certes significative entre le mode d'accès au logement et le type d'emploi, mais elle est, quelle que soit la mesure utilisée, de *faible intensité* (Tableau 2.6).

2.5 Autres tests et mesures symétriques

Les mesures basées sur le χ^2 sont largement répandues et disponibles dans les logiciels. Il existe d'autres indicateurs et tests basées sur des formulations différentes. Dans ce chapitre, nous nous limiterons aux mesures symétriques c.-à-d. nous obtiendrons exactement les mêmes valeurs si elles sont calculées sur la transposée du tableau de contingence.

Mesure	Val.Min	Val.Max	Valeur
χ^2	0	$+\infty$	32.4125
ϕ	0	1.4142	0.1800
V	0	1.0000	0.1273
T	0	0.9036	0.1150
C	0	0.8165	0.1772
C^*	0	1.0000	0.2170

Tableau 2.6. Récapitulatif des mesures - "Housing vs. Job"

2.5.1 Test du rapport de vraisemblance

Le test du rapport de vraisemblance[2] consiste à calculer le ratio des vraisemblances de deux configurations correspondants aux hypothèses à confronter.

Pour rappel, la vraisemblance pour n observations i.i.d. (indépendants et identiquement distribués) de la variable aléatoire X suivant une distribution de probabilité $P(\cdot)$ s'écrit :

$$L(X_1, X_2, \ldots, X_n) = \Pi_{i=1}^{n} P(X_i) \tag{2.9}$$

Dans la pratique, on manipule plutôt le log de la fonction de vraisemblance, entre autres parce que traiter des additions est toujours plus facile à manipuler que les multiplications.

Dans le cadre qui nous intéresse, il s'agit de calculer la vraisemblance à partir des paramètres estimés o_k, et de la confronter avec la vraisemblance lorsque l'hypothèse nulle (H_0 : Indépendance des variables en ligne et colonne du tableau de contingence) est vraie c.-à-d. e_k. Si les vraisemblances diffèrent significativement, on peut penser que la configuration actuelle (l'échantillon observé) est incompatible avec l'hypothèse nulle avancée.

La statistique du test s'écrit ([5], page 153) :

$$G = 2 \sum_{k=1}^{K} o_k \ln(\frac{o_k}{e_k}) \tag{2.10}$$

Sous l'hypothèse nulle, elle suit une loi du [χ^2 à $(L-1) \times (C-1)$] degrés de libertés.

Test d'indépendance du χ^2 et test du rapport de vraisemblance

Pour des échantillons de taille modérée, le test du rapport de vraisemblance et le test de Pearson donnent des résultats similaires. Il semble que l'approximation à l'aide de la loi du χ^2 de la distribution de la statistique G soit satisfaisante dès lors que $|o_k - e_k| > e_k$ pour toutes les cases du tableau[3].

On note aussi les ravages que peuvent causer de trop petites valeurs de o_k et e_k sur la statistique G. Lorsque que $o_k = 0$, nous n'intégrons pas la case dans le calcul de la somme.

2. Voir http://en.wikipedia.org/wiki/Likelihood_ratio_test
3. Voir http://en.wikipedia.org/wiki/G-test

Pour l'anecdote, le test du χ^2 de Pearson est en fait une approximation du test de rapport de vraisemblance. La statistique du test était plus facile à calculer du temps où les calculettes (et les ordinateurs) n'étaient pas encore répandues. Qui se rappelle encore des règles à calculer et des tables des logarithmes ? A l'époque, obtenir le logarithme d'un nombre n'était pas aisé. De nos jours, certes la question ne se pose plus, mais la popularité du test du χ^2 de Pearson est telle qu'il paraît difficile que le test du rapport de vraisemblance le supplante un jour. Même si, après vérification, je me suis rendu compte que tous les grands logiciels (commerciaux) de statistique le proposaient en standard.

Exemple "Housing vs. Job"

Housing × Job	high qualif/selp emp/mgm	skilled	unemp/unskilled non res	unskilled res
for free	$23.92 = 33 \times \ln \frac{33}{16.0}$	−4.85	2.08	−7.95
own	−10.87	2.82	−2.44	11.84
rent	−4.88	2.25	1.19	2.27

Tableau 2.7. "Housing" vs. "Job" - Tableau de calcul de la statistique G

Dans notre exemple, nous formons le tableau des $[o_k \ln(\frac{o_k}{e_k})]$ (Tableau 2.7; nous remarquons que certaines cases du tableau recèlent des valeurs négatives). Puis nous effectuons la somme que nous multiplions par 2 pour obtenir

$$G = 2 \times 15.3942 = 30.7903$$

La probabilité critique du test est $p_c = 0.000028$. Ces résultats sont très proches de ceux test du χ^2 de Pearson. Les conclusions sont cohérentes.

Remarque 10 (Test sur les petits effectifs). La statistique G peut être corrigée par le coefficient de Williams lorsque les effectifs deviennent faibles (voir [7], page 286). Cette version s'écrit :

$$G_{corr} = \frac{G}{1 + \frac{[(n \sum_l \frac{1}{n_{l.}}) - 1] \times [(n \sum_c \frac{1}{n_{.c}}) - 1]}{6n(L-1)(C-1)}} \tag{2.11}$$

La distribution est inchangée.

2.5.2 Mesure basée sur l'information mutuelle - Théorie de l'information

Principe

En théorie de l'information, l'information mutuelle quantifie le degré de dépendance entre 2 variables aléatoires. Formellement [4],

4. http://en.wikipedia.org/wiki/Mutual_information

$$I(Y,X) = \sum_l \sum_c \pi_{lc} \times \log_2 \frac{\pi_{lc}}{\pi_{l.} \times \pi_{.c}} \tag{2.12}$$

Elle indique dans quelle mesure les probabilités conjointes peuvent être déduites des probabilités conditionnelles[5]. Lorsque les variables sont indépendantes, $\pi_{lc} = \pi_{l.} \times \pi_{.c}$, par conséquent, $I(Y,X) = 0$.

La meilleure lecture de l'information mutuelle est certainement la lecture à partir de l'entropie de Shannon. On peut écrire :

$$\begin{aligned} I(Y,X) &= H(Y,X) - [H(Y) + H(X)] \\ &= H(Y) - H(Y/X) \\ &= H(X) - H(X/Y) \end{aligned}$$

où

— $H(Y,X) = -\sum_l \sum_c \pi_{lc} \log_2 \pi_{lc}$ est l'entropie conjointe, elle quantifie la quantité d'information nécessaire pour connaître les valeurs prises conjointement par les variables Y et X ;

— $H(Y) = -\sum_l \pi_{l.} \log_2 \pi_{l.}$ est l'entropie marginale, c'est la quantité d'information nécessaire pour connaître la valeur prise par Y (idem pour X : $H(X) = -\sum_c \pi_{.c} \log_2 \pi_{.c}$) ;

— $H(Y/X) = -\sum_c \pi_{.c} \sum_l \pi_{l/c} \log_2 \pi_{l/c}$ est l'entropie conditionnelle, elle indique la quantité d'information nécessaire pour connaître la valeur prise par Y *sachant la valeur prise par* X (idem $H(X/Y)$ pour X sachant Y).

Si Y et X sont indépendants, la connaissance de la valeur prise par X ne donne aucune indication sur la valeur prise par Y, et inversement.

Test d'indépendance

L'information mutuelle est toujours positive ($I(Y,X) \geq 0$). Le test d'indépendance consiste à évaluer si elle s'écarte *significativement* de la valeur 0, ce qui entraînerait le rejet de l'hypothèse d'indépendance.

La statistique du test est calculée à partir des estimations des probabilités obtenues à l'aide du tableau de contingence c.-à-d. les fréquences observées. Elle s'écrit

$$G' = 2 \times n \times \ln(2) \times I(Y,X) \tag{2.13}$$

Sous H_0, G' suit une loi du χ^2 à $[(L-1) \times (C-1)]$ degrés de liberté.

Exemple "Housing vs. Job"

Nous formons le tableau des $f_{lc} \log_2 \frac{f_{lc}}{f_{l.} f_{.c}}$ (Tableau 2.8) pour obtenir l'information mutuelle $I(Y,X) = 0.0222$. Nous en déduisons la statistique $G' = 30.7903$, la p-value du test est $p_c = 0.000028$. L'hypothèse d'indépendance entre "Housing" et "Job" est rejetée.

5. La notion d'information mutuelle peut être développée également pour les variables continues, nous manipulons dans ce cas les fonctions de distributions.

Housing × Job	high qualif/selp emp/mgm	skilled	unemp/unskilled non res	unskilled res
for free	$0.035 = 0.033 \times \ln \frac{0.033}{0.148 \times 0.108}$	-0.007	0.003	-0.011
own	-0.016	0.004	-0.004	0.017
rent	-0.007	0.003	0.002	0.003

Tableau 2.8. "Housing" vs. "Job" - Tableau de calcul de la statistique G'

Remarque 11 (Test du rapport de vraisemblance et information mutuelle). Nous constatons que les valeurs de G et G' sont strictement identiques. Ce n'est pas fortuit. Si leur essence n'est pas la même, la première est une procédure statistique, la seconde s'appuie sur la théorie de l'information, les formules concordent complètement et, par conséquent, **les deux tests sont équivalents**.

3

Indicateurs asymétriques - Mesures PRE

Les mesures étudiées jusqu'à présent étaient toutes symétriques c.-à-d. si nous transposons le tableau de contingence en mettant la variable X en ligne et Y en colonne, l'indicateur calculé restait identique. Cela est approprié si nous nous contentons de mesurer le degré de liaison entre 2 variables.

Lorsque nous voulons étudier la causalité, les mesures symétriques ne sont plus adaptées. Savoir si la connaissance de X améliore la connaissance des valeurs prises par Y est différent du schéma inverse. Autrement dit, l'étude du profil ligne relève d'une analyse différente de l'étude du profil colonne.

Prenons deux autres variables de notre fichier "GERMAN CREDIT". Nous nous intéressons au statut final du crédit "CLASS", le client présente un "good" ou un "bad" risque, et "CREDIT HISTORY, l'historique de crédit du client, les valeurs possibles sont "A30 : no credits taken, all credits paid back duly", "A31 : all credits at this bank paid back duly", "A32 : existing credits paid back duly till now", "A33 : delay in paying off in the past", "A34 : critical account/other credits existing (not at this bank)".

Pour un banquier, il est plus intéressant d'évaluer le risque d'un client à partir de son comportement passé. En revanche, constater l'insolvabilité d'un client pour s'intéresser par la suite à ses antécédents ne paraît pas vraiment pertinent. Du moins si l'objectif de l'étude est de produire des règles d'attribution ou non d'un crédit à un client. Si nous mettons en ligne la variable "CLASS", en colonne "CREDIT HISTORY", nous nous intéresserons donc avant tout au profil colonne (Figures 3.1 et 3.2).

NB class	credit_history					
class	all paid	critical/other existing	delayed previously	existing paid	no credits/all paid	Total
good	21	243	60	361	15	700
bad	28	50	28	169	25	300
Total	49	293	88	530	40	1000

Fig. 3.1. Tableau de contingence "Class" vs. "Credit history"

NB class	credit_history					
class	all paid	critical/other existing	delayed previously	existing paid	no credits/all paid	Total
good	42.9%	82.9%	68.2%	68.1%	37.5%	70.0%
bad	57.1%	17.1%	31.8%	31.9%	62.5%	30.0%
Total	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Fig. 3.2. Tableau de contingence "Class" vs. "Credit history" - Profil colonne

Les valeurs brutes présentent un intérêt anecdotique, les profils par contre donnent des indications très intéressantes sur le comportement des demandeurs de crédits :

— 70% des clients sont fiables (class = good) ;
— ce pourcentage passe à 82.94% chez les clients qui ont des crédit par ailleurs ("other credit") ;
— la situation est dégradée, 37.50% chez ceux qui n'ont aucun crédit ou qui ont remboursé normalement leurs crédits" [1].

Selon leur comportement dans le passé, le niveau de risque associé aux demandeurs de crédit est différent. L'analyse est clairement non symétrique. Il nous faut des mesures qui tiennent compte de cette spécificité.

3.1 Les mesures PRE - Proportional Reduction in Error

Les mesures PRE indiquent dans quelle proportion la connaissance de X permet de réduire l'erreur de prédiction des valeurs de la variable Y. La définition générique de la mesure est le rapport normalisé

$$PRE(Y/X) = \frac{e_Y - e_{Y/X}}{e_Y} \tag{3.1}$$

où e_Y est l'erreur de prédiction de Y si on n'a aucune connaissance de X ; $e_{Y/X}$ est l'erreur de prédiction si on utilise la connaissance de X, c'est l'erreur de prédiction conditionnelle.

Par définition, la mesure PRE varie entre 0 et 1 :

— Lorsque X n'amène aucune information sur Y, $e_{Y/X} = e_Y$, par conséquent $PRE(Y/X) = 0$.
— Lorsque la connaissance de X permet de prédire avec certitude la valeur de Y, $e_{Y/X} = 0$, et $PRE(Y/X) = 1$.

Exemple 1 (Titre et sexe). Prenons un exemple de la vie courante pour bien comprendre la nature asymétrique de la mesure. Connaître le titre d'une personne (Monsieur, Madame, Mademoiselle) permet de connaître avec certitude le sexe d'une personne. Dans ce cas, la mesure PRE doit être égale à 1. A l'inverse, connaître le sexe ne permet pas de deviner avec certitude le titre. Certes, à "Sexe = Masculin", on peut associer "Monsieur" ; mais à "Sexe = Féminin", on ne sait pas s'il faut attribuer "Madame" ou "Mademoiselle". La mesure doit traduire cette relative incertitude, elle doit être inférieure à 1.

Dans tous les cas, l'erreur conditionnelle $e_{Y/X}$ s'exprime comme une moyenne pondérée de l'erreur associée à chaque valeur de x_c

$$e_{Y/X} = \sum_{c=1}^{C} f_{.c} \times e_{Y/X=x_c} \tag{3.2}$$

1. Sans être un banquier émérite, on peut se poser des questions quand même quant à la pertinence de ces résultats. Les clients qui remboursent sont classés comme des gens à problème ? En l'absence d'informations supplémentaires, nous mettrons avant tout l'accent sur les formules et les procédures statistiques dans ce support. Ce fichier est pourtant très connu. Je trouve pour ma part étonnant de n'avoir trouvé, sur le Web, aucune réflexion critique concernant les résultats issus de ce fichier...

De fait, les mesures que nous étudierons dans ce chapitre se distinguerons essentiellement par la quantification de l'erreur de prédiction e. Dans chaque situation, nous indiquerons la formule de e, son interprétation, l'expression de la mesure PRE qui en découle, l'estimation par intervalle pour un niveau de confiance $(1-\alpha)$ et le test de significativité pour un risque α.

Remarque 12 (Mesures PRE symétriques). Les mesures PRE sont par définition asymétriques. Cependant, il est toujours possible d'en dériver des indicateurs symétriques en combinant $e_{Y/X}$ et $e_{X/Y}$. Nous y reviendrons plus loin (section 3.6).

3.2 λ de Goodman et Kruskal

3.2.1 Définition

Le λ de Goodman et Kruskal est défini de la manière suivante

$$\lambda_{Y/X} = \frac{\sum_{c=1}^{C} \max_l(n_{lc}) - \max_l(n_{l.})}{n - \max_l(n_{l.})} \tag{3.3}$$

λ indique la réduction de l'erreur de prédiction de la valeur de Y si nous connaissons la valeur de X.

Exemple

Reprenons l'exemple "CLASS vs. CREDIT HISTORY" (Figure 3.1). Nous calculons λ assez facilement :

$$\lambda_{Y/X} = \frac{28 + 243 + 60 + 361 + 25 - 700}{1000 - 700} = 0.0567$$

La formule en tant que telle ne pose pas de problème. La lecture des résultats est en revanche un peu ardue. Essayons de détailler le raisonnement. La meilleure prédiction de "CLASS" en dehors de toute connaissance de "CREDIT HISTORY" est "CLASS = GOOD" (c'est la modalité la plus fréquente), l'erreur de prédiction est dans ce cas $e_Y = 1 - 0.7 = 0.3$.

L'erreur de prédiction moyenne si l'on considère la valeur de X est

$$e_{Y/X} = 0.049 \times \frac{21}{49} + 0.293 \times \frac{50}{293} + 0.088 \times \frac{28}{88} + 0.530 \times \frac{169}{530} + 0.040 \times \frac{15}{40} = 0.283$$

On en déduit $\lambda_{Y/X} = \frac{0.3-0.283}{0.3} = 0.0567$: on réduit l'erreur de prédiction de 5.67% si on utilise la connaissance de "CREDIT HISTORY" pour prédire la valeur de "CLASS".

3.2.2 Variance asymptotique et intervalle de confiance

Il est possible d'obtenir un intervalle de confiance de λ en calculant la variance asymptotique et en s'appuyant sur la normalité asymptotique de la distribution.

Variance asymptotique

La variance asymptotique est égale à

$$\sigma_\lambda^2 = \frac{\sum_c \sum_l n_{lc} \left(\delta_{lc} - \delta_{l.} + \lambda \times \delta_{l.}\right)^2 - n \times \lambda^2}{[n - \max_l(n_{l.})]^2} \tag{3.4}$$

où

$$\delta_{lc} = \begin{cases} 1 & \text{si } l = \arg\max_i n_{ic} \\ 0 & \text{sinon} \end{cases}$$

Exemple sur les données "CLASS" vs. "CREDIT HISTORY"

On voudrait nous dégoûter de la statistique qu'on s'y prendrait pas autrement [2]. Arrêtons nous un instant sur cette formule pour bien comprendre son mécanisme. Nous prendrons comme référence le tableau de contingence croisant "CLASS" et "CREDIT HISTORY". Les calculs sont détaillés dans une feuille de calcul EXCEL que nous reproduisons ici (Figure 3.3).

A

class	all paid	critical/other existing	delayed previously	existing paid	no credits/all paid	Total
good	21	243	60	361	15	700
bad	28	50	28	169	25	300
Total	49	293	88	530	40	1000

B

class	all paid	critical/other existing	delayed previously	existing paid	no credits/all paid	Total
good	0	1	1	1	0	1
bad	1	0	0	0	1	0

C

class	all paid	critical/other existing	delayed previously	existing paid	no credits/all paid
good	0.8899	0.0032	0.0032	0.0032	0.8899
bad	1.0000	0.0000	0.0000	0.0000	1.0000

D

class	all paid	critical/other existing	delayed previously	existing paid	no credits/all paid
good	18.6874	0.7803	0.1927	1.1592	13.3482
bad	28.0000	0.0000	0.0000	0.0000	25.0000

Lambda	0.056667
Variance	0.000933
Ecart-type	0.030543

E

class	all paid	critical/other existing	delayed previously	existing paid	no credits/all paid
good	21.0000	0.0000	0.0000	0.0000	15.0000
bad	28.0000	0.0000	0.0000	0.0000	25.0000

Variance (H0)	0.000986
Ecart-type (H0)	0.031396
z	1.804929
p-value	0.071086

Fig. 3.3. Tableau de contingence "Class" vs. "Credit history" - Calcul du λ et tests associés

Le point de départ est toujours le tableau de contingence, c'est la partie **A** de la feuille de calcul.

Dans la distribution marginale de "CLASS" ($n_{i.}$, le maximum est la valeur 700 et elle est située sur la ligne numéro $l = \arg\max_i n_{i.} = 1$. On en déduit alors $\delta_{1.} = 1$ et $\delta_{2.} = 0$. Nous procédons à l'identique pour

2. Tout ceci est à but pédagogique bien entendu. Dans la pratique, les logiciels se chargent de manipuler, correctement si possible, toutes ces formules. Ouf !

chaque distribution conditionnelle (pour chaque colonne c) pour obtenir les valeurs δ_{lc}. Nous obtenons la partie **B**.

Nous formons en **C** la portion de l'équation correspondant à $(\delta_{lc} - \delta_{l.} + \lambda \times \delta_{l.})^2$; puis en **D**, la fraction associée à $n_{lc}\,(\delta_{lc} - \delta_{l.} + \lambda \times \delta_{l.})^2$. En effectuant la somme et en complétant l'équation, nous obtenons

$$\sigma^2_\lambda = \frac{87.1678 - 1000 \times 0.056667}{(1000 - 700)^2} = 0.0009333$$

L'écart-type est $\sigma_\lambda = \sqrt{0.0009333} = 0.030543$.

Remarque 13 (Gestion des ex-aequos). Le calcul ci-dessus devient épineux si nous avons des ex-aequos parmi les valeurs maximales situés dans chaque colonne du tableau de contingence. Les résultats peuvent différer d'un logiciel à l'autre, tout dépend de la règle de gestion adoptée. Dans SAS par exemple (http://v8doc.sas.com/sashtml/stat/chap28/sect20.htm), on prend comme valeur de référence le plus petit indice associé au maximum.

Intervalle de confiance

En utilisant la normalité asymptotique de λ, l'intervalle de confiance au niveau $(1 - \alpha)$ est défini par

$$\left[\lambda - u_{1-\frac{\alpha}{2}} \times \sigma_\lambda ; \lambda + u_{1-\frac{\alpha}{2}} \times \sigma_\lambda\right]$$

où $u_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi normale centrée et réduite.

Exemple sur les données "CLASS" vs. "CREDIT HISTORY"

Pour "CLASS vs. CREDIT HISTORY", au niveau de confiance de 90%, le quantile est $u_{0.95} = 1.644853$, la borne basse de l'intervalle de variation de λ est $0.0064 = 0.056667 - 1.644853 \times 0.030543$, la borne haute 0.1069.

3.2.3 Test de significativité

Au delà de l'intervalle de variation, on pourrait également utiliser les informations ci-dessus (variance et distribution asymptotique) pour mettre en place le test d'hypothèses

$$H_0 : \lambda = \lambda_0$$
$$H_1 : \lambda \neq \lambda_0$$

La situation est plus complexe lorsqu'il s'agit de vérifier si λ est significatif ($H_0 : \lambda = 0$) c.-à-d. la variable X n'amène aucune information sur la connaissance de Y :

1. Puisque $0 \leq \lambda \leq 1$ par définition, l'hypothèse alternative s'écrit obligatoirement $H_1 : \lambda > 0$, nous mettons en place un test unilatéral ;

2. La variance asymptotique calculée de manière générale (Équation 3.4) n'est plus valable, il faut utiliser une autre estimation en rapport avec l'hypothèse nulle. Elle s'écrit,

$$\sigma^2_\lambda(0) = \frac{\sum_c \sum_l n_{lc}(\delta_{lc} - \delta_{l.})^2 - \frac{1}{n}[\sum_c \max_l(n_{lc}) - \max_l(n_{l.})]^2}{[n - \max_l(n_{l.})]^2} \tag{3.5}$$

On forme le rapport suivant pour réaliser le test,

$$z_\lambda = \frac{\lambda}{\sigma_\lambda(0)}$$

En nous appuyant toujours sur la normalité asymptotique, la région critique du test, au risque α, est définie par

$$R.C. : z_\lambda > u_{1-\alpha}$$

Il est également possible d'utiliser la probabilité critique (p-value) $p_c = 1 - \Phi(z_\lambda)$, où $\Phi(\cdot)$ est la fonction de répartition de la loi normale centrée réduite. On rejette l'hypothèse nulle lorsque $(p_c < \alpha)$.

Exemple sur les données "CLASS" vs. "CREDIT HISTORY"

Nous reprenons notre feuille de calcul sous EXCEL (Figure 3.3). La partie **E** correspond à $\cdots n_{lc}(\delta_{lc} - \delta_{l.})^2 \cdots$. Nous complétons les calculs pour obtenir la variance asymptotique

$$\sigma^2_\lambda(0) = \frac{89.0000 - \frac{1}{1000}[707 - 700]^2}{[1000 - 700)]^2} = 0.000986$$

On en déduit l'écart type $\sigma_\lambda(0) = 0.031396$ et le rapport $z_\lambda = \frac{0.056667}{0.031396} = 1.804929$. Au risque $\alpha = 5\%$, nous pouvons comparer cette valeur avec $u_{0.95} = 1.644853$. Nous pouvons aussi calculer directement la probabilité critique $p_c = 0.035543$.

Dans notre exemple, au risque $\alpha = 5\%$, nous considérons que λ est significativement différent de 0. La connaissance de l'historique d'un client, en matière de crédit, amène de l'information sur le succès de la demande actuelle d'emprunt.

Remarque 14 (Formule générale de la variance et variance sous H_0). Nous noterons que les variances asymptotiques standard et sous l'hypothèse nulle $(H_0 : \lambda = 0)$ sont finalement assez proches. Les logiciels font peu la distinction. SPSS est un des rares logiciels à indiquer clairement les estimations utilisées dans ses sorties.

3.3 τ de Goodman et Kruskal

3.3.1 Définition

Le λ de Goodman et Kruskal impose d'affecter, pour chaque valeur de X, la valeur de Y la plus fréquente. L'erreur de prédiction est la probabilité d'affecter à tort cette modalité. Cette approche est un

peu fruste. Lorsque les modalités les plus fréquentes sont toutes situées sur la même ligne, nous obtenons mécaniquement $\lambda = 0$, alors même que les profils colonnes sont très différents.

Prenons l'exemple du croisement des variables "CLASS" et "PURPOSE" (l'objet de la demande de crédit) de notre fichier. La classe majoritaire est toujours "CLASS = GOOD" dans toutes les colonnes, elle est aussi la modalité globalement majoritaire (dans la marge). Naturellement, $\lambda = 0$, annonçant que la connaissance de PURPOSE ne donne indication sur "CLASS". Pourtant les profils colonnes diffèrent de manière importante d'une modalité de PURPOSE à l'autre. Nous ne pouvons pas décemment dire qu'il y a indépendance ici (Figure 3.4).

	radio/tv	education	furniture/ equipmen	new car	used car	business	domestic appliance	repairs	other	retraining	Sum
good	218	28	123	145	86	63	8	14	7	8	700
	77.86%	*56.00%*	*67.96%*	*61.97%*	*83.50%*	*64.95%*	*66.67%*	*63.64%*	*58.33%*	*88.89%*	*70%*
bad	62	22	58	89	17	34	4	8	5	1	300
	22.14%	*44.00%*	*32.04%*	*38.03%*	*16.50%*	*35.05%*	*33.33%*	*36.36%*	*41.67%*	*11.11%*	*30%*
Sum	280	50	181	234	103	97	12	22	12	9	1000
	100%	*100%*	*100%*	*100%*	*100%*	*100%*	*100%*	*100%*	*100%*	*100%*	*100%*

Fig. 3.4. Tableau de contingence "Class" vs. "Purpose" - Profil colonne

On peut enrichir l'indicateur en étudiant *la structure de l'erreur*. C'est le propos du $\tau_{Y/X}$ de Goodman et Kruskall. Si on affecte à un individu la modalité y_l, la probabilité de le faire à tort est $(1-\pi_l)$. Si nous adoptons une stratégie d'affectation proportionnelle à la présence à la distribution des modalités de Y, la probabilité de conclure à tort à la modalité c_l dans la population est $\pi_l \times (1-\pi_l)$. La probabilité d'erreur est la moyenne pondérée sur l'ensemble des modalités de Y.

Si on estime la probabilité à l'aide des fréquences, l'erreur marginale s'écrit

$$S_Y = \sum_{l=1}^{L} f_{l.}(1 - f_{l.}) \tag{3.6}$$

Si toutes les valeurs sont centrées sur une des modalités de Y, $S_Y = 0$; s'il y a équi-répartition c.-à-d. $f_{l.} = \frac{1}{L}$, $\forall l$, alors $S_Y = 1 - \frac{1}{L}$.

L'erreur conditionnelle est la moyenne pondérée des erreurs pour chaque modalité x_c de X

$$S_{Y/X} = \sum_{c=1}^{C} f_{.c} \sum_{l=1}^{L} f_{l/c}(1 - f_{l/c}) \tag{3.7}$$

Comme toutes les mesures PRE, le τ de Goodman et Kruskal est défini par

$$\tau_{Y/X} = \frac{S_Y - S_{Y/X}}{S_Y} \tag{3.8}$$

Le τ de Goodman et Kruskal varie entre 0, la connaissance de X ne donne aucune indication sur Y, et 1, la connaissance de X permet de connaître avec certitude la valeur de Y. La mesure est asymétrique.

A

class	all paid	critical/other existing	delayed previously	existing paid	no credits/all paid	Total
good	21	243	60	361	15	700
bad	28	50	28	169	25	300
Total	49	293	88	530	40	1000

B

class	all paid	critical/other existing	delayed previously	existing paid	no credits/all paid	Total
good	0.429	0.829	0.682	0.681	0.375	0.700
bad	0.571	0.171	0.318	0.319	0.625	0.300

C

class	all paid	critical/other existing	delayed previously	existing paid	no credits/all paid	Total
good	0.245	0.142	0.217	0.217	0.234	0.210
bad	0.245	0.142	0.217	0.217	0.234	0.210

D

Somme	0.490	0.283	0.434	0.434	0.469	0.420
Poids	0.049	0.293	0.088	0.530	0.040	1.000
Produit	0.024	0.083	0.038	0.230	0.019	0.420

S(Y)	0.420000
S(Y/X)	0.394090
Tau	0.061691
sigma(Tau)	0.015238
u(0.95)	1.644853

	Borne basse	Borne haute
I.C. à 90%	0.0366	0.0868

C(tau)	61.629706
DDL	4
p-value	1.31798E-12

Fig. 3.5. Tableau de contingence "Class" vs. "Credit history" - Calcul du τ et tests associés

Exemple sur les données "CLASS" vs. "CREDIT HISTORY"

Le détail des calculs est retracé dans une feuille Excel (Figure 3.5) :

1. Le point de départ est toujours le tableau de contingence **A**.
2. Nous formons les profils colonnes **B**.
3. Pour chaque colonne, nous calculons les valeurs $f_{l/c}(1 - f_{l/c})$. Dans notre exemple, Y ne possède que 2 modalités, $f_{1/c} = 1 - f_{2/c}$, voilà pourquoi nous avons des valeurs identiques sur les 2 lignes du tableau **C**.
4. Reste à effectuer les sommes pour chaque colonne $\sum_l f_{l/c}(1 - f_{l/c})$, à extraire les poids relatifs de chaque colonne $f_{.c}$ et calculer le produit $f_{.c} \times \sum_l f_{l/c}(1 - f_{l/c})$ (Tableau **D**).

Le τ de Goodman et Kruskal découle de

$$\tau_{Y/X} = \frac{0.42 - 0.394090}{0.42} = 0.061691$$

τ étant défini entre 0 et 1, on se rend compte que le lien est de faible intensité... s'il existe, cela reste à confirmer avec le test de significativité,.

Remarque 15 (Indice de Gini et ses multiples interprétations). L'indice S_Y se prête à de nombreuses interprétations. C'est ce qui fait son charme. Dans de nombreuses références, on le rencontre sous l'appellation *indice de Gini*. Il se lit dans ce cas comme un indice de concentration (ou de diversité, tout dépend du point de vue). En statistique, on peut le concevoir comme une variance calculée sur variables nominales. En théorie de l'information, il s'interprète comme un indice d'incertitude, on parle d'entropie quadratique.

On fera le lien avec le coefficient d'incertitude de Theil que nous verrons dans la section suivante. En tous les cas, voilà une autre point de vue sur les mesures PRE, on peut les interpréter comme une analyse de la diversité de Y lorsque l'on connaît les valeurs de X c.-à-d. tout simplement une analyse de variance.

3.3.2 Variance asymptotique et intervalle de confiance

Nous nous servons de la normalité asymptotique pour calculer l'intervalle de confiance. Nous devons au préalable produire une estimation de la variance.

Variance asymptotique

La variance asymptotique correspond encore à une autre formule à coucher dehors. Il faut la lire une fois et s'empresser de la faire calculer par les logiciels. Elle s'écrit :

$$\sigma_\tau^2 = \frac{4}{\delta^4}\sum_c\sum_l n_{lc}\left[(\nu-\delta)\left(\frac{1}{n_{.c}}\sum_l n_{lc}n_{l.} - n_{l.}\right) - n\delta\left(\frac{1}{n_{.c}^2}\sum_l n_{lc}^2 - \frac{1}{n_{.c}}n_{lc}\right)\right]^2 \tag{3.9}$$

avec

$\delta = n^2 - \sum_l n_{l.}^2$

$\nu = n\sum_c\sum_l \frac{n_{lc}^2}{n_{.c}} - \sum_l n_{l.}^2$

Nous passerons sur les détails des calculs, nous ferons confiance à SPSS qui nous fourni, pour le croisement "CLASS vs. CREDIT HISTORY", la valeur $\sigma_\tau = 0.015238$.

Intervalle de confiance

Pour un niveau de confiance $1-\alpha$, nous pouvons calculer l'intervalle de confiance

$$\left[\tau - u_{1-\frac{\alpha}{2}} \times \sigma_\tau ; \tau + u_{1-\frac{\alpha}{2}} \times \sigma_\tau\right]$$

Pour le cas "CLASS vs. CREDIT HISTORY" de notre fichier, l'intervalle $[0.0366; 0.0868]$ a 90% de chances de contenir la "vraie" valeur de τ (Figure 3.5).

3.3.3 Test de significativité

Pour tester la significativité du coefficient τ, nous utilisons un schéma différent. La variance n'est pas mise à contribution. Il a été établi que sous l'hypothèse H_0 d'indépendance ($\tau = 0$), la statistique

$$C(\tau) = (n-1)(L-1)\tau_{Y/X} \tag{3.10}$$

Suit asymptotiquement une loi du χ^2 à $(L-1)(C-1)$ degrés de liberté.

C'est un test unilatéral. La région critique du test s'écrit :

$$R.C. : C(\tau) > \chi^2_{1-\alpha}[(L-1)(C-1)]$$

où $\chi^2_{1-\alpha}[(L-1)(C-1)]$ est le fractile d'ordre $1-\alpha$ de la loi du χ^2 à $(L-1)(C-1)$ degrés de liberté.

Nous pouvons également nous baser sur la probabilité critique $p_c = P\left(\chi^2 > C(\tau)\right)$ pour décider du rejet ou de l'acceptation de l'hypothèse nulle.

Exemple sur les données "CLASS" vs. "CREDIT HISTORY"

Nous formons la statistique du test

$$C(\tau) = (1000-1) \times (2-1) \times 0.061691 = 61.629706$$

Le degré de liberté est égal à $ddl = (2-1)(5-1) = 4$, la p-value du test est $p_c = 1.31798 \cdot 10^{-12}$. Le τ de Goodman et Kruskal est très fortement significatif (Figure 3.5).

3.3.4 Cas de "CLASS vs. PURPOSE"

L'association "CLASS vs. PURPOSE" est très particulière. En effet, rappelons-nous, si nous utilisons le λ de Goodman et Kruskal, nous obtiendrons la valeur 0, laissant à croire qu'il n'y aucun lien entre les deux variables. En passant au τ de Goodman et Kruskal, la conclusion est tout autre (Figure 3.6), montrant, si besoin était, l'intérêt de cette seconde mesure d'association lorsque la modalité majoritaire est la même dans toutes les colonnes du tableau de contingence.

Goodman & Kruskal's Tau for nominal attributes

Y	X	Tau	Chi²	d.f.	p-value
class	purpose	0.033356	33.32	9	0.0001

Fig. 3.6. Tableau de contingence "Class" vs. "Purpose" - τ de Goodman et Kruskal

3.4 U de Theil

3.4.1 Définition

Le U de Theil (Uncertainty Coefficient - Coefficient d'incertitude) indique la réduction proportionnelle d'incertitude lorsque l'on essaie de prédire les valeurs de Y à l'aide de X. Il est basé sur la théorie de l'information. De la même manière que le τ de Goodman et Kruskal, il confronte les distributions conditionnelles avec la distribution marginale.

En vérité, τ de Goodman et Kruskal et U de Theil reposent sur les mêmes mécanismes. Ces indicateurs utilisent une entropie pour mesurer l'incertitude, la première utilise l'entropie quadratique (connu également sous le nom d'indice de Gini), la second s'appuie sur l'entropie de Shannon. Par conséquent, ils fournissent des résultats très similaires. On préférera peut-être le U de Theil dans la pratique car

il présente l'insigne avantage d'être plus connu et plus présent dans les logiciels, la variance qui lui est associée est également un peu plus facile à calculer (ça nous changera tiens !).

L'entropie marginale [3] s'écrit

$$H(Y) = -\sum_l f_{l\cdot} \ln f_{l\cdot} \tag{3.11}$$

L'entropie conditionnelle

$$H(Y/X) = -\sum_c f_{\cdot c} \sum_l f_{l/c} \ln f_{l/c} \tag{3.12}$$

Le coefficient d'incertitude de Theil est une mesure PRE,

$$U_{Y/X} = \frac{H(Y) - H(Y/X)}{H(Y)} \tag{3.13}$$

Remarque 16 (Fréquence nulle). $f \ln f = 0$ car $\lim_{a\to 0} a \ln a = 0$.

Le coefficient d'incertitude varie entre 0 et 1. Lorsque $U = 1$, la connaissance de X permet de déterminer avec certitude la valeur prise par Y.

Exemple sur les données "CLASS" vs. "CREDIT HISTORY"

Nous retraçons le calcul du coefficient d'incertitude dans une feuille de calcul (Figure 3.7) :

1. Le point de départ est toujours le tableau de contingence **A**.
2. Nous formons les profils colonnes **B**.
3. Pour chaque colonne, nous calculons les valeurs $f_{l/c} \ln(f_{l/c})$ (Tableau **C**).
4. Reste à effectuer les sommes pour chaque colonne $\sum_l f_{l/c} \ln(f_{l/c})$, à extraire les poids relatifs de chaque colonne $f_{\cdot c}$ et à calculer le produit $f_{\cdot c} \times \sum_l f_{l/c} \ln(f_{l/c})$ (Tableau **D**).

Le U de Theil est obtenu avec

$$U_{Y/X} = \frac{0.610864 - 0.580631}{0.610864} = 0.049493$$

3.4.2 Variance asymptotique et intervalle de confiance

Nous nous servons de la normalité asymptotique pour calculer l'intervalle de confiance. Nous devons au préalable produire une estimation de la variance.

3. Nous utilisons le logarithme népérien pour être en accord avec les notations utilisées dans les logiciels de statistique courants. Le résultat serait le même si nous utilisions le logarithme à base 2, en effet le coefficient d'incertitude est un ratio.

A

class	all paid	critical/other existing	delayed previously	existing paid	no credits/all paid	Total
good	21	243	60	361	15	700
bad	28	50	28	169	25	300
Total	49	293	88	530	40	1000

B

class	all paid	critical/other existing	delayed previously	existing paid	no credits/all paid	Total
good	0.429	0.829	0.682	0.681	0.375	0.700
bad	0.571	0.171	0.318	0.319	0.625	0.300

C

class	all paid	critical/other existing	delayed previously	existing paid	no credits/all paid	Total
good	-0.363	-0.155	-0.261	-0.262	-0.368	-0.250
bad	-0.320	-0.302	-0.364	-0.364	-0.294	-0.361

D

Somme	-0.683	-0.457	-0.625	-0.626	-0.662	-0.611
Poids	0.049	0.293	0.088	0.530	0.040	1.000
Produit	-0.033	-0.134	-0.055	-0.332	-0.026	-0.611

H(Y)	0.610864
H(Y/X)	0.580631
U	0.049493

E

Calcul H(X)	-0.147780814	-0.359681722	-0.213876825	-0.336485484	-0.128755033

H(X)	1.186580

F

class	all paid	critical/other existing	delayed previously	existing paid	no credits/all paid
good	0.021	0.243	0.060	0.361	0.015
bad	0.028	0.050	0.028	0.169	0.025

G

class	all paid	critical/other existing	delayed previously	existing paid	no credits/all paid
good	-0.081	-0.344	-0.169	-0.368	-0.063
bad	-0.100	-0.150	-0.100	-0.300	-0.092

H(Y,X)	1.767211

H

class	all paid	critical/other existing	delayed previously	existing paid	no credits/all paid
good	2.024	2.093	0.043	0.273	2.306
bad	3.573	7.259	0.000	0.000	4.243

sigma^2(U)	0.000157	21.813499
sigma(U)	0.012516	0.012516
u(0.95)	1.644853	
	Borne basse	Borne haute
I.C. à 90%	0.0289	0.0701

I

class	all paid	critical/other existing	delayed previously	existing paid	no credits/all paid
good	5.055	6.987	0.042	0.270	5.844
bad	11.625	15.915	0.097	0.629	13.468

P	59.929950
sigma^2(U)(0)	0.000158

Tests de significativité

Loi normalé	
sigma(U)[H0]	0.012576
z(U)	3.935544
p-value	0.000083

Loi du KHI-2	
C(U)	60.467109
DDL	4
p-value	2.313958E-12

Fig. 3.7. Tableau de contingence "Class" vs. "Credit history" - Calcul du U et tests associés

Variance asymptotique

La formule de la variance est un peu plus simple (on peut la reproduire sur un tableur en tous cas) :

$$\sigma_U^2 = \frac{1}{n^2 \times H(Y)^4} \times \left[\sum_c \sum_l n_{lc} \left\{ H(Y) \ln f_{l/c} + [H(X) - H(Y,X)] \ln f_{l.} \right\}^2 \right] \tag{3.14}$$

Exemple sur les données "CLASS" vs. "CREDIT HISTORY"

Détaillons les calculs à l'aide de notre feuille EXCEL (Figure 3.7) :

1. Nous formons le tableau de calcul des $H(X)$ (Tableau **E**), nous avons $H(X) = 1.186580$.
2. Nous calculons les fréquences conjointes (Tableau **F**) en vue d'obtenir le tableau de calcul de $H(Y,X) = 1.767211$ (Tableau **G**).
3. Enfin, le dernier tableau **H** correspond aux termes dans la double somme de l'équation 3.14.

Nous en extrayons la variance et l'écart type asymptotique

$$\sigma_U^2 = \frac{1}{1000^2 \times 0.610864^4} \times 21.813499 = 0.000157$$
$$\sigma_U = \sqrt{0.000157} = 0.012516$$

Intervalle de confiance

Pour un niveau de confiance $(1-\alpha)$, nous pouvons calculer l'intervalle de confiance

$$\left[U - u_{1-\frac{\alpha}{2}} \times \sigma_U; U + u_{1-\frac{\alpha}{2}} \times \sigma_U\right]$$

Pour le cas "CLASS vs. CREDIT HISTORY" de notre fichier, l'intervalle $[0.0289; 0.0701]$ a 90% de chances de contenir la "vraie" valeur de U (Figure 3.7).

3.4.3 Test de significativité basé sur la loi normale

Pour tester la significativité du coefficient ($H_0 : U_{Y/X} = 0$ vs. $H_1 : U_{Y/X} > 0$), nous pouvons nous appuyer sur la normalité asymptotique. Il faut bien entendu utiliser une estimation de la variance en rapport avec l'hypothèse nulle, elle est grandement simplifiée :

$$\sigma_U^2(0) = \frac{P - n \times [H(Y) + H(X) - H(Y,X)]^2}{n^2 \times H(Y)^2} \tag{3.15}$$

avec

$$P = \sum_l \sum_c n_{lc} \times \left(\ln \frac{n_{l.} \times n_{.c}}{n \times n_{lc}}\right)^2 \tag{3.16}$$

On forme le rapport

$$z_U = \frac{U}{\sigma_U(0)} \tag{3.17}$$

Et la région critique du test pour un risque α est définie par

$$R.C. : z_U > u_{1-\alpha}$$

Nous pouvons également lire directement la p-value du test pour accepter ou rejeter l'hypothèse nulle.

Exemple sur les données "CLASS" vs. "CREDIT HISTORY"

Reprenons notre feuille de calcul (Figure 3.7). Nous construisons le tableau **I** pour calculer la grandeur P. L'écart type sous l'hypothèse nulle est égale à

$$\sigma_U(0) = \sqrt{\frac{59.929950 - 1000 \times [0.610864 + 1.186580 - 1.767211]^2}{1000^2 \times 0.610864^2}} = 0.012576$$

Nous en déduisons la z-value $z(U) = \frac{0.049493}{0.012576} = 3.935544$, que nous comparons au quantile de la loi normale d'ordre 0.95, c.-à-d. 1.644853, pour un test à 5%. Nous concluons au rejet de l'hypothèse nulle.

3.4.4 Test de significativité basé sur la loi du χ^2

Il existe un autre test de significativité, **plus puissant**, basé sur la loi du χ^2. La statistique du test transforme le U de Theil de la manière suivante [4]

$$C(U) = 2 \times n \times H(Y) \times U_{Y/X} \tag{3.18}$$

Elle suit asymptotiquement une loi du χ^2 à $(L-1)(C-1)$ degrés de liberté.

C'est un test unilatéral. La région critique du test s'écrit :

$$R.C. : C(U) > \chi^2_{1-\alpha}[(L-1)(C-1)]$$

où $\chi^2_{1-\alpha}[(L-1)(C-1)]$ est le fractile d'ordre $1-\alpha$ de la loi du χ^2 à $(L-1)(C-1)$ degrés de liberté.

Nous pouvons également nous baser sur la probabilité critique $p_c = P\left(\chi^2 > C(U)\right)$ pour décider du rejet ou de l'acceptation de l'hypothèse nulle.

Exemple sur les données "CLASS" vs. "CREDIT HISTORY"

Nous calculons la statistique $C(U)$,

$$C(U) = 2 \times 1000 \times 0.610864 \times 0.049493 = 60.467109$$

La probabilité critique, avec une loi du χ^2 à $4 = (2-1) \times (5-1)$ degrés de liberté, est $p_c = 2.313958 \cdot 10^{-12}$. L'association est très significative.

3.5 Récapitulatif sur les mesures PRE

Essayons de reprendre tout cela pour y voir un peu plus clair. Toutes ces formules finissent par donner le tournis :

1. Les mesures PRE sont basées sur le principe de la réduction proportionnelle de l'erreur. Elles se distinguent par la manière avec laquelle elles quantifient l'erreur de prédiction.

4. Attention, l'entropie $H(Y)$ est bien calculée avec le logarithme népérien dans cette section!!! La transformation serait un peu différente si toutes nos formules étaient basées sur un logarithme à base 2.

2. Ce sont des mesures asymétriques par nature. On pourrait les rendre symétriques en calculant la moyenne des indicateurs calculés sur le tableau de contingence et sa transposée (section 3.6).
3. Elles varient varie forcément entre 0 (X n'apporte aucune information sur la connaissance de Y) et 1 (la connaissance de X permet de déterminer à coup sûr la valeur de Y).
4. Elles suivent asymptotiquement une loi normale. Nous pouvons nous appuyer sur cette propriété pour calculer les intervalles de confiance et mettre en place les tests de significativité.
5. Attention, la variance calculée est différente dans les deux cas. Nous devons utiliser la variance en accord avec l'hypothèse nulle pour le test de significativité.
6. Pour certaines mesures, il existe une transformation qui, généralement, suit une loi du χ^2. Lorsqu'elle est disponible, on préférera cette procédure car elle est plus puissante.
7. Pour apprécier et commenter les valeurs des coefficients PRE, en plus des tests de significativité, on peut leur associer une grille de lecture (Tableau 3.1).

PRE	"Force" du lien
$PRE < 0.1$	Faible
$0.1 \leq PRE < 0.2$	Modérée
$0.2 \leq PRE < 0.3$	Modérément forte
$0.3 \leq PRE$	Forte

Tableau 3.1. Grille de lecture des valeurs des mesures PRE

Très souvent, τ de Goodman-Kruskal et U de Theil fournissent des résultats similaires. Le λ de Goodmal-Kruskal est perturbé lorsque la modalité majoritaire est située sur la même ligne, quelle que soit la valeur de X. Dans ce cas il est mécaniquement égal à 0 même si les variables ne sont pas indépendantes. Cette mesure sera d'autant plus à éviter que la distribution initiale de Y est très déséquilibrée c.-à-d. une des modalités de Y est fortement présente dans le fichier, avec une fréquence marginale élevée.

Exemple sur les données "CLASS" vs. "CREDIT HISTORY"

Reprenons les résultats obtenus sur notre croisement fétiche "CLASS vs. CREDIT HISTORY" (Tableau 3.2). Les 3 mesures fournissent des résultats cohérents.

3.6 Mesures PRE symétriques

Les mesures PRE sont par nature asymétriques. Mais il est possible de déduire des versions symétriques en combinant simplement les indicateurs dans les deux sens, $PRE(Y/X)$ et $PRE(X/Y)$. La variante pour le U de Theil (section 3.4) est peut-être la plus naturelle de par sa construction.

Mesure	Valeur	90% I.C.	Significativité
λ	0.057	$[0.0064; 0.1069]$	Approx. Normale $z = 1.085$ $p_c = 0.071$
τ	0.062	$[0.0366; 0.0868]$	Approx. χ^2 $C = 61.63$ $p_c = 1.318 \cdot 10^{-12}$
U	0.049	$[0.0282; 0.0701]$	Approx. χ^2 $C = 60.47$ $p_c = 2.314 \cdot 10^{-12}$

Tableau 3.2. Récapitulatif, mesures PRE sur "CLASS vs. CREDIT HISTORY"

3.6.1 Coefficient d'incertitude symétrique

Le coefficient d'incertitude symétrique s'écrit :

$$U = 2\left[\frac{H(X) + H(Y) - H(X,Y)}{H(X) + H(Y)}\right] \tag{3.19}$$

Où $H()$ est l'entropie telle que définie dans la section 3.4, et $H(X,Y)$ l'entropie conjointe.

$$H(X,Y) = -\sum_l \sum_c f_{lc} \ln f_{lc} \tag{3.20}$$

Le coefficient U varie entre 0 et 1. La liaison est maximale lorsqu'elle tend vers 1. En cas d'indépendance, $H(X,Y) = H(X) + H(Y)$, et U vaut par conséquent 0.

"Class" vs. "Credit Duration"

Nous reprenons l'exemple du croisement entre "class" et "credit duration" de la section 3.4, mais avec un nouveau tableau de calcul (Figure 3.8).

Voyons en le détail :

— **"A"** est le tableau initial des effectifs n_{lc}.

— Nous avons les fréquences marginales et les logarithmes correspondants pour X et Y en respectivement **"B"** et **"C"**.

— En **"D"** sont représentés les fréquences conjointes et en **"D'"** son logarithme.

— A partir de ces informations, nous pouvons calculer successivement $H(X)$, $H(Y)$ et $H(X,Y)$. Le coefficient d'incertitude symétrique est égal à $U = 2\left[\frac{1.186580+0.610864-1.767211}{1.186580+0.610864}\right] = 0.033641$.

3.6.2 Intervalle de confiance

L'indicateur U est distribuée asymptotiquement selon la loi normale. Nous avons besoin de l'estimation de son écart-type pour calculer l'intervalle de confiance. Elle n'est pas des plus simples :

A : n_{lc}

class	all paid	critical/other existing	delayed previously	existing paid	no credits/all paid	Total
good	21	243	60	361	15	700
bad	28	50	28	169	25	300
Total	49	293	88	530	40	1000

B : f_{.l}

class	all paid	critical/other existing	delayed previously	existing paid	no credits/all paid
f_{.c}	0.049	0.293	0.088	0.530	0.040
ln(f_{.c})	-3.016	-1.228	-2.430	-0.635	-3.219

C : f_{c.}

class	f_{c.}	ln(f_{c.})
good	0.700	-0.3567
bad	0.300	-1.2040

D : f_{lc}

class	all paid	critical/other existing	delayed previously	existing paid	no credits/all paid
good	0.021	0.243	0.060	0.361	0.015
bad	0.028	0.050	0.028	0.169	0.025

D' : ln(f_{lc})

class	all paid	critical/other existing	delayed previously	existing paid	no credits/all paid
good	-3.863	-1.415	-2.813	-1.019	-4.200
bad	-3.576	-2.996	-3.576	-1.778	-3.689

H(X)	1.186580	H(X,Y)	1.767211
H(Y)	0.610864	U	0.033641

E : n_{c.} x n_{.l}

34300	205100	61600	371000	28000
14700	87900	26400	159000	12000

E' : H(X,Y) x ln[(Tableau E) / n^2]

-5.9601	-2.7997	-4.9254	-1.7523	-6.3188
-7.4575	-4.2971	-6.4227	-3.2496	-7.8161

F : [H(X) + H(Y) ln[(Tableau A) / n]

-6.9439	-2.5428	-5.0569	-1.8314	-7.5487
-6.4269	-5.3847	-6.4269	-3.1956	-6.6306

G : (Tableau E' - Tableau F)^2

0.96793	0.06599	0.01731	0.00626	1.51286
1.06216	1.18285	0.00002	0.00292	1.40553

sigma_U	0.008462	u(0.975)	1.96
		borne basse	0.017055
		borne haute	0.050226

H : [ln (Tableau E / (n x Tableau A)]^2

0.24071	0.02875	0.00069	0.00075	0.38957
0.41520	0.31830	0.00346	0.00372	0.53871

P	59.9299	sigma_U(0)	0.008548
		z_u	3.935544
		p-value	8.30085E-05

Fig. 3.8. Tableau de contingence "Class" vs. "Credit history" - Calcul du U symétrique et tests associés

$$\sigma_U = \frac{2}{n \times [H(X)+H(Y)]^2} \sqrt{\sum_l \sum_c n_{lc} \times \left\{ H(X,Y) \ln \frac{n_{c.} \times n_{.l}}{n^2} - [H(X)+H(Y)] \ln \frac{n_{lc}}{n} \right\}^2} \tag{3.21}$$

"Class" vs. "Credit Duration"

Le calcul a été décomposé en plusieurs étapes dans notre feuille Excel (Figure 3.8) :

— En **"E"**, nous avons le tableau sous indépendance (au facteur $\frac{1}{n}$ près).

— En **"E'"**, nous formons une partie de la formule située dans la double somme. La seconde est en **"F"**. **"G"** permet de réaliser la différence.

— La somme, le passage à la racine, puis la multiplication par le facteur précédent la racine sont regroupés dans la cellule associée à $\sigma_U = 0.008462$.

Pour un intervalle de confiance à $(1-\alpha)$ avec un quantile de la loi normale égal à $u_{1-\frac{\alpha}{2}}$, les bornes basses et hautes sont définies par :

$$U \pm u_{1-\frac{\alpha}{2}} \times \sigma_U \tag{3.22}$$

Soit, pour un intervalle à 95% :

$$borne\,basse = 0.033641 - 1.96 \times 0.008462 = 0.017055$$
$$borne\,haute = 0.033641 + 1.96 \times 0.008462 = 0.050226$$

3.6.3 Test de significativité

Pour le test de significativité, l'estimation de l'écart-type de l'indicateur est comme de coutume largement simplifié sous l'hypothèse nulle d'absence de liaison. Elle devient :

$$\sigma_U(0) = \frac{2}{n \times [H(X) + H(Y)]} \sqrt{P - n \times [H(Y) + H(X) - H(Y,X)]^2} \tag{3.23}$$

Où P est définie dans l'équation 3.16 plus haut.

La statistique de test $z_U = \frac{U}{\sigma_U(0)}$ contient au rejet de l'hypothèse nulle pour les valeurs excessivement élevées en valeur absolue.

"Class" vs. "Credit Duration"

Revenons à notre exemple (Figure 3.8). Le sous-tableau **"H"** sert à calculer la quantité $P = 59.9299$. Nous obtenons alors l'écart-type sous H_0 : $\sigma_U(0) = 0.008548$). Il est très légèrement plus élevé que la version précédente ($\sigma_U = 0.008462$). Ainsi, sur des petits effectifs, il peut y avoir des divergences entre cette procédure et l'utilisation de l'intervalle de confiance (recouvrement ou non de la valeur 0) pour les petits échantillons.

La statistique de test $z_U = \frac{0.033641}{0.008548} = 3.935544$, avec une p-value < 0.0001, nous indique le coefficient est largement significativement différent de zéro.

Partie II

Cas particuliers

4

Tables 2 × 2 - Cas des variables binaires

Le cas des variables binaires, variables nominales à 2 modalités, présente des particularités intéressantes. Le codage d'une telle variable en variable numérique est naturelle, on utilise souvent le codage 0/1 c.-à-d. une des modalités est codée 1, souvent il s'agit de la modalité qui nous intéresse (ex. pour les médecins, les personnes malades ; pour le banquier, le "bon" client ; pour les assurances, le client qui va contracter une nouvelle police ; etc.), la seconde modalité est codée 0. Il n'y a aucune perte d'information lors du passage d'un codage à l'autre.

Nous verrons dans ce chapitre que la manipulation des variables binaires, outre le fait que nous travaillons maintenant sur un tableau de contingence 2×2, introduit des particularités que nous pouvons exploiter.

La première est que nous pouvons déjà simplifier grandement les notations puisque nous avons un simple tableau 2×2 (Tableau 4.1).

Y vs. X	1	0
1	a	b
0	c	d

Tableau 4.1. Tableau générique 2×2

La statistique χ^2 peut s'écrire :

$$\chi^2 = \frac{n \times (ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)} \tag{4.1}$$

Croisement "Own telephone vs. Foreign Worker"

Pour illustrer ce chapitre, nous croiserons 2 variables binaires : "OWN TELEPHONE", qui indique si le client demandeur de crédit possède un numéro de téléphone enregistré à son nom, et "FOREIGN WORKER", qui indique si le demandeur de crédit est travailleur étranger ou non.

Nous présentons les 50 premières observations du fichier (Figure 4.1), avec le tableau de contingence (Tableau **A**, calculé sur les 1000 individus), le tableau sous hypothèse d'indépendance (Tableau **B**), et le tableau de calcul de la statistique du χ^2 (Tableau **C**).

Obs.	own_telephone	foreign_worker	Ytelephone	Xforeign
1	yes	yes	1	1
2	none	yes	0	1
3	none	yes	0	1
4	none	yes	0	1
5	none	yes	0	1
6	yes	yes	1	1
7	none	yes	0	1
8	yes	yes	1	1
9	none	yes	0	1
10	none	yes	0	1
11	none	yes	0	1
12	none	yes	0	1
13	yes	yes	1	1
14	none	yes	0	1
15	none	yes	0	1
16	none	yes	0	1
17	none	yes	0	1
18	none	yes	0	1
19	yes	yes	1	1
20	yes	yes	1	1
21	yes	yes	1	1
22	none	yes	0	1
23	none	no	0	0
24	none	yes	0	1
25	none	no	0	0
26	yes	yes	1	1
27	none	yes	0	1
28	none	yes	0	1
29	none	yes	0	1
30	yes	yes	1	1
31	yes	yes	1	1
32	none	yes	0	1
33	yes	yes	1	1
34	none	yes	0	1
35	yes	yes	1	1
36	none	yes	0	1
37	yes	yes	1	1
38	none	yes	0	1
39	yes	yes	1	1
40	none	yes	0	1
41	none	yes	0	1
42	yes	yes	1	1
43	yes	yes	1	1
44	none	yes	0	1
45	none	yes	0	1
46	none	yes	0	1
47	none	yes	0	1
48	yes	yes	1	1
49	none	yes	0	1
50	none	yes	0	1

Tableau de contingence (A)

NB own_telephone	foreign_worker	
own_telephone	no	yes	Total
none	32	564	596
yes	5	399	404
Total	37	963	1000

Tableau sous indépendance (B)

	no	yes	Total
none	22.052	573.948	596
yes	14.948	389.052	404
Total	37	963	1000

Calcul du KHI-2 (C)

	no	yes
none	4.487697442	0.172424512
yes	6.620464544	0.254368835

KHI-2 et PHI

KHI-2	11.534955
DDL	1
p-value	0.000683
PHI	0.107401

Coefficient de corrélation

r	0.107401
t	3.412656
ddl	998
p-value	0.000669

Calcul du KHI-2 - Correction de Yates (D)

	no	yes
none	4.047918738	0.155527511
yes	5.971682098	0.229441576

KHI-2 (Corrigé)

KHI-2	10.404570
DDL	1
p-value	0.001257

Fig. 4.1. "Own telephone vs. Foreign worker" - Traitement des variables binaires

Indicateur	Valeur
χ^2	11.534955
ddl	1
p-value	0.000683

Tableau 4.2. "OWN TELEPHONE vs. FOREIGN WORKER" - Résultats du test d'indépendance

La valeur du χ^2 peut être obtenue avec la formule simplifiée :

$$\chi^2 = \frac{1000 \times (32 \times 399 - 5 \times 564)^2}{596 \times 404 \times 37 \times 963} = 11.534955$$

Nous reprenons l'essentiel des résultats dans un tableau (Tableau 4.2). Le lien entre les 2 variables est largement significatif. Il semble y avoir une relation entre le fait d'avoir un téléphone enregistré à son nom et le fait d'être un travailleur étranger. En étudiant les contributions au χ^2, ce lien reposerait en grande partie sur une répulsion (-57%) entre le fait d'être natif (Foreign Worker = No) et la possession d'un téléphone enregistré à son nom (Own Telephone = yes). Quand je disais que ce fichier était bien étrange...

Les autres informations présentes dans la figure 4.2 seront commentés ultérieurement.

4.1 Coefficient ϕ et coefficient de corrélation

Avec des variables binaires, nous travaillons nécessairement sur des tableaux de contingence 2×2. Le degré de liberté est égal à 1. Une grande partie des indicateurs dérivés du χ^2 sont identiques, c'est le cas en particulier ϕ, V de Cramer et T de Tschuprow. De fait, ϕ maintenant varie entre 0 et 1.

Dans l'exemple "TELEPHONE vs. FOREIGN WORKER", nous avons $\phi = 0.10701$.

Autre information importante, et c'est quelque chose de tout à fait nouveau, $\phi = \sqrt{\frac{\chi^2}{n}}$ **est équivalent à la valeur absolue du coefficient de corrélation entre les variables codées 0/1**. Cela enrichit les possibilités d'interprétation ; cela peut nous ouvrir des portes lorsque nous voulons mettre en oeuvre des techniques plus complexes, nous verrons cela lorsque nous aborderons les associations partielles.

Arrêtons-nous un instant sur l'équivalence entre le coefficient de corrélation et le coefficient ϕ lorsque les variables sont binaires.

4.1.1 Coefficient de corrélation

Estimé sur un échantillon de taille n, le coefficient de corrélation empirique de Pearson entre deux variables continues Y et X est défini de la manière suivante ([8], section 2.3)

$$r = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_i (y_i - \bar{y})^2} \cdot \sqrt{\sum_i (x_i - \bar{x})^2}} \tag{4.2}$$

Il indique de degré de dépendance *linéaire* entre 2 variables quantitatives. C'est l'estimateur du coefficient de corrélation

$$\rho = \frac{COV(Y, X)}{\sigma_Y \times \sigma_X} \tag{4.3}$$

Lorsque les 2 variables Y et X sont indépendantes, nous avons $\rho = 0$, l'inverse n'est pas toujours vrai.

Pour mettre en place le test d'hypothèses

$$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases}$$

La statistique du test s'écrit :

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \tag{4.4}$$

Sous H_0, elle suit une loi de Student à $(n-2)$ degrés de liberté. La région critique du test au risque α est définie par

$$R.C. : |t| > t_{1-\frac{\alpha}{2}(n-2)}$$

Nous pouvons toujours exploiter la p-value pour décider du rejet ou de l'acceptation de l'hypothèse nulle.

Croisement "Own telephone vs. Foreign Worker"

Reprenons notre exemple "TELEPHONE vs. FOREIGN WORKER", nous avons rajouté 2 colonnes dans notre tableau de données. Nous avons adopté le codage suivant :

$$Y = \begin{cases} 1 & \text{, si OWN TELEPHONE = "yes"} \\ 0 & \text{, si OWN TELEPHONE = "none"} \end{cases}$$

$$X = \begin{cases} 1 & \text{, si FOREIGN WORKER = "yes"} \\ 0 & \text{, si FOREIGN WORKER = "no"} \end{cases}$$

Puis nous avons calculé le coefficient de corrélation [1] $r = 0.107401$. Nous observons effectivement que r est exactement identique à ϕ (Figure 4.2).

Nous avons formé la statistique $t = 3.412656$ et la probabilité critique du test est égale à $p_c = 0.000669$.

Les formulations des tests sont différentes, avec des lois de distribution différentes, il est normal que nous n'ayons pas exactement la même probabilité critique. Nous remarquerons néanmoins que les résultats sont très proches et cohérents. Que ce soit avec le test d'indépendance du χ^2 ou le test du coefficient de corrélation linéaire, nous concluons à l'existence d'un lien significatif entre "OWN TELEPHONE" et "FOREIGN WORKER".

4.1.2 Coefficient de corrélation : un pas vers les variables ordinales

En utilisant le codage 0/1 et en calculant le coefficient de corrélation, nous introduisons subrepticement une information supplémentaire : le sens de la variation. Pour preuve, le coefficient de corrélation peut prendre des valeurs négatives alors que ϕ est toujours positif. Nous mesurons, lorsque X passe de 0 à 1, la propension de Y à passer de 0 à 1 ($r > 0$) ou de 1 à 0 ($r < 0$). Cela ne porte pas trop à conséquence lorsque les variables sont binaires. Il nous faut tout simplement garder à l'esprit que le signe du coefficient de corrélation empirique obtenu sur les données dépend du codage adopté : si nous inversons le codage d'une des variables, le signe de r est modifié.

Nous étudions plus en détail le traitement des variables ordinales dans la partie IV de ce support.

4.2 χ^2 et correction de continuité pour les petits effectifs

Lors de la présentation du test d'indépendance du χ^2 (Section 2.2), nous mettions l'accent sur les conditions de validité du test. La modélisation à l'aide de la loi du χ^2 des valeurs de la statistique est dégradée dès lors que les effectifs théoriques, sous l'hypothèse d'indépendance, sont trop faibles ($e_k < 5$ approximativement pour au moins une des cases du tableau). Il ne semblait pas y avoir de véritable solution mis à part les recommandations de prudences adressés aux praticiens.

1. Nous avons utilisé la fonction COEFFICIENT.CORRELATION(.) du tableur EXCEL

Dans le cas particulier des tables 2×2, il existe une manière de remédier à cela en introduisant une correction de continuité, dite correction de Yates[2]. Elle consiste en une modification de la statistique du test. La loi associée sous l'hypothèse nulle et les degrés de libertés restent identiques. Attention, il est entendu que cette correction n'a de sens que pour les tables 2×2. La statistique corrigée s'écrit :

$$\chi^2_{Yates} = \sum_k \frac{(|o_k - e_k| - 0.5)^2}{e_k} \tag{4.5}$$

Les avis sont partagés quant à l'opportunité d'introduire cette correction dans la pratique. D'autant plus qu'elle n'est théoriquement justifiée que lorsque les marges sont fixées, c.-à-d. si l'on répète l'échantillonnage, on devrait obtenir les mêmes marges lignes et colonnes. Cela n'est bien entendu possible que dans les expérimentations cliniques, etc. (Howell, page 167).

Pour ou contre la correction de Yates, l'information la plus importante à mon sens est qu'elle rend le test plus conservateur c.-à-d. elle a tendance à favoriser l'hypothèse nulle. Lorsque les effectifs sont élevés, cette correction est inopérante.

Croisement "Own telephone vs. Foreign Worker"

Nous avons introduit la correction de Yates dans notre exemple, nous avons formé la statistique du test et calculé la p-value (Figure 4.2, dernière partie de la feuille de calcul, en bas à droite). Nous constatons que le χ^2 calculé a été diminuée $\chi^2_{Yates} = 10.404570$, la p-value associée est plus grande $p_c = 0.001257$. La correction reste très mineure sur un fichier de $n = 1000$ observations.

4.3 Coefficient $\mathcal{Q}$ de Yule basé sur les comparaisons par paires

Le coefficient $\mathcal{Q}$ de Yule, tout comme le coefficient de corrélation, s'appuie sur le codage 0/1 pour mesurer l'association entre 2 variables binaires. A la différence qu'il exploite différemment les informations en introduisant le principe de *la comparaison par paires*. Ce coefficient est un cas particulier du γ de Goodman et Kruskal destiné à mesurer l'association entre les variables ordinales.

4.3.1 Comparaison par paires

La "comparaison par paires" consiste à comparer 2 à 2 les observations du fichier à partir des valeurs prises par les variables Y et X, toutes les deux codées 0/1 dans cette section. Nous définissons alors 3 types de situations :

1. Les paires concordantes sont les couples d'observations (i, j) où (a) $(y_i > y_j)$ et $(x_i > x_j)$ ou (b) $(y_i < y_j)$ et $(x_i < x_j)$. Nous noterons P le nombre de paires concordantes.
2. Les paires discordantes sont les couples d'observations où (a) $(y_i > y_j)$ et $(x_i < x_j)$ ou (b) $(y_i < y_j)$ et $(x_i > x_j)$. Nous noterons Q le nombre de paires discordantes[3].
3. Les ex-aequos sont les couples pour lesquelles $(y_i = y_j)$ ou $(x_i = x_j)$

Numéro	Y	X
1	1	0
2	1	1
3	0	0
4	1	0
5	0	1

Tableau 4.3. Exemple de données pour les comparaisons par paires

Prenons l'exemple du fichier de données avec 5 observations (Tableau 4.3). Nous pouvons former $5^2 = 25$ couples d'observations. Nous annoterons les paires (Tableau 4.4), nous observons $P = 2$ paires concordantes, $Q = 4$ discordantes et 19 ex-aequos.

Paire	État	Paire	État
1,1	*	3,4	*
1,2	*	3,5	*
1,3	*	4,1	*
1,4	*	4,2	*
1,5	D	4,3	*
2,1	*	4,4	*
2,2	*	4,5	D
2,3	C	5,1	D
2,4	*	5,2	*
2,5	*	5,3	*
3,1	*	5,4	D
3,2	C	5,5	*
3,3	*	-	-

Tableau 4.4. Décompte des ex-aequos (*), des paires concordantes (C) et discordantes (D)

4.3.2 Coefficient $\mathcal{Q}$ de Yule

Définition

Dans la population, le coefficient $\mathbb{Q}$ de Yule est le rapport de probabilité (Siegel, 1988 ; page 292)

$$\begin{aligned}\mathbb{Q} &= \frac{P(\text{Y et X concordants}) - P(\text{Y et X discordants})}{1 - P(\text{Y et X ex-aequo})} \\ &= \frac{P(\text{Y et X concordants}) - P(\text{Y et X discordants})}{P(\text{Y et X concordants}) + P(\text{Y et X discordants})}\end{aligned}$$

2. http://en.wikipedia.org/wiki/Yates'_correction_for_continuity
3. La notation est un peu malheureuse, j'en conviens, mais c'est celle qui fait référence dans les ouvrages.

Pour 2 individus extraits au hasard dans la population, si $\mathbb{Q} = 1$ (resp. $\mathbb{Q} = -1$), il est certain qu'ils présentent des valeurs concordantes (resp. discordantes) c.-à-d. Y est positivement (resp. négativement) lié à X. Lorsque les variables Y et X sont indépendants, $\mathbb{Q} = 0$.

Sur un échantillon, le coefficient $\mathcal{Q}$ de Yule est défini par le rapport :

$$\mathcal{Q} = \frac{P - Q}{P + D} \tag{4.6}$$

Il s'interprète comme le surplus de paires concordantes sur les paires discordantes sur l'ensemble des paires, en excluant les ex-aequos.

Dans notre exemple ci-dessus, $\mathcal{Q} = \frac{2-4}{2+4} = -0.33$.

Calcul sur un tableau de contingence 2×2

Il est évident qu'utiliser la procédure de calcul ci-dessus sur un fichier comportant des centaines d'observations n'est pas tenable. De plus, les données sont souvent directement fournies sous forme de tableau de contingence, il faut proposer une formule qui s'applique directement sur la table 2×2.

En partant du tableau générique 2×2 (Tableau 4.1), le nombre de paires concordantes est égal à $P = 2ad$ et le nombre de paires discordantes $Q = 2bc$. Nous en déduisons le coefficient $\mathcal{Q}$ de Yule, après simplification,

$$\mathcal{Q} = \frac{ad - bc}{ad + bc} \tag{4.7}$$

Reprenons notre jeu de données (Tableau 4.3), nous formons le tableau de contingence (Tableau 4.5) et nous obtenons

$$\mathcal{Q} = \frac{1 \cdot 1 - 2 \cdot 1}{1 \cdot 1 + 2 \cdot 1} = \frac{-1}{3} = -0.33$$

Y vs. X	1	0
1	1	2
0	1	1

Tableau 4.5. Tableau de contingence pour les données exemples (Tableau 4.3)

Exemple "Own telephone vs. Foreign Worker"

Dans cet exemple (Figure 4.2), réorganisé suivant le codage 0/1 (Tableau 4.6), il devient plus facile de calculer le $\mathcal{Q}$ de Yule, il est égal à

$$\mathcal{Q} = \frac{399 \cdot 32 - 5 \cdot 564}{399 \cdot 32 + 5 \cdot 564} = \frac{9948}{15588} = 0.638183$$

Tel. x F.Wker	1	0
1	399	5
0	564	32

Tableau 4.6. Tableau réorganisé pour le croisement "Telephone vs. Foreign Worker"

La liaison est positive. *Compte tenu de notre codage*, cela indique que les travailleurs étrangers (Foreign worker = yes) ont plus tendance à avoir un téléphone enregistré à leur nom (Own telephone = yes). Nous étions parvenu à la même conclusion avec le coefficient de corrélation.

Grille de lecture des valeurs de $\mathcal{Q}$

Certains auteurs [4] proposent une grille de lecture de la valeur de $\mathcal{Q}$ (Tableau 4.7). Pourquoi pas. Cela nous permet d'avoir des repères. Il reste que la véritable évaluation passe par les techniques de statistique inférentielle.

Plage de valeurs	Intensité de la liaison
$\lvert\mathcal{Q}\rvert = 0$	nulle
$0.01 \leq \lvert\mathcal{Q}\rvert \leq 0.09$	négligeable
$0.10 \leq \lvert\mathcal{Q}\rvert \leq 0.49$	légère
$0.50 \leq \lvert\mathcal{Q}\rvert \leq 0.69$	forte
$0.70 \leq \lvert\mathcal{Q}\rvert \leq 1$	très forte

Tableau 4.7. Grille de lecture du coefficient $\mathcal{Q}$ de Yule

Intervalle de confiance

Le coefficient $\mathcal{Q}$ de Yule suit asymptotiquement une loi normale. En produisant une estimation de la variance, nous avons la possibilité de calculer l'intervalle de variation.

Le coefficient de Yule étant un cas particulier du γ de Goodman et Krusal, il en est de même en ce qui concerne la variance. L'expression générique de cette dernière est pour le moins compliquée. Fort heureusement, puisque nous travaillons sur un tableau 2×2, nous pouvons la simplifier.

La variance s'écrit

$$\sigma^2_{\mathcal{Q}} = \frac{16}{(P+Q)^4}\left[a(Qd)^2 + b(Pc)^2 + c(Pb)^2 + d(Qa)^2\right] \tag{4.8}$$

L'intervalle de confiance au niveau $(1-\alpha)$ est tout naturellement

$$\left[\mathcal{Q} - u_{1-\frac{\alpha}{2}} \times \sigma_{\mathcal{Q}}; \mathcal{Q} + u_{1-\frac{\alpha}{2}} \times \sigma_{\mathcal{Q}}\right] \tag{4.9}$$

où $u_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi normale centrée réduite.

4. http://www.aly-abbara.com/utilitaires/statistiques/khi_carre_rr_odds_ratio_ic.html

Exemple "Own telephone vs. Foreign Worker"

Dans notre exemple (Tableau 4.6), nous avions obtenu $\mathcal{Q} = 0.638183$. Le nombre de paires concordantes (resp. discordantes) est $P = 2 \cdot 399 \cdot 32 = 25536$ (resp. $Q = 2 \cdot 564 \cdot 5 = 5640$). Nous estimons la variance

$$\begin{aligned}\sigma^2_{\mathcal{Q}} &= \frac{16}{(25536 + 5640)^4}\left[399 \cdot (5640 \cdot 32)^2 + 5 \cdot (25536 \cdot 564)^2 + 564 \cdot (25536 \cdot 5)^2 + 32 \cdot (5640 \cdot 399)^2\right] \\ &= 0.020687\end{aligned}$$

L'écart type, $\sigma_{\mathcal{Q}} = \sqrt{0.020687} = 0.143828$, nous permet d'élaborer l'intervalle de confiance à 90% dont les bornes basses et hautes sont

$$\begin{aligned}bb(\mathcal{Q}) &= 0.638183 - 1.644853 \times 0.143828 = 0.401607 \\ bh(\mathcal{Q}) &= 0.638183 + 1.644853 \times 0.143828 = 0.874759\end{aligned}$$

Test de significativité

Pour le test de significativité, il nous faut produire une autre estimation de la variance, en accord avec l'hypothèse nulle d'indépendance entre Y et X. Le test à mettre en oeuvre est le suivant

$$\begin{aligned}H_0 &: \mathbb{Q} = 0 \\ H_1 &: \mathbb{Q} \neq 0\end{aligned}$$

Il s'agit bien d'un test bilatéral puisque $\mathbb{Q}$ est défini sur l'intervalle $[-1; +1]$.

La variance asymptotique sous l'hypothèse nulle s'écrit

$$\sigma^2_{\mathcal{Q}}(0) = \frac{4}{(P+Q)^2}\left[a(d)^2 + b(c)^2 + c(b)^2 + d(a)^2 - \frac{1}{n}(P-Q)^2\right] \tag{4.10}$$

La statistique $z_{\mathcal{Q}} = \frac{\mathcal{Q}}{\sigma_{\mathcal{Q}}(0)}$ suit une loi normal centrée réduite. Au risque α, la région critique est définie par

$$R.C. : |z_{\mathcal{Q}}| > u_{1-\frac{\alpha}{2}}$$

Plus simplement, nous pouvons comparer la p-value du test avec le risque α.

Exemple "Own telephone vs. Foreign Worker"

Dans notre exemple, nous obtenons la variance sous H_0

$$\begin{aligned}\sigma^2_{\mathcal{Q}}(0) &= \frac{4}{(25536+5640)^2}\left[399(32)^2+5(564)^2+564(5)^2+32(399)^2-\frac{1}{1000}(25536-5640)^2\right] \\ &= 0.027622\end{aligned}$$

Nous en déduisons la statistique $z_{\mathcal{Q}} = \frac{0.638183}{\sqrt{0.027622}} = \frac{0.638183}{0.166198} = 3.839886$ et la p-value du test $p_c = 0.000123$, largement en-deçà du risque $\alpha = 10\%$ que l'on s'est choisi. Il y a un lien très significatif entre les variables étudiées.

4.4 Test de Mc Nemar - Comparaison de proportions pour échantillons appariés

Le test de Mc Nemar s'applique aux comparaisons de proportions sur des échantillons appariés.

On l'utilise souvent dans les situations où l'on évalue l'état d'une variable binaire (ex. apprécier (1) ou pas (0) un produit, bien dormir ou pas, etc.) avant X et après Y un traitement (ex. avant et après avoir regardé une publicité à la TV, selon que l'on ingurgite ou pas une infusion de camomille au moment de se coucher, etc.).

Mais il s'applique en réalité dans toute situation où l'appariement est pertinent. Nous pouvons l'utiliser par exemple pour comparer la proportion des hauts salaires (ou des sportifs invétérés, ou des fan de hard rock, etc.) chez les hommes et les femmes qui sont en couple.

Formellement, on oppose les probabilité d'occurrence des cas favorables dans les deux configurations c.-à-d.

$$H_0 : P(Y = 1) = P(X = 1)$$

En pratique, les données peuvent d'exprimer sous la forme d'un tableau de contingence 2×2 (Tableau 4.1) qui identifie les changements d'états. Si l'on reprend l'exemple de l'impact de la publicité sur l'appréciation d'un produit :

— "a" et "d" sont les personnes qui sont restés sur leur avis initial.

— "b" sont les personnes pour lesquels la publicité a eu un impact positif (passées de $X = 0$ à $Y = 1$).

— "c" ont été dégoûtés par la publicité (passées de $X = 1$ à $Y = 0$).

De fait, la statistique de test s'appuie exclusivement sur les changements d'états :

$$\chi^2_{MN} = \frac{(b-c)^2}{b+c} \tag{4.11}$$

Sous l'hypothèse nulle, elle suit une loi du $^c hi^2$ à 1 degré de liberté. La région critique est située du côté des valeurs excessivement élevées.

Pour les petits effectifs, une correction de continuité peut être introduite :

$$\chi^2_{MN} = \frac{(|b-c|-1)^2}{b+c} \tag{4.12}$$

Approbation de l'action du Président

Nous prenons un exemple tiré de l'ouvrage d'Agresti ([1], section 10.1.3). Un sondage a été effectué auprès de 1600 personnes en âge de voter. 944 disent approuver l'action du Président. Un mois plus tard, pour la même question posée auprès des mêmes personnes, elles sont 880 à avoir un avis positif. Peut-on en conclure qu'il y a un revirement significatif de l'opinion des personnes vis-à-vis du Président ?

Opinion	Avant = Approuve	Avant = Désapprouve	Total
Après = Approuve	794	86	880
Après = Désapprouve	150	570	720
Total	944	656	1600

Tableau 4.8. Approbation du Président (Voir [1], Présentation transposée par rapport à Table 10.1)

Nous calculons la statistique de test, sans correction de continuité.

$$\chi^2_{MN} = \frac{(150-86)^2}{150+86} = 17.36$$

Avec une p-value < 0.001, nous pouvons conclure à un changement d'opinion manifeste au sein de la population. Le mois passé a dû être désastreux pour l'équipe présidentielle.

Remarque 17 (Une statistique Mc Nemar signée). Il est possible de travailler avec

$$z_{MN} = \frac{b-c}{\sqrt{b+c}}$$

Qui suit asymptotiquement une loi normale. Les conclusions sont équivalentes. L'intérêt ici est que l'on a en plus le sens de l'évolution. z_{MN} est positif lorsque le changement d'état est favorable, il est négatif sinon. Pour notre exemple, nous avons $z_{MN} = \frac{86-150}{\sqrt{86+150}} = -4.17$.

Remarque 18 (Une statistique alternative pour le test de comparaison de proportions pour échantillons appariés). Dans ([13], section 15.2.1), un test du rapport de vraisemblance basé sur la statistique déviance est proposé :

$$d = 2(b \times \ln \frac{2b}{b+c} + c \times \ln \frac{2c}{b+c}) \tag{4.13}$$

Sous H_0, elle est également distribuée selon une loi du χ^2 à 1 degré de liberté.

Pour nous exemple des adoubements présidentiels, nous obtenons :

$$d = 2(86 \times \ln \frac{2 \times 86}{86+150} + 150 \times \ln \frac{2 \times 150}{86+150}) = 17.58$$

Avec les mêmes conclusions.

5

Risque relatif, Odds et Odds-Ratio

5.1 Association Facteur / Maladie - Tableau de contingence 2×2

Il existe une autre manière d'exploiter les tableaux de contingence 2×2, très populaire en statistique épidémiologique (Renaud, 1986). Il s'agit de l'association "Facteur - Maladie". L'objectif est de quantifier dans quelle mesure l'exposition d'un individu à un facteur de risque (ex. tabagisme, alcool, amiante, etc.) entraîne une apparition accrue de la maladie (ex. cancer, maladie cardio-vasculaire, etc.). Dans le tableau 5.1, nous effectuons un croisement entre le fait de fumer et l'occurrence d'une maladie cardio-vasculaire. Le $\chi^2 = 20.23$, il est largement significatif avec une p-value $p_c < 0.001$. Il y a un lien entre ces deux variables. Le tout est de caractériser au mieux la nature de ce lien. Certes, avec l'étude des contributions, nous pourrions déjà affiner l'analyse. Mais nous verrons qu'avec les notions abordés dans ce chapitre, l'exploitation des résultats est mieux adaptée au contexte.

Tableau	Fumeur (1)	Non-fumeur (0)	Total
Malade (+)	72	32	104
Non-malade (-)	36	60	96
Total	108	92	200

Tableau 5.1. Association "Facteur d'exposition (Fumer) - Maladie (Cardiaque)" (Renaud, 1986 ; page 56)

Même si l'épidémiologie est un cadre privilégié, ce type d'analyse peut s'appliquer dans beaucoup d'autres domaines dès lors que nous voulons évaluer l'occurrence (le surcroît d'occurrence devrait-on dire) d'une modalité particulière (la modalité positive "+") de la variable d'intérêt Y chez un groupe d'individu particulier (le groupe des exposés - Groupe 1) par rapport à un groupe de référence (le groupe témoin - Groupe 0). Nous pouvons résumer la situation dans un tableau de contingence générique (Tableau 5.2).

Remarque 19 (Formats des tableaux). **Attention.** Par rapport aux standards de l'étude des associations "facteurs vs. maladies", notre tableau de travail est transposé. Tout simplement parce que dans ce support, nous avons pour coutume de mettre en ligne la variable dépendante Y, en colonne la variable indépendante X (Voir par ex. les mesures asymétriques). Il m'est paru hasardeux de déroger à cela pour un seul chapitre,

au risque de perdre le lecteur. Dans les supports associés à l'épidémiologie, l'habitude est de mettre en ligne le facteur d'exposition, en colonne l'occurrence ou non de la maladie. Il faudra y penser lorsque vous essayerez de rapprocher les formules décrites ici et ailleurs.

Tableau	Exposé (1)	Témoin (0)	Total
Positif (+)	a	b	a+b
Négatif (-)	c	d	c+d
Total	a+c	b+d	n = a+c+b+d

Tableau 5.2. Table générique pour le calcul des odds et odds-ratio

Données CREDIT GERMAN : association "CLASS vs. PURPOSE"

Pour illustrer notre propos, revenons aux données GERMAN CREDIT. Admettons [1] que historiquement la banque assumait uniquement certains crédits à la consommation, plus particulièrement les crédits pour les équipements ("purpose = furniture / equipment") et l'achat de matériel hi-fi ("purpose = radio /tv"). L'arrivée d'une nouvelle direction a un peu chamboulé tout ça. L'entreprise s'est diversifiée et assume maintenant d'autres types de crédits. Les vieux grognards (les éléphants) montent au créneau et soutiennent que ces nouveaux crédits sont plus risqués, c.-à-d. emmènent plus souvent des clients défaillants (class = bad) et mèneront l'entreprise à la faillite. Ils nous demandent de démontrer cela avec les données [2].

Tableau de contingence initial (A)

class	business	domestic appliance	education	new car	other	repairs	retraining	used car	furniture/equipment	radio/tv	Total
bad	34	4	22	89	5	8	1	17	58	62	300
good	63	8	28	145	7	14	8	86	123	218	700
Total	97	12	50	234	12	22	9	103	181	280	1000

Fig. 5.1. "Class vs. Purpose" - Tableau original

Notre premier réflexe est de construire le tableau de contingence croisant CLASS et PURPOSE (Figure 5.1). Même s'il contient toutes les données susceptibles de nous intéresser, sa lecture est difficile car les informations importantes sont éparpillées. Dans notre étude, le groupe témoin est "purpose = furniture / equipment + radio /tv". La modalité positive à analyser est "class = bad", les clients défaillants. Nous effectuons la réorganisation adéquate de manière à obtenir un second tableau de contingence (Figure 5.2).

Le χ^2 calculé sur ce tableau est égal à $\chi^2 = 6.417904$, il est significatif à 5% puisque la p-value $p_c = 0.011298$. Il semble y avoir un lien entre l'objet du crédit et la défaillance. En étudiant les contributions, nous constaterons essentiellement une attraction +32.27% entre "Groupe Exposé" et "Class = bad",

1. Tout ceci est bien entendu totalement fictif mais les statistiques sont tellement plus agréables lorsqu'on fait raconter une histoire aux chiffres...

2. Et voilà, je me demandais à quel moment ça allait nous retomber dessus. "Quand les éléphants se battent, c'est le gazon qui est piétiné..." (Vieux proverbe africain).

Tableau de contingence après regroupement (B)			
	Groupe(Exposé = 1)	Groupe(Témoin=0)	Total
bad (+)	180	120	300
good (-)	359	341	700
Total	539	461	1000

Fig. 5.2. "Class vs. Purpose" - Tableau après regroupement

et une répulsion -37.73% entre "Groupe témoin" et "Class = bad". Ces deux cases portant 70% de l'information utile dans le tableau. Ces résultats semblent accréditer les déductions de nos grognards : les nouveaux crédits contractés posent problème.

Pour intéressants qu'ils soient, ces résultats sont très difficiles à faire passer. Allez expliquer à des néophytes des χ^2 et des contributions. Les personnes hostiles auront vite fait de tout rejeter parce qu'ils ne comprennent pas, ou parce que ça les arrange de ne pas comprendre. Ils remettront en cause vos résultats, et enchaîneront rapidement pour remettre en cause votre existence même au sein de l'entreprise. Il nous faut donc utiliser des notions simples pour expliciter nos résultats et convaincre l'auditoire. La notion de **risque** est justement une notion que tout un chacun peut comprendre aisément [3].

Dans ce chapitre, nous présentons plusieurs indicateurs basés principalement sur des rapports de probabilités.

5.2 Probabilités conditionnelles

Les probabilités conditionnelles permettent de comparer les situations. Dans ce cas, nous nous intéresserons aux profils colonnes. Pour les estimations, notre référence sera le tableau des effectifs (Tableau 5.2), nous utiliserons l'exemple du croisement "CLASS vs. PURPOSE Regroupé" pour illustrer notre propos (Figure 5.2).

Nous distinguons les probabilités suivantes :

— $P(+)$ la probabilité d'être positif dans la population, estimée par $f_+ = \frac{a+b}{a+b+c+d}$. Dans notre exemple, la probabilité d'être défaillant est $f_+ = \frac{300}{1000} = 30\%$. Attention, cette estimation n'est valable que si le fichier a été construit à partir d'un échantillonnage aléatoire simple dans la population. Dans tout autre cas, cette estimation ne reflète que le mode de constitution du fichier.

— $P(+/0)$ la probabilité d'être positif sachant que l'on est dans le groupe témoin. Si l'on se réfère au tableau de contingence observé, elle est estimée par $f_{+/0} = \frac{b}{b+d}$. Dans notre exemple, elle est égale à $f_{+/0} = \frac{120}{461} = 0.2603$. Les personnes du groupe témoin, c.- -à-d. qui contractent les crédits "classiques", ont 26.03% de "chances" d'être défaillants.

— $P(+/1)$ est la probabilité d'être positif sachant que l'on est dans le groupe exposé. Il est estimé par $f_{+/1} = \frac{a}{a+c} = \frac{180}{539} = 0.3340$. La probabilité d'être défaillant est de 33.40% chez les personnes contractant les nouveaux crédits.

— De la même manière, nous pouvons déduire $f_{-/0} = 1 - f_{+/0} = 73.97\%$ et $f_{-/1} = 1 - f_{+/1} = 66.60\%$.

3. Ex. Si on prend le volant après avoir vidé une bouteille de Whisky, on augmente les risques d'avoir un accident..., j'imagine qu'il n'est pas nécessaire de faire des études mirobolantes pour comprendre cela.

Premier constat intéressant, la proportion de personnes défaillantes paraît plus élevé dans le groupe exposé (aux nouveaux crédits). On pourrait mettre en place un test de comparaison de proportions *classique*[4] pour évaluer la significativité de la différence. On préférera cependant construire une autre statistique qui permet, d'une part, de réaliser ce test de comparaison, et d'autre part, propose une interprétation intuitive.

5.3 Risque relatif

5.3.1 Définition et estimation

Le risque relatif compare les deux probabilités $P(+/1)$ et $P(+/0)$ en utilisant le ratio :

$$RR = \frac{P(+/1)}{P(+/0)} \tag{5.1}$$

Son principal intérêt est son interprétation : **il indique le surcroît de chances d'être positif du groupe exposé par rapport au groupe témoin**. S'il est égal à 1, on n'a pas plus de chances d'être positif dans le groupe exposé que dans le groupe témoin. S'il est significativement différent de 1, cela indique qu'il y a une différence entre les probabilités.

Sur un échantillon (Tableau 5.2), il est estimé par

$$rr = \frac{a/(a+c)}{b/(b+d)} = \frac{a \times (b+d)}{b \times (a+c)} \tag{5.2}$$

Dans notre exemple "CLASS vs. PURPOSE regroupé", il est égal à $rr = \frac{0.3340}{0.2603} = 1.2829$. On le lit : "par rapport au groupe témoin, les individus du groupe exposé ont 1.2829 fois plus de chances d'être positif", ou encore, "il y a 1.2829 fois plus de chances d'être positif dans le groupe exposé que dans le groupe témoin".

Reste à savoir maintenant si c'est significativement $\neq 1$ pour qu'on s'en inquiète.

5.3.2 Intervalle de confiance et test de significativité

Distribution asymptotique et intervalle de confiance

L'estimateur rr du risque relatif est défini sur $]0; +\infty[$. On préfère généralement manipuler le logarithme $\ln(rr)$, qui varie entre $]-\infty; +\infty[$. En effet, il suit asymptotiquement une loi normale de moyenne et de variance

$$E[\ln(rr)] = \ln(RR) \tag{5.3}$$

$$V[\ln(rr)] = \frac{1 - f_{+/1}}{a} + \frac{1 - f_{+/0}}{b} \tag{5.4}$$

4. http://www.eao.chups.jussieu.fr/polys/biostats/poly/POLY.Chp.12.1.2.html

La variance peut s'écrire d'une autre manière

$$\sigma^2_{\ln(rr)} = V[\ln(rr)] = \frac{1}{a} \times \frac{c}{a+c} + \frac{1}{b} \times \frac{d}{b+d} \tag{5.5}$$

Ou encore

$$\sigma^2_{\ln(rr)} = V[\ln(rr)] = \frac{1}{a} - \frac{1}{a+c} + \frac{1}{b} - \frac{1}{b+d} \tag{5.6}$$

Nous disposons d'un estimateur, elle est distribuée selon une loi normale, nous avons de plus une estimation de la variance, il est facile de produire l'intervalle de confiance au niveau $(1-\alpha)$ de $\ln(rr)$

$$[\ln(rr) - u_{1-\frac{\alpha}{2}} \times \sigma_{\ln(rr)}; \ln(rr) + u_{1-\frac{\alpha}{2}} \times \sigma_{\ln(rr)}] \tag{5.7}$$

Pour le ramener à l'estimation rr, nous passons à l'exponentielle, l'intervalle de variation devient

$$rr \in [e^{\ln(rr) - u_{1-\frac{\alpha}{2}} \times \sigma_{\ln(rr)}}; e^{\ln(rr) + u_{1-\frac{\alpha}{2}} \times \sigma_{\ln(rr)}}] \tag{5.8}$$

"Class vs. Purpose regroupé

Dans notre exemple, pour un niveau de confiance de 90%, nous retraçons le détail des calculs dans le tableau ci-après

Indicateur	Valeur
rr	1.2829
$\ln(rr)$	0.2491
$\sigma_{\ln(rr)}$	$\sqrt{\frac{1}{180} \times \frac{259}{539} + \frac{1}{120} \times \frac{341}{461}} = \sqrt{0.0099} = 0.0993$
$u_{0.95}$	1.6449
Borne basse ($\ln(rr)$)	0.0858
Borne haute ($\ln(rr)$)	0.4125

Pour obtenir au final l'intervalle de variation de rr

$$[1.0896; 1.5106]$$

Information importante, l'intervalle ne contient pas la valeur caractéristique 1. Il semble que les nouveaux crédits soient plus risqués que les crédits traditionnels. Au pire cas (ou dans le meilleur des cas, tout dépend du point de vue), au niveau de confiance 90%, ils ont 1.0896 fois plus de chances d'emmener un client défaillant.

Test d'indépendance

L'autre procédure pour tester le risque relatif est d'utiliser un test d'hypothèses. Lorsque le facteur d'exposition (X) et l'occurrence de la maladie (Y) sont indépendant, $RR = 1$, et par conséquent $\ln(RR) = 0$.

Tester l'indépendance entre X et Y revient ainsi à vérifier la significativité du risque relatif

$$H_0 : \ln(RR) = 0$$
$$H_1 : \ln(RR) \neq 0$$

L'expression de la variance de l'estimateur $\ln(rr)$ n'est pas modifiée, la statistique z du test est définie par

$$z_{\ln(rr)} = \frac{\ln(rr)}{\sqrt{\frac{1}{a} \times \frac{c}{a+c} + \frac{1}{b} \times \frac{d}{b+d}}} \tag{5.9}$$

Et la région critique du test pour un niveau de risque α

$$R.C. : |z_{\ln(rr)}| > u_{1-\frac{\alpha}{2}}$$

"Class vs. Purpose regroupé

En utilisant les valeurs déjà calculées précédemment, nous obtenons $z_{\ln(rr)} = \frac{0.2491}{0.0993} = 2.5085$, à comparer avec $u_{0.95} = 1.6449$ pour un test à 10%. Nous pouvons aussi calculer la p-value du test $p_c = 0.0121$. Au risque de 10%, nous considérons que, effectivement, les grognards de l'organisme de crédit ont de bonnes raisons de s'inquiéter.

5.3.3 Travailler sur les différences de proportions

Comparaison de proportions de deux échantillons indépendants

Plutôt que de travailler sur le ratio des probabilités $P(+/1)$ et $P(+/0)$ qu'est le risque relatif, nous pouvons travailler sur la différence pour le test d'indépendance. Nous avons ni plus ni moins qu'un test de différence entre deux proportions ([9], section 3.1).

Les hypothèses à confronter s'écrivent :

$$H_0 : P(+/1) = P(+/0)$$
$$H_1 : P(+/1) \neq P(+/0)$$

Nous pouvons former la statistique de test avec :

$$\begin{aligned} DT &= f_{+/1} - f_{+/0} \\ &= \frac{a}{a+c} - \frac{b}{b+d} \end{aligned}$$

Dont la variance, sous l'hypothèse d'égalité des proportions, est :

$$V_0(DT) = \frac{f_{+/1}(1-f_{+/1})}{a+c} + \frac{f_{+/0}(1-f_{+/0})}{b+d} \tag{5.10}$$

$$= f_+(1-f_+)(\frac{1}{a+c} + \frac{1}{b+d}) \tag{5.11}$$

Où $f_+ = \frac{a+b}{n}$ est l'estimation commune d'être positif sous l'hypothèse nulle.

"Class vs. Purpose regroupé

Nous reprenons notre exemple ci-dessus. La fréquence commune d'être positif est fournie par

$$f_+ = \frac{180+120}{1000} = 0.3$$

Nous en déduisons la variance de DT :

$$V_0(DT) = 0.3 \times (1-0.3) \times (\frac{1}{180+359} + \frac{1}{120+341}) = 0.000845$$

La statistique de test est égale à :

$$z_{DT} = \frac{|\frac{180}{539} - \frac{120}{461}|}{\sqrt{0.000845}} = 2.5334$$

Pour un test bilatéral, la p-value est égale à 0.0113. Les données contredisent l'hypothèse d'égalité des proportions au risque 5%.

Intervalle de confiance de différence de proportions

Dans le cas du rejet de l'hypothèse d'égalité des proportions, il peut être utile d'estimer l'intervalle de confiance de la différence des deux proportions.

Par la rapport à la section précédente, l'estimation de la variance de DT est modifiée puisque l'idée d'une proportion commune aux deux populations (témoin et exposé) n'est plus valable. Elle s'écrit :

$$V(DT) = \frac{f_{+/1}(1-f_{+/1})}{a+c} + \frac{f_{+/0}(1-f_{+/0})}{b+d} \tag{5.12}$$

$$= \frac{a \times c}{(a+c)^3} + \frac{b \times d}{(b+d)^3} \tag{5.13}$$

"Class vs. Purpose regroupé

Pour notre exemple, voici la variance de la différence de proportions :

$$V(DT) = \frac{180 \times 359}{(180+359)^3} + \frac{120 \times 341}{(120+341)^3} = 0.000830$$

Et l'intervalle de confiance au niveau 95% de la différence

$$[0.0736 - u_{0.975} \times \sqrt{0.000830} \ ; \ 0.0736 + u_{0.975} \times \sqrt{0.000830}]$$
$$[0.0172 \ ; \ 0.1301]$$

5.4 Odds

Le Odds ou *rapport de chances* est défini pour chaque groupe, il est égal, pour le groupe exposé, à

$$odds(+/1) = \frac{P(+/1)}{P(-/1)} \tag{5.14}$$

où $P(+/1)$ est la probabilité d'être positif sachant qu'on est dans le groupe exposé, $P(-/1)$ celle d'être négatif dans le groupe exposé.

Sur un tableau de données (Tableau 5.2), on l'estime avec

$$\widehat{odds(+/1)} = \frac{\frac{a}{a+c}}{\frac{c}{a+c}} = \frac{a}{c} \tag{5.15}$$

Dans l'exemple "CLASS vs. PURPOSE" (Figure 5.2), il est égal à $\frac{180}{359} = 0.5014$. Cela veut dire que dans le groupe exposé, "il y a 0.5014 fois plus de chances d'être défaillant (que de ne pas l'être)" ; ou encore, "les défaillants sont ($\frac{1}{0.5104} = 1.99$) fois moins nombreux que les non-défaillants.

De la même manière, nous pouvons calculer le rapport de chances pour le groupe témoin

$$odds(+/0) = \frac{P(+/0)}{P(-/0)} \tag{5.16}$$

Il est égal à 0.3519 dans notre fichier exemple.

Les "odds" n'ont pas vraiment d'intérêt en tant que tel. Ils permettent en revanche d'introduire un nouvel indicateur, autrement plus performant.

5.5 Odds ratio

5.5.1 Définition et estimation

L'odds-ratio est le rapport entre l'odds du groupe exposé et l'odds du groupe témoin[5]

$$OR(1/0) = \psi_{1/0} = \frac{odds(+/1)}{odds(+/0)} = \frac{\frac{P(+/1)}{P(-/1)}}{\frac{P(+/0)}{P(-/0)}} \tag{5.17}$$

Il mesure la même chose que le risque relatif. Mais il présente d'excellentes propriétés qui font que son utilisation est plus avantageuse dans de nombreuses situations. Il est entre autres directement

5. Dans le texte, tant qu'il ne peut pas y avoir d'ambiguïté, nous omettrons le double indice de $\psi_{1/0}$. Nous les retrouverons lorsque nous aurons à comparer plusieurs groupes.

calculé par des techniques statistiques très populaires telles que la régression logistique ([10] ; [2], page 51) ou les modèles log-linéaires.

Sur un tableau de données, il est estimé par

$$\hat{\psi} = \frac{a/c}{b/d} = \frac{a \times d}{b \times c} \tag{5.18}$$

Dans l'exemple "Class vs. Purpose", nous obtenons $\hat{\psi} = \frac{180 \times 341}{120 \times 359} = 1.4248$. Les crédits ont 1.4248 fois plus de chances d'être défaillants dans le groupe exposé que dans le groupe témoin.

5.5.2 Équivalence avec le risque relatif

Il est toujours un peu gênant d'avoir 2 indicateurs censés dire la même chose et qui prennent des valeurs différentes. En réalité, dans de nombreuses circonstances, l'odds-ratio est très proche du risque relatif. Explicitons cela.

Dans les études réelles, la modalité positive de la variable Y est très peu fréquente. Fort heureusement, les malades sont rares ; les fraudeurs ne sont pas légions ; les crédits qui défaillent surviennent peu souvent. Dans notre tableau de contingence (Tableau 5.2), $(a+b)$ est très petit par rapport à $(c+d)$, a par rapport à c (par conséquent $a + c \approx c$) et b face à d ($b + d \approx d$). Si nous reprenons l'expression du risque relatif, nous remarquerons alors

$$rr = \frac{a \times (b+d)}{b \times (a+c)} \approx \frac{a \times d}{b \times c} \tag{5.19}$$

On retrouve l'expression de l'odds-ratio estimé sur le tableau de données (Equation 5.18).

Dans notre exemple "Class vs. Purpose", les estimations différent sensiblement, $rr = 1.2829$ contre $\hat{\psi} = 1.4248$, car la modalité positive est relativement fréquente ($f_+ = 30\%$).

Dans un second exemple (fictif) où $f_+ = 0.014$ est faible (Tableau 5.3), $rr = 1.2308$ et $\hat{\psi} = 1.2344$ sont très proches.

Tableau	Exposé (1)	Témoin (0)	Total
Positif (+)	16	12	28
Négatif (-)	1024	948	1972
Total	1040	960	2000

Tableau 5.3. Exemple de table pour calcul du rr et $\hat{\psi}_{1/0}$ - Petite valeur de f_+

5.5.3 Intervalle de confiance et test d'hypothèses

Intervalle de confiance

Le logarithme de l'estimateur de l'odds-ratio $\ln(\hat{\psi})$ suit asymptotiquement une loi normale de paramètres

$$E[\ln(\hat{\psi})] = \ln(\psi) \tag{5.20}$$

$$V[\ln(\hat{\psi})] = \sigma^2_{\ln(\hat{\psi})} = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \tag{5.21}$$

Nous pouvons calculer l'intervalle de confiance au niveau $(1-\alpha)$. En suivant le même cheminement que pour le risque relatif, il s'écrit

$$\ln(\psi) \in [\ln(\hat{\psi}) \pm u_{1-\frac{\alpha}{2}} \times \sigma_{\ln(\hat{\psi})}] \tag{5.22}$$

Nous en déduisons l'intervalle de confiance pour ψ.

Reprenons notre exemple "Class vs. Purpose" (Figure 5.2) et détaillons les calculs :

Indicateur	Valeur
$\hat{\psi}$	1.4248
$\ln(\hat{\psi})$	0.3540
$\sigma_{\ln(\hat{\psi})}$	$\sqrt{\frac{1}{180} + \frac{1}{120} + \frac{1}{359} + \frac{1}{341}} = \sqrt{0.0196} = 0.1400$
$u_{0.95}$	1.6449
Borne basse ($\ln(\psi)$)	0.1237
Borne haute ($\ln(\psi)$)	0.5843

L'intervalle de confiance au niveau 90% est

$$\psi \in [1.1317; 1.7938]$$

Test de significativité

A l'instar du risque relatif, nous avons la possibilité de tester la réalité du lien entre le facteur d'exposition X et la modalité positive de la variable d'intérêt Y. Il s'agit de comparer l'odds-ratio avec la valeur de référence 1 ou, c'est équivalent, comparer $\ln(\psi)$ à 0.

$$H_0 : \ln(\psi) = 0$$
$$H_1 : \ln(\psi) \neq 0$$

On utilise pour cela la statistique standardisée

$$z_{\ln(\hat{\psi})} = \frac{\ln(\hat{\psi})}{\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}} \tag{5.23}$$

Et la région critique du test au risque α est

$$R.C. : |z_{\ln(\hat{\psi})}| > u_{1-\frac{\alpha}{2}}$$

Dans l'exemple "Class vs. Purpose", nous obtenons $z_{\ln(\hat{\psi})} = \frac{0.3540}{0.1400} = 2.5283$; la probabilité critique du test est $p_c = 0.0115$. La conclusion est tout à fait cohérente avec celle du test basé sur le risque relatif : les nouveaux crédits semblent plus risqués, les défaillances sont proportionnellement (dans le groupe exposé relativement au groupe témoin) plus nombreuses.

Remarque 20 (Comparaison à un standard - Comparaison entre 2 *odds-ratio).* Le test de significativité est finalement un cas particulier de la comparaison à un standard. Il est possible de comparer l'odds-ratio à n'importe quelle valeur, il est également possible de procéder à un test unilatéral. Plus intéressant encore, nous pouvons envisager la comparaison entre 2 odds-ratios calculés sur des échantillons différents.

Remarque 21 (Odds-ratio et petits effectifs). Pour les petits effectifs ou lorsqu'une des cases du tableau contient la valeur 0, une formule alternative de l'odds-ratio peut être proposée ([1], section 3.4.1).

$$\hat{\psi} = \frac{(a+0.5) \times (d+0.5)}{(b+0.5) \times (c+0.5)} \tag{5.24}$$

De même, la variance du logarithme de l'estimateur peut être améliorée avec ([1], équation 3.15) :

$$V[\ln(\hat{\psi})] = \frac{1}{(a+0.5)} + \frac{1}{(b+0.5)} + \frac{1}{(c+0.5)} + \frac{1}{(d+0.5)} \tag{5.25}$$

Mode de tirage et odds-ratio

Par rapport au risque relatif, l'odds-ratio se justifie par certaines de ses propriétés disions-nous. Dans cette section, nous en explicitons une qui est tout à fait remarquable.

Bien souvent, la modalité positive est rare, voire très rare. Dans l'analyse des fraudes à la carte bancaire, les cas litigieux sont infimes par rapport au nombre de transactions normales. Effectuer un tirage aléatoire d'un échantillon conséquent ne garantit même pas la présence d'au moins une observation appartenant à la modalité positive de la variable d'intérêt. Pour éviter cet écueil, on procède à un tirage dit *rétrospectif* (Celeux, 1994 ; page 5), il consiste à définir a priori le nombre d'observations positives et négative désiré, puis à effectuer un tirage aléatoire dans ces sous-ensembles pour constituer l'échantillon. Dans ce cas, il est évident que la fréquence observée des positifs $f_+ = \frac{a+b}{c+d}$ n'est en rien une estimation de la probabilité $P(+)$.

Dans ce contexte, l'odds-ratio présente une propriété remarquable : il est invariant par rapport au mode d'échantillonnage. On peut le calculer dans le cadre d'un tirage rétrospectif, nous obtiendrons la même valeur si nous l'avions calculé sur un échantillon issu d'un tirage aléatoire simple.

Exemple "Class vs. Purpose" - Tirage équilibré

Reprenons notre exemple, nous avons conservé les 300 observations positives, nous voulons tirer aléatoirement 300 observations négatives parmi les 700 disponibles. Pour les besoins pédagogiques, nous

Tableau de contingence (Tirage retrospectif) (G)

	Groupe(Exposé = 1)	Groupe(Ref=0)	Total
bad (+)	180	120	300
good (-)	153.86	146.14	300
Total	333.86	266.14	600

Fig. 5.3. "Class vs. Purpose" - Tableau avec tirage rétrospectif équilibré

avons recalculé les effectifs à l'aide d'une simple règle de trois afin que les indicateurs obtenus soient complètement comparables, d'où les valeurs avec des décimales (Figure 5.3).

Nous récapitulons les indicateurs obtenus (Tableau 5.4). C'est particulièrement édifiant, seul l'odds-ratio conserve sa valeur dans l'échantillon volontairement équilibré. Cette propriété fait en grande partie sa popularité.

Indicateur	Tirage aléatoire (Fig. 5.2)	Tirage équilibré (Fig. 5.3)
Risque relatif	1.2829	1.1958
Odds(+/1)	0.5014	1.1699
Odds(+/0)	0.3519	0.8211
Odds-ratio	1.4248	1.4248

Tableau 5.4. Quelques indicateurs sur le croisement "Class vs. Purpose" selon le mode d'échantillonnage

5.6 Récapitulatif pour le croisement "CLASS vs. PURPOSE regroupé"

Tous les calculs sur l'exemple "Class vs. Purpose regroupé" sont centralisés dans une feuille de calcul du classeur EXCEL qui accompagne ce support. Nous la décrivons rapidement ici pour que la lecture en soit facilitée (Figure 5.4).

La feuille de calcul est subdivisée en une série de sous-tableaux :

[A] Tableau de contingence initial.

[B] Tableau de contingence après regroupement en groupes "exposé" et "témoin".

[C] Tableau des effectifs théoriques sous l'hypothèse d'indépendance.

[D] Tableau de calcul de la statistique du χ^2, on aperçoit en dessous la p-value du test qui nous indique que les lignes et les colonnes ne sont pas indépendantes. Il y a de l'information à glaner dans ce croisement, encore faut-il pouvoir la caractériser correctement.

[E] Tableau pour les notations.

[F] Calcul des odds, odds-ratio et risque relatif pour le tableau **B**.

[G] Tableau de contingence après ré-équilibrage des effectifs pour les catégories positives et négatives.

[H] Calcul des odds, odds-ratio et risques relatifs sur le tableau **G**.

Tableau de contingence initial (A)

class	business	domestic appliance	education	new car	other	repairs	retraining	used car	furniture/equipment	radio/tv	Total
bad	34	4	22	89	5	8	1	17	58	62	300
good	63	8	28	145	7	14	8	86	123	218	700
Total	97	12	50	234	12	22	9	103	181	280	1000

Tableau de contingence après regroupement (B)

	Groupe(Exposé = 1)	Groupe(Témoin=0)	Total
bad (+)	180	120	300
good (-)	359	341	700
Total	539	461	1000

Tableau sous indépendance (C)

	Groupe(Exposé = 1)	Groupe(Témoin=0)	Total
bad (+)	161.7	138.3	300
good (-)	377.3	322.7	700
Total	539	461	1000

Notations des cases du tableau (E)

	Groupe(Exposé = 1)	Groupe(Témoin=0)	Total
bad (+)	a	b	a+b
good (-)	c	d	c+d
Total	a+c	b+d	n

Tableau de calcul du KHI-2 (D)

	Groupe(Exposé = 1)	Groupe(Témoin=0)	Total
bad (+)	2.07	2.42	300
good (-)	0.89	1.04	700
Total	539	461	1000

KHI-2	6.417904
DDL	1
p-value	0.011298

Calcul des odds et odds-ratio (F)

Odds(+/1)	0.5014
Odds(+/0)	0.3519
RR(+/-)	1.2829
Odds-Ratio(+/-)	1.4248

Ln(odds-ratio)	0.3540
sigma[ln(OR)]	0.1400

z	2.5283
p-value	0.0115

Tableau de contingence (Tirage retrospectif) (G)

	Groupe(Exposé = 1)	Groupe(Ref=0)	Total
bad (+)	180	120	300
good (-)	153.86	146.14	300
Total	333.86	266.14	600

Calcul des odds et odds-ratio (H)

Odds(+/1)	1.1699
Odds(+/0)	0.8211
RR(+/-)	1.1958
Odds-Ratio(+/-)	1.4248

Fig. 5.4. "Class vs. Purpose" - Feuille de calcul pour les odds, odds-ratio et risque relatif

5.7 Odds-ratio et coefficient $\mathcal{Q}$ de Yule

Le coefficient $\mathcal{Q}$ de Yule peut être déduit de l'odds-ratio. Cela peut paraître étonnant tant ils paraissent différents, le premier est fondé sur les comparaisons par paires, le second sur des rapports de probabilités. Mais d'un autre côté, puisqu'il s'agit avant tout d'indicateurs calculés sur 4 cases d'un tableau de contingence, il est normal finalement que nous puissions les recouper. Les possibilités de combinaison de ces cases sont réduites.

La formule de passage est la suivante :

$$\mathcal{Q} = \frac{\hat{\psi} - 1}{\hat{\psi} + 1} \tag{5.26}$$

Dans notre exemple "Class vs. Purpose regroupé", nous avons ainsi

$$\mathcal{Q} = \frac{1.4248 - 1}{1.4248 + 1} = 0.1752$$

Si nous utilisons la formule initiale (Équation 4.7), nous obtenons exactement la même valeur

$$\mathcal{Q} = \frac{180 \cdot 341 - 120 \cdot 359}{180 \cdot 341 + 120 \cdot 359} = 0.1752$$

5.8 Odds-ratio dans les tableaux $2 \times C$

L'odds-ratio est initialement calculé pour les tableaux 2×2. En réalité, nous pouvons le généraliser pour des tableaux $2 \times C$ en opposant 2 modalités de la variable X.

Pour préciser les idées, reprenons notre étude de crédits risqués "Class vs. Purpose". Les décideurs veulent identifier les nouveaux prêts qui posent le plus problème, en les comparant avec les crédits traditionnels témoins. Le tableau croisé est organisé de manière différente (Figure 5.5).

Tableau de contingence

class	Témoin	business	domestic appliance	education	new car	other	repairs	retraining	used car	Total
bad (+)	120	34	4	22	89	5	8	1	17	180
good (-)	341	63	8	28	145	7	14	8	86	359
Total	461	97	12	50	234	12	22	9	103	539

Fig. 5.5. "Class vs. Purpose" - Tableau $2 \times C$ pour le calcul des odds-ratio individuels

Supposons, sans nuire à la généralité du propos, que le groupe témoin est toujours dans la première colonne du tableau de contingence. L'odds-ratio opposant la modalité x_c avec le groupe témoin x_1 s'écrit de la manière suivante

$$\hat{\psi}_{x_c/x_1} = \frac{n_{+c}/n_{-c}}{n_{+1}/n_{-1}}$$

Ce qui revient à élaborer $(C-1)$ tableaux 2×2 à partir du tableau initial. Cela peut être particulièrement intéressant dans les études facteurs - maladies lorsque nous étudions différents degrés d'exposition (ex. le groupe témoin est constitué des non-fumeurs, le second groupe fume 1 paquet par jour, le troisième 2 paquets, etc.). L'idée au final est de pouvoir classer les groupes selon le risque qui leur est associé (Reynaud, 1986 ; page 75). Ce type d'analyse est d'autant plus important que le regroupement des modalités en "groupe témoin" vs. "groupe exposé" peut masquer les différences significatives dans certains sous-groupes.

Nous pouvons compléter l'analyse avec un test de significativité. Mais il faut être prudent car nous sommes en situation de *comparaisons multiples* : à force de tester différentes configurations, nous finissons par détecter à tort des odds-ratio significatifs. Le risque des tests individuels doit être corrigé. La technique la plus connue, mais peut-être pas la meilleure, est la *correction de Bonferroni*[6] : elle consiste à diviser le risque α par le nombre K de tests.

Nous avons complété la feuille de calcul dans notre exemple (Figure 5.6). En sus du tableau de contingence, nous observons tour à tour :

1. Les odds calculés pour chaque groupe.
2. Les odds-ratio entre chaque groupe et le groupe témoin placé dans la première colonne.
3. Le rang des odds-ratio pour détecter les groupes où le risque est le plus élevé. Nous constatons que les crédits destinés à l'éducation semblent les plus risqués (1-er), suivis des "autres crédits"

6. Voir http://en.wikipedia.org/wiki/Bonferroni_correction, avec les critiques qui vont avec.

Tableau de contingence

class	Témoin	business	domestic appliance	education	new car	other	repairs	retraining	used car	Total
bad (+)	120	34	4	22	89	5	8	1	17	180
good (-)	341	63	8	28	145	7	14	8	86	359
Total	461	97	12	50	234	12	22	9	103	539
Odds(+/-)	0.3519	0.5397	0.5000	0.7857	0.6138	0.7143	0.5714	0.1250	0.1977	

Odds-ratio

Odds-ratio(+/-; Groupe/Témoin)	1.5336	1.4208	2.2327	1.7442	2.0298	1.6238	0.3552	0.5617
Rang	5	6	1	3	2	4	8	7

Test de significativité

Ln(Odds-ratio)	0.4276	0.3512	0.8032	0.5563	0.7079	0.4848	-1.0351	-0.5767
Ecart type	0.2378	0.6215	0.3040	0.1715	0.5951	0.4557	1.0660	0.2859
z	1.7982	0.5652	2.6419	3.2445	1.1896	1.0637	-0.9710	-2.0176

Fig. 5.6. "Class vs. Purpose" - Feuille de calcul pour les odds-ratio individuels dans un tableau $2 \times C$

(2-ème), puis des crédits destinés à l'achat de véhicules neufs (3-ème). A l'opposé, les crédits pour les véhicules d'occasion (7-ème) ou les crédits pour formation de reconversion (8-ème) sont les plus sûrs.

4. Il faut détecter les odds-ratio significativement différents de 1. Cela dépend bien sûr de l'odds-ratio mais aussi de l'écart type (en filigrane, de l'effectif associés aux groupes si l'on se penche un instant sur la formule de l'écart type).
   — Nous calculons tout d'abord le logarithme de l'odds-ratio,
   — Puis les écart types associés.
   — Nous formons la statistique $z_{\ln(\hat{\psi})}$.
   — Nous effectuons $K = 8$ comparaisons. Pour un risque $\alpha = 10\%$, nous introduisons la correction de Bonferroni, le risque individuel est $\frac{\alpha}{K} = \frac{0.10}{8} = 0.0125$, la valeur critique $u_{1-0.0125} = 2.2414$. Tous les groupes pour lesquels $|z|$ est supérieur à ce seuil présentent un odds-ratio significativement différent de 1,
   — C.-à-d. les crédits pour voitures neuves et les crédits pour l'éducation[7].

7. Encore une fois, on ne s'étendra pas sur les bizarreries de ce fichier...

6

Coefficient de concordance pour les variables nominales

6.1 Concordance de jugements - Le coefficient κ

L'étude de la concordance de jugements consiste à évaluer la cohérence du classement de n objets (individus) en C catégories par plusieurs juges. Dans l'exemple du crédit bancaire, il y a 1000 dossiers à étudier. Admettons que chaque dossier est évalué de manière indépendante par 2 experts qui les classent dans 3 catégories distinctes : bon dossier, mauvais dossier, ne se prononce pas. Nous pouvons dès lors nous poser questions : Est-ce que, individuellement, le dossier i a été classé de la même manière par les 2 experts ? Est-ce que, globalement, les experts classent les dossiers de la même manière ?

Nous nous intéresserons à la seconde question. Contrairement aux autres chapitres de ce support, il ne s'agit donc pas d'étudier la relation existant entre 2 variables. Nous ne manipulons pas un tableau de contingence au sens où nous l'avions défini initialement (Section 1.2). Le tableau est défini par le croisement entre : en ligne, les individus à classer ; en colonne, les catégories à attribuer ; à l'intérieur du tableau, x_{ic} correspond au nombre de fois où la catégorie c a été assignée à l'individu i (Tableau 6.1).

Remarque importante, il s'agit d'une analyse sur échantillons appariés puisque qu'une observation est traitée par plusieurs juges.

Individu	1	$\cdots$	c	$\cdots$	C	Somme
1	x_{11}	$\cdots$	x_{1c}	$\cdots$	x_{1C}	$x_{1.}$
$\vdots$						$\vdots$
i	x_{i1}	$\cdots$	x_{ic}	$\cdots$	x_{iC}	$x_{i.}$
$\vdots$						$\vdots$
n	x_{n1}	$\cdots$	x_{nc}	$\cdots$	x_{nC}	$x_{n.}$
Somme	$x_{.1}$	$\cdots$	$x_{.c}$	$\cdots$	$x_{.C}$	-

Tableau 6.1. Tableau de données pour le calcul du coefficient de concordance

La marge ligne $x_{i.}$ correspond au nombre de juges qui a évalué l'individu numéro i. La marge colonne $x_{.c}$ correspond au nombre de fois où la catégorie c a été attribuée à un individu.

Le coefficient de concordance pour données nominales, que nous désignerons sous l'appellation générique "coefficient κ", peut être rapproché avec coefficient de concordance de Kendall[1]. A la différence qu'il ne s'agit pas de classer les individus (c.-à-d. de leur affecter une valeur qui permet de les trier), mais plutôt de les affecter à des catégories, qui peuvent par ailleurs correspondre à un ordonnancement. Il assouplit la contrainte du coefficient de Kendall. Dans notre exemple des crédits bancaires, il paraît inconcevable de demander à un expert de classer 1000 crédits selon un ordre de préférence. En revanche leur demander d'attribuer aux dossiers une étiquette selon leur intérêt à chaque expertise (bon, mauvais, ne se prononce pas) ne devrait pas poser de problèmes.

Les applications du coefficient de concordance sont multiples :

— Vérifier que plusieurs tests médicaux aboutissent à des diagnostics identiques.
— Vérifier que les médecins conseillent les mêmes traitements aux patients.
— Évaluer la cohérence des conseillers d'orientation pour les élèves entrant au lycée.
— Mesurer la diversité des modèles dans les méthodes ensemblistes en apprentissage supervisé.
— Etc.

On peut voir de différentes manières un accord fort entre les appréciations. Si les experts classent de manière identique les objets, cela conforte le jugement, cela veut dire qu'ils perçoivent la même information sous-jacente lorsqu'ils affectent un objet à une catégorie. Cela indiquerait aussi que, finalement, une seule expertise suffirait puisque, de toute manière, ils préconisent le même classement.

Deux situations extrêmes sont opposés pour définir l'indicateur : (1) pour chaque observation, les juges attribuent la même catégorie, il y a un accord parfait entre les jugements ; (2) pour chaque observation, les juges attribuent au hasard la catégorie. Le coefficient κ indique l'intensité de l'accord, il correspond au rapport

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{6.1}$$

où $P(A)$ est la probabilité que les juges s'accordent, $P(E)$ la probabilité que les juges s'accordent s'ils classaient les individus totalement au hasard.

Sur un échantillon de n observations, les probabilités sont estimés par des proportions, nous obtenons la statistique $\hat{\kappa}$

$$\hat{\kappa} = \frac{p_a - p_e}{1 - p_e} \tag{6.2}$$

où p_a est la proportion d'accords entre juges dans l'échantillon observé, p_e la proportion d'accords que l'on observerait si les juges classaient les n objets au hasard.

Le coefficient est défini sur l'intervalle -1 (désaccord total entre les juges) et 1 (accord parfait entre les juges). En situation d'indépendance entre les jugements c.-à-d. les affectations se font au hasard, il sera égal à 0.

1. Voir http://www.cons-dev.org/elearning/stat/stat7/st7.html

κ	Interprétation
< 0.00	Désaccord
$0.00 - 0.2$	Très faible accord
$0.21 - 0.40$	Faible accord
$0.41 - 0.60$	Accord modéré
$0.61 - 0.80$	Accord fort
> 0.80	Accord excellent

Tableau 6.2. Grille de lecture des valeurs du coefficient κ

Certains auteurs[2] proposent une grille de lecture des valeurs du coefficient (Tableau 6.2). Comme toujours avec ce type de référence, il faut surtout s'y rapporter à titre indicatif. Plus importantes sont sans nul doute les indications fournies par la statistique inférentielle (intervalle de confiance et test d'hypothèses).

Fidèle à notre démarche, nous adoptons une présentation assez succincte et en relation constante avec un exemple dans ce support. Si le lecteur souhaite une étude approfondie du coefficient κ, nous conseillons le site "http://kappa.chez-alice.fr/" qui est remarquable. Il fait le tour quasi-exhaustif du coefficient de concordance. Un logiciel est de plus proposé pour effectuer des calculs sur un jeu de données. Ce document nous a beaucoup inspiré pour ce chapitre.

Exemple "CREDIT GERMAN"

Nous apprenons que tous les dossiers de demande de crédit ont été, au fur et à mesure, évalués par plusieurs experts indépendants qui leur attribuaient l'étiquette "bon", "mauvais", "indéterminé". La direction, qui avaient conservé ces données sous le coude[3] pendant longtemps, veut savoir si ces expertises sont cohérentes c.-à-d. s'il y a un consensus dans ces jugements. Elle extrait les nouvelles colonnes de l'entrepôt de données et nous les soumettent. Ainsi, dans notre fichier initial viennent s'ajouter des variables catégorielles correspondant au classement des experts. Chaque variable supplémentaire peut prendre 3 modalités possibles, celles énumérées ci-dessus.

6.2 Cas de 2 juges - κ de Cohen

6.2.1 Confrontation de 2 juges : croisement de 2 variables

Définition de la statistique

On nous demande de confronter l'étiquetage de 2 cabinets d'expertise très célèbres. Le cas particulier de 2 juges nous intéresse car nous pouvons former une autre représentation des données, un tableau de

2. http://kappa.chez-alice.fr/Kappa_2juges_Def.htm et http://en.wikipedia.org/wiki/Fleiss'_kappa

3. Qu'est-ce qu'il ne faut pas inventer pour intéresser le lecteur, je vous jure...

contingence qui croise : en ligne, le classement effectué par le premier juge, en colonne le classement effectué par le second juge (Tableau 6.3). Avant même les calculs, un coup d'oeil très rapide permet de constater qu'aucun individu classé en "bad" par le premier juge n'a été classé en "good" par le second, et inversement. Cela rassure la banque quant au sérieux de ces deux fameux experts. Il reste néanmoins qu'il y a des valeurs en dehors de la diagonale principale. L'accord n'est pas parfait, il nous faut quantifier son intensité.

Juge 1 x Juge 2	bad	good	indet.	Somme
bad	188	0	4	192
good	0	417	6	423
indet.	59	67	259	385
Somme	247	484	269	1000

Tableau 6.3. Croisement de 2 juges sur le fichier GERMAN CREDIT

Nous retrouvons une configuration qui correspond à un tableau de contingence. Dans la situation d'indépendance, le contenu du tableau est bien connu, il est égal au produit des marges (qui devrait être divisé par la marge totale, nous simplifions un peu ici). La statistique κ, dit κ de Cohen dans le cas de 2 juges [4], compare les diagonales principales du tableau observé et du tableau théorique. En respectant les notations associées au tableau de contingence (Tableau 1.1), elle s'écrit :

$$\begin{aligned}\hat{\kappa}_c &= \frac{p_a - p_e}{1 - p_e} \\ &= \frac{\frac{1}{n}\sum_l n_{ll} - \frac{1}{n^2}\sum_l n_{l.}n_{.l}}{1 - \frac{1}{n^2}\sum_l n_{l.}n_{.l}} \\ &= \frac{n\sum_l n_{ll} - \sum_l n_{l.}n_{.l}}{n^2 - \sum_l n_{l.}n_{.l}}\end{aligned}$$

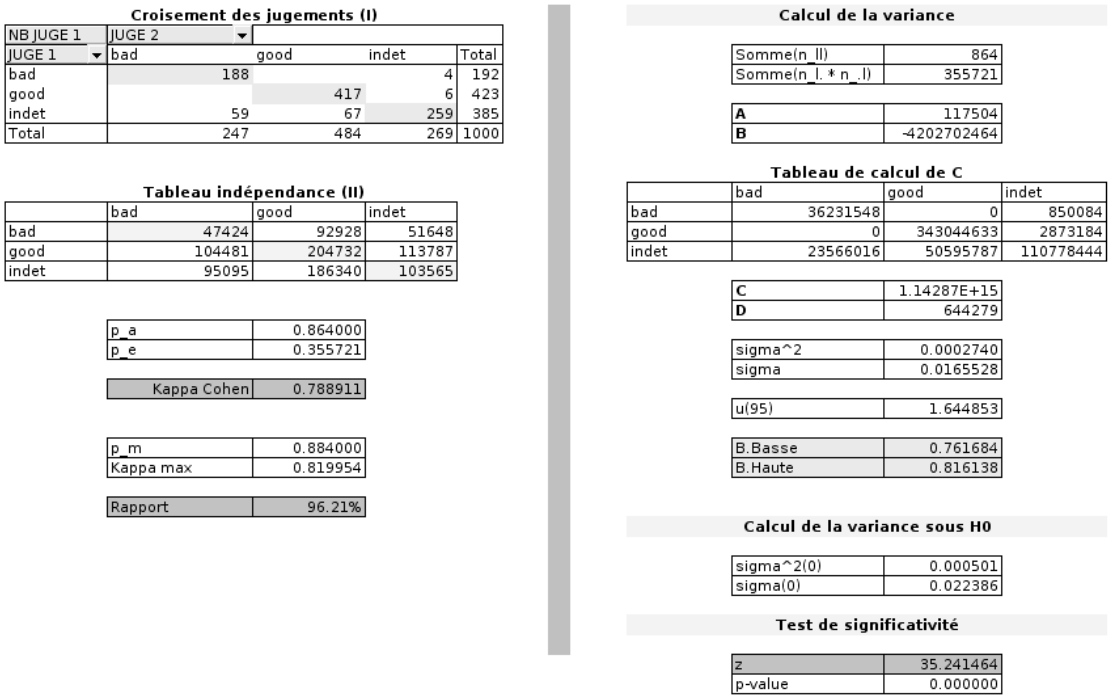
Exemple "CREDIT GERMAN"

Dans notre exemple, nous reproduisons la séquence de calcul (Figure 6.1), nous nous intéressons uniquement à la partie gauche à ce stade :

1. Le tableau **(I)** croise les deux variables.
2. A partir du produit des marges, nous formons le tableau **(II)** sous indépendance.
3. Nous calculons la proportion des accords $p_a = \frac{188+417+259}{1000} = 0.864000$.
4. De la même manière, la proportion des accords sous l'hypothèse d'indépendance $p_e = \frac{47424+204732+103565}{1000^2} = 0.355721$
5. La statistique s'obtient par le rapport $\hat{\kappa}_c = \frac{0.864-0.355721}{1-0.355721} = 0.788911$

Si l'on se réfère à notre grille de lecture (Tableau 6.2), nous avons un accord fort entre les 2 juges. Ce que laissait déjà entrevoir la lecture rapide du tableau croisant les affectations.

4. http://kappa.chez-alice.fr/Kappa_2juges_Def.htm



Croisement des jugements (I)

NB JUGE 1	JUGE 2			
JUGE 1	bad	good	indet	Total
bad	188		4	192
good		417	6	423
indet	59	67	259	385
Total	247	484	269	1000

Tableau indépendance (II)

	bad	good	indet
bad	47424	92928	51648
good	104481	204732	113787
indet	95095	186340	103565

p_a	0.864000
p_e	0.355721
Kappa Cohen	0.788911
p_m	0.884000
Kappa max	0.819954
Rapport	96.21%

Calcul de la variance

Somme(n_ll)	864
Somme(n_l. * n_.l)	355721
A	117504
B	-4202702464

Tableau de calcul de C

	bad	good	indet
bad	36231548	0	850084
good	0	343044633	2873184
indet	23566016	50595787	110778444

C	1.14287E+15
D	644279
sigma^2	0.0002740
sigma	0.0165528
u(95)	1.644853
B.Basse	0.761684
B.Haute	0.816138

Calcul de la variance sous H0

sigma^2(0)	0.000501
sigma(0)	0.022386

Test de significativité

z	35.241464
p-value	0.000000

Fig. 6.1. CREDIT GERMAN - Feuille de calcul pour la statistique $\hat{\kappa}_c$ - 2 juges

Coefficient corrigé

On compare le coefficient calculé par rapport à l'accord maximal pour évaluer son intensité. C'est d'ailleurs un peu l'idée de la grille de lecture ci-dessus. Or $\hat{\kappa}_c$ n'a la possibilité d'atteindre la valeur 1 que dans des circonstances très particulières, lorsque les marges en lignes et en colonne sont identiques c.-à-d. $n_{l.} = n_{.l}$. Dans les autres cas, sa valeur maximale est inférieure à 1. Une manière de corriger la statistique est donc de calculer l'accord maximal que l'on peut atteindre sur une configuration donnée, de calculer le coefficient $\hat{\kappa}_c$ correspondant et, de lui comparer le coefficient obtenu.

La proportion d'accord maximal que l'on peut obtenir pour un tableau donné est

$$p_m = \frac{1}{n} \sum_l \min(n_{l.}; n_{.l})$$

On en déduit alors le coefficient κ, il est égal à

$$\hat{\kappa}_c(max) = \frac{p_m - p_e}{1 - p_e} \tag{6.3}$$

Nous pouvons construire une statistique corrigée qui serait le rapport entre $\hat{\kappa}_c$ et $\hat{\kappa}_c(max)$.

Malheureusement, dans ce cas, nous perdons l'interprétation du coefficient en termes d'accords et de désaccords. Pour relativiser tout cela, nous dirons surtout que cet ajustement fait partie des stratagèmes destinés à corriger certaines carences du coefficient κ.

Exemple "CREDIT GERMAN"

Reprenons notre exemple, nous obtenons aisément $p_e = \frac{\min(192;247)+\min(423;484)+\min(385;269)}{1000} = \frac{192+423+269}{1000} = 0.884000$; le coefficient associé est $\hat{\kappa}_c(max) = 0.819954$.

Notre association correspond à $\frac{0.788911}{0.819954} = 96.21\%$ de l'accord maximal qu'elle pourrait atteindre.

6.2.2 Statistique inférentielle

Intervalle de confiance

Le coefficient $\hat{\kappa}_c$ suit asymptotiquement une loi normale. Pour calculer son intervalle de confiance, il nous faut produire une estimation de sa variance :

$$\sigma^2_{\hat{\kappa}_c} = n\left[\frac{A}{D^2} + \frac{B}{D^3} + \frac{C}{D^4}\right] \tag{6.4}$$

où, toujours en utilisant les notations du tableau de contingence,

— $A = (\sum_l n_{ll})(n - \sum_l n_{ll})$

— $B = 2(n - \sum_l n_{ll})\left[2\sum_l n_{ll}\sum_l n_{l.}n_{.l} - n\sum_l n_{ll}(n_{l.} + n_{.l})\right]$

— $C = (n - \sum_l n_{ll})^2\left[n\sum_{l,c} n_{lc}(n_{l.} + n_{.l})^2 - 4(\sum_l n_{l.}n_{.l})^2\right]$

— $D = n^2 - \sum_l n_{l.}n_{.l}$

Nous pouvons alors écrire l'intervalle de variation au niveau de confiance $(1-\alpha)$

$$\kappa_c \in \left[\hat{\kappa}_c \pm u_{1-\frac{\alpha}{2}} \times \sigma_{\hat{\kappa}_c}\right] \tag{6.5}$$

Remarque 22 (Autre écriture de la variance). La variance ci-dessus est fort complexe. Il existe une écriture alternative qui utilise les fréquences relatives. Elle n'est pas plus simple mais elle nous a permis au moins de recouper les calculs. Elle est disponible, entre autres, sur le site de documentation du logiciel SAS (http://v8doc.sas.com/sashtml/stat/chap28/sect26.htm). Nous obtenons exactement les mêmes résultats. Ce qui est plutôt rassurant.

Test de significativité

Nous utilisons toujours la normalité asymptotique pour mettre en place le test

$$H_0 : \kappa_c = 0$$
$$H_1 : \kappa_c \neq 0$$

Sauf que, bien entendu, il nous faut maintenant produire une estimation de la variance sous l'hypothèse nulle. Elle est assez simplifiée par rapport à la précédente :

$$\sigma^2_{\hat{\kappa}_c}(0) = \frac{1}{n\left(n^2 - \sum_l n_{l.}n_{.l}\right)^2}\left[n^2\sum_l n_{l.}n_{.l} + \left(\sum_l n_{l.}n_{.l}\right)^2 - n\left(\sum_l n_{l.}n_{.l}(n_{l.}+n_{.l})\right)\right] \tag{6.6}$$

Nous formons la statistique

$$z_{\hat{\kappa}_c} = \frac{\hat{\kappa}_c}{\sigma_{\hat{\kappa}_c}(0)} \tag{6.7}$$

Et, pour notre test bilatéral au risque α, la région critique du test s'écrit pour

$$R.C. : |z_{\hat{\kappa}_c}| > u_{1-\frac{\alpha}{2}}$$

La p-value du test nous permet aussi de décider.

Exemple des données CREDIT-GERMAN

Pour expliciter toutes ces formules fastidieuses sur un exemple, reprenons la copie d'écran de la feuille EXCEL (Figure 6.1). Dans la partie de droite, nous retrouvons principalement les éléments de calculs de l'intervalle de confiance et du test de significativité.

Intervalle de confiance

Le premier enjeu est le calcul de la variance asymptotique. Nous formons tour à tour :

— Les sommes intermédiaires $\sum_l n_{ll} = 864$ et $\sum_l n_{l.}n_{.l} = 355721$.

— Nous calculons alors $A = 117504$, $B = -4202702464$.

— Pour C, nous sommes obligés de créer un tableau intermédiaire, avant de pouvoir former $C = 1.14287 \cdot 10^{15}$.

— $D = 644279$.

— Ce sont les étapes les plus difficiles. La variance est obtenue simplement avec $\sigma^2_{\hat{\kappa}_c} = 1000 \times \left[\frac{117504}{644279^2} + \frac{(-4202702464)}{644279^3} + \frac{1.14287 \cdot 10^{15}}{644279^4}\right] = 0.0002740$, et l'écart type $\sigma_{\hat{\kappa}_c} = 0.0165528$.

Pour un niveau de confiance 90%, nous utilisons le quantile $u_{0.95} = 1.644853$, nous obtenons l'intervalle de confiance

$$\kappa_c \in [0.788911 - 1.644853 \times 0.0165528 \ ; 0.788911 + 1.644853 \times 0.0165528]$$

$$\kappa_c \in [0.761684 \ ; 0.816138]$$

Test de significativité

Dans ce cas, la formule de la variance est simplifiée, nous obtenons l'écart type $\sigma_{\hat{\kappa}_c}(0) = 0.022386$, la statistique $z_{\hat{\kappa}_c} = \frac{0.788911}{0.022386} = 35.24$.

La p-value du test est $p_c < 0.000001$. Sans aucun doute, le coefficient est très significativement différent de 0.

6.3 Cas de m juges - κ de Fleiss

6.3.1 Confrontation de m juges - Nouveau tableau de données

Dans le cas où l'on veut évaluer la cohérence de m juges, il n'est plus question d'utiliser un tableau de contingence croisant les affectations. Nous devons donc revenir au tableau présenté initialement (Tableau 6.1).

Reprenons la confrontation de 2 juges dans notre exemple "CREDIT GERMAN", nous affichons le tableau de données pour les 20 premiers individus (Figure 6.2). Pour l'observation numéro 1, 2 affectations ont été réalisées, elles sont relatives à la catégorie "good". L'observation numéro 2 a été jugée 2 fois "bad"... L'individu numéro 13 a été jugée une fois "good", une fois "indéterminée", etc. Dans cette configuration, la marge colonne est constante, elle est égale au nombre de juges ($m = x_{i.}$).

Tableau de calcul

Numéro	bad	good	indet	Total
1	0	2	0	2
2	2	0	0	2
3	0	2	0	2
4	0	0	2	2
5	2	0	0	2
6	0	0	2	2
7	0	2	0	2
8	0	0	2	2
9	0	2	0	2
10	2	0	0	2
11	2	0	0	2
12	2	0	0	2
13	0	1	1	2
14	0	0	2	2
15	2	0	0	2
16	1	0	1	2
17	0	2	0	2
18	2	0	0	2
19	1	0	1	2
20	0	2	0	2

Fig. 6.2. CREDIT GERMAN, 20 premières observations - Organisation des données pour le calcul de $\hat{\kappa}_f$

La généralisation à m juges est le coefficient κ_f de Fleiss[5]. Il résulte toujours de la confrontation entre la proportion de concordance observée et la proportion que l'on observerait si les juges affectait les catégories au hasard. Reste à expliciter la démarche et les formules de dénombrement (Siegel, 1988 ; pages 284 à 291).

Notons f_c la proportion d'individus affectés à la catégorie c, c'est le rapport $f_c = \frac{x_{.c}}{m \times n}$. En effet, chaque individu a été affecté m fois, nous travaillons sur un échantillon apparié. Si les juges affectent les catégories au hasard, la proportion totale d'accord pour la catégorie c est f_c^2, sur l'ensemble des catégories nous obtenons

$$p_e = \sum_c f_c^2 \tag{6.8}$$

Voyons maintenant la configuration observée. Pour un individu i, la proportion d'accord entre les m juges est le ratio entre le nombre de paires pour lesquelles cet accord a lieu et le nombre de paires que l'on peut former

5. http://en.wikipedia.org/wiki/Fleiss'_kappa

$$\begin{aligned} s_i &= \frac{\sum_c C_{x_{ic}}^2}{C_m^2} \\ &= \frac{1}{m(m-1)} \sum_c x_{ic}(x_{ic}-1) \end{aligned}$$

La proportion totale d'accord est obtenue sur l'ensemble des observations

$$\begin{aligned} p_a &= \frac{1}{n} \sum_i s_i \\ &= \left[\frac{1}{nm(m-1)} \sum_i \sum_c x_{ic}^2 \right] - \frac{1}{m-1} \end{aligned}$$

La statistique de Fleiss est toujours un coefficient de concordance

$$\hat{\kappa}_f = \frac{p_a - p_e}{1 - p_e} \tag{6.9}$$

Remarque 23 (Écriture alternative de la formule). Il existe une écriture alternative de la formule de Fleiss que l'on retrouve dans la littérature (http://kappa.chez-alice.fr/Kappa_plusjuges_cstjug.htm). Nous la donnons à titre indicatif, le plus important est que le résultat calculé soit exactement le même dans les 2 cas. Ce que nous avons vérifié bien entendu.

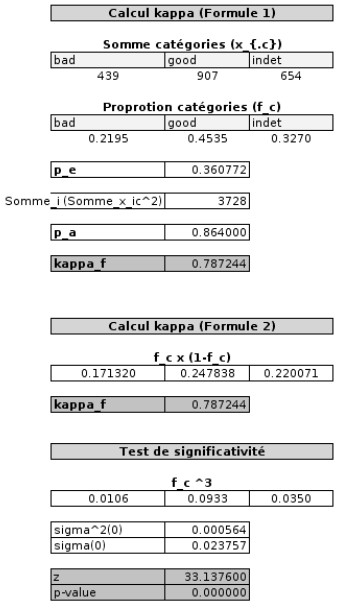
$$\hat{\kappa}_f = 1 - \frac{nm^2 - \sum_i \sum_c x_{ic}^2}{nm(m-1) \sum_c f_c(1-f_c)} \tag{6.10}$$

Exemple des données CREDIT-GERMAN

Nous reprenons le même exemple que tout à l'heure, les dossiers ont été classés par 2 juges. L'intérêt est de pouvoir comparer les résultats du κ de Cohen et du κ de Fleiss. La différence bien entendu est que la deuxième technique peut être mise en oeuvre pour un nombre quelconque de juges ($m > 2$).

Les étapes de calcul sont détaillés dans la feuille EXCEL (Figure 6.3 ; **Formule 1**) qui partent du nouveau format du tableau de données (Tableau 6.1) et dont un extrait est affiché dans la figure 6.2 :

1. Première étape, nous calculons les sommes des affectations à chaque catégorie ("bad", "good", "indet") $x_{.c} = \{439, 907, 654\}$.
2. Nous en déduisons les proportions $f_{.c} = \{0.2195, 0.4535, 0.3270\}$.
3. Nous pouvons calculer la proportion d'accord sous l'hypothèse d'affectation au hasard $p_e = 0.2195^2 + 0.4535^2 + 0.3270^2 = 0.360772$
4. Dans un deuxième temps, nous cherchons à calculer p_a. Pour ce faire, nous effectuons la somme $\sum_i \sum_c x_{ic}^2 = 3278$.
5. La proportion d'accord est $p_a = \frac{1}{1000 \cdot 2 \cdot (2-1)} 3278 - \frac{1}{2-1} = 0.864000$.
6. L'estimation du coefficient κ_f de Fleiss est $\hat{\kappa}_f = 0.787244$.



Calcul kappa (Formule 1)		
Somme catégories (x_{.c})		
bad	good	indet
439	907	654
Proprotion catégories (f_c)		
bad	good	indet
0.2195	0.4535	0.3270
p_e	0.360772	
Somme_i (Somme_x_ic^2)	3728	
p_a	0.864000	
kappa_f	0.787244	

Calcul kappa (Formule 2)		
f_c x (1-f_c)		
0.171320	0.247838	0.220071
kappa_f	0.787244	

Test de significativité		
f_c ^3		
0.0106	0.0933	0.0350
sigma^2(0)	0.000564	
sigma(0)	0.023757	
z	33.137600	
p-value	0.000000	

Fig. 6.3. CREDIT GERMAN - Calcul de $\hat{\kappa}_f$ de Fleiss et test de significativité

7. En utilisant la formulation alternative (Équation 6.10), nous obtenons exactement la même valeur (**Formule 2** dans la feuille de calcul).

Surprise, la formule de Cohen et de Fleiss n'aboutissent pas au même résultat. La principale raison est qu'elles dénombrent différemment la proportion d'accord p_e sous l'hypothèse d'indépendance. Il reste toutefois que les coefficients estimés sont assez proches. Ils ne remettent pas en cause leurs conclusions respectives : les 2 experts financiers sont très cohérents.

6.3.2 Statistique inférentielle - Test de significativité

La formule de la variance de $\hat{\kappa}_f$ valable quelle que soit l'hypothèse nulle n'est pas disponible, il n'est donc pas possible de calculer des intervalles de confiance[6].

Sous l'hypothèse d'affectation aléatoire des individus, il est néanmoins possible d'en fournir une estimation assez approximative (Siegel, 1988 ; page 289, formule 9.30). Nous avons la possibilité de mettre en place un test de significativité pour opposer $H_0 : \kappa_f = 0$ vs. $H_1 : \kappa_f \neq 0$.

$$\sigma^2_{\hat{\kappa}_f}(0) \approx \frac{2}{nm(m-1)} \frac{p_e - (2m-3)p_e^2 + 2(m-2)\sum_c f_c^3}{(1-p_e)^2} \tag{6.11}$$

En utilisant l'approximation normale, la statistique $z_{\hat{\kappa}_f}$ suit une loi normale centrée réduite et la région critique du test s'écrit

6. Il semble qu'il faille passer par des formulations différentes du coefficient pour pouvoir produire une estimation viable de la variance. Voir http://kappa.chez-alice.fr/kappa_plusjuges_sign.htm, formules de Landis et Koch (Landis J.R., Koch G.G., *A one-way components of variance model for categorical data* , Biometrics, 1977, 33, 671-679.)

$$R.C. : |z_{\hat{\kappa}_f}| > u_{1-\frac{\alpha}{2}}$$

Nous pouvons également utiliser la p-value.

Exemple des données CREDIT-GERMAN

Considérons la dernière partie de la figure 6.3 :

1. Nous calculons tout d'abord les f_c^3 (bien qu'en réalité nous n'en aurons pas besoin dans le cas précis puisque nous avons $m = 2$ juges et donc $(m-2) = 0$, cette partie de la formule est égale à 0.
2. Nous calculons la variance sous l'hypothèse nulle $\sigma^2_{\hat{\kappa}_f}(0) = 0.000564$, d'où l'écart type $\sigma_{\hat{\kappa}_f}(0) = 0.023757$.
3. La statistique z devient $z_{\hat{\kappa}_f} = 33.137600$,
4. Et la p-value du test $p_c < 0.000001$.

Les experts sont très cohérents lorsqu'ils classent les dossiers. Nous remarquerons que, sans être les mêmes, les valeurs sont très similaires à ceux du κ_c de Cohen. Les conclusions sont très tranchées.

6.4 Nombre de juges quelconque - Formule généralisée

Très souvent, en situation réelle, il est illusoire de pouvoir disposer constamment des mêmes juges pour classer les individus. Les experts d'un organisme de crédit, par exemple, sont des gens précieux, très occupés, ils changent au cours du temps, ils peuvent être en nombre différent. Il nous faudrait une formulation du coefficient de concordance κ qui soit valable lorsque le nombre de juges $m_i = x_{i.}$ est différent d'un individu à l'autre.

Au delà de son intérêt pratique, le coefficient que nous présentons dans cette section revêt une importance particulière car il introduit une notion nouvelle : la concordance associée aux catégories[7]. Le coefficient κ global (de Landis et Koch) est la moyenne pondérée de ces concordances.

6.4.1 Concordance associées à la catégorie c

Le nombre de juges n'étant plus constant, il nous faut redéfinir dans un premier temps la proportion d'individus affectés à la catégorie c. Elle est égale maintenant à

$$f_c = \frac{\sum_i x_{ic}}{\bar{m}} \tag{6.12}$$

où $\bar{m} = \frac{\sum_i m_i}{n}$ est le nombre moyen de juges par observation.

Le coefficient de concordance associée à la catégorie c s'écrit

$$\hat{\kappa}(c) = 1 - \frac{\sum_i \frac{x_{ic}(m_i - x_{ic})}{m_i}}{n(\bar{m}-1)f_c(1-f_c)} \tag{6.13}$$

7. Voir http://kappa.chez-alice.fr/Kappa_plusjuges_dmod.htm

6.4.2 Concordance globale

On dérive le coefficient global de concordance entre les juges en effectuant une moyenne pondérée des κ associées aux catégories [8]

$$\hat{\kappa}_g = \frac{\sum_c f_c(1-f_c)\hat{\kappa}(c)}{\sum_c f_c(1-f_c)} \tag{6.14}$$

Remarque qui a son importance, si le nombre de juges est finalement constant $m_i = m$, cette formule permet de retomber sur le coefficient de Fleiss. Il s'agit bien d'une généralisation.

6.4.3 Application aux données CREDIT-GERMAN

Reprenons toujours le même exemple. Le nombre de juges est constant mais nous ne sommes pas censés le savoir. Nous appliquons directement les formules de Landis et Koch. Nous vérifierons à la fin si nous retombons sur les mêmes résultats que précédemment (κ de Fleiss). Les calculs sont résumés dans la feuille EXCEL (Figure 6.4). Nous nous consacrons dans un premier temps aux coefficients associés aux catégories.

	Tableau de calcul				x_ic(m_i - x_ic)/m_i		
Numér	bad	good	indet	m_i	bad	good	indet
1	0	2	0	2	0	0	0
2	2	0	0	2	0	0	0
3	0	2	0	2	0	0	0
4	0	0	2	2	0	0	0
5	2	0	0	2	0	0	0
6	0	0	2	2	0	0	0
7	0	2	0	2	0	0	0
8	0	0	2	2	0	0	0
9	0	2	0	2	0	0	0
10	2	0	0	2	0	0	0
11	2	0	0	2	0	0	0
12	2	0	0	2	0	0	0
13	0	1	1	2	0	0.5	0.5
14	0	0	2	2	0	0	0
15	2	0	0	2	0	0	0
16	1	0	1	2	0.5	0	0.5
17	0	2	0	2	0	0	0
18	2	0	0	2	0	0	0
19	1	0	1	2	0.5	0	0.5
20	0	2	0	2	0	0	0
21	0	2	0	2	0	0	0
22	0	1	1	2	0	0.5	0.5
23	0	2	0	2	0	0	0
24	0	2	0	2	0	0	0
25	0	2	0	2	0	0	0

Calcul kappa par catégorie

Somme catégories (x_{.c})

bad	good	indet
439	907	654

Moyenne des juges	2

Proprotion catégories (f_c)

bad	good	indet
0.2195	0.4535	0.3270

kappa par catégorie

bad	good	indet
0.816133	**0.852726**	**0.691009**

Calcul kappa global

Pondération f_c (1 - f_c)

0.171320	0.247838	0.220071

Produit : Kappa x pondération

0.139820	0.211338	0.152071

Kappa généralisé	0.787244

Fig. 6.4. CREDIT GERMAN - Calcul du $\hat{\kappa}$ généralisé, nombre de juges variable (25 premières observations)

1. Notre tableau de données au format (Tableau 6.1) a été complété par 3 nouvelles colonnes pour obtenir les quantités $\frac{x_{ic}(m_i - x_{ic})}{m_i}$. Nous ne reproduisons que les valeurs pour les 25 premières observations ici. Il est dès lors possible de constituer les formules consolidées dans la partie droite de la feuille EXCEL.

8. Voir http://kappa.chez-alice.fr/Kappa_plusjuges_pmod.htm

2. Nous formons les sommes par catégories $\sum_i x_{ic} = \{439, 907, 654\}$, pour obtenir les proportions, il nous faut au préalable calculer le nombre moyen de juges $\bar{m} = 2$, viennent alors les proportions $f_c = \{0.2195, 0.4535, 0.3270\}$.
3. Nous pouvons calculer les coefficients de concordance par catégorie $\hat{\kappa}(c) = \{0.816133, 0.852726, 0.691009\}$.

Premier résultat très intéressant : les experts s'accordent pour attribuer l'étiquette "good" aux clients demandeurs de crédit. Nous l'avions perçu déjà dans le tableau de contingence croisant les deux juges (Tableau 6.3). Mais ce type de tableau est impossible à établir dès que le nombre de juges augmente et/ou lorsqu'il est irrégulier, différent pour chaque dossier à évaluer. Voilà un résultat qui doit rassurer le banquier, la désignation des "bons" clients repose sur un consensus fort.

Il en est de même pour les clients défaillants. La concordance pour la catégorie "bad" est 0.816133. Ici également, il y a un consensus fort lorsque les experts apposent l'étiquette "mauvais client" à un individu demandeur de crédit.

Enfin, et ça paraît normal après coup, lorsque les experts ne savent pas trop quoi penser d'un client, qu'ils classent en "indéterminé", les avis sont partagés $\hat{\kappa}(indet) = 0.691009$, moins tranché en tous les cas que pour les 2 autres catégories. C'est un résultat qui paraît de bon sens.

Reste à savoir maintenant ce qu'il en est de la concordance globale :

1. Nous déterminons le coefficient de pondération $f_c(1 - f_c)$;
2. Le coefficient global devient

$$\hat{\kappa}_g = \frac{0.171320 \cdot 0.816133 + 0.247838 \cdot 0.852726 + 0.220071 \cdot 0.691009}{0.171320 + 0.247838 + 0.220071} = 0.787244$$

Nous aboutissons avec cette démarche généralisée à une valeur identique à celle de la méthode de Fleiss qui, elle, ne s'applique que dans le cas particulier du nombre de juges constant.

Partie III

Association partielle

TODO

Partie IV

Traitement des variables ordinales

7

Caractériser les associations ordinales - Inadéquation de la corrélation

Cette partie est un peu à part dans notre support. Nous traitons toujours des variables qualitatives, les valeurs sont dénombrables, sauf que cette fois-ci, les modalités sont ordonnées. Les exemples sont nombreux, une échelle de satisfaction de la clientèle (ex. mécontent, indifférent, satisfait, très satisfait), les risques de défaillance attribuées à un demandeur de crédit par un expert (ex. peu risquée, risquée, très risquée), etc. Parfois, la variable ordinale traduit une information que l'on pourrait véhiculer à l'aide d'une variable quantitative (ex. la taille est transformée en graduation : petit, moyen, grand). Il ne faut pas y voir forcément un appauvrissement de l'information manipulée dans ce cas. Il peut s'agir d'une stratégie pour appréhender la non-linéarité d'une relation avec une autre variable par exemple.

Puisque les valeurs sont ordonnées, pourquoi ne pas utiliser les techniques dédiées aux traitement des variables quantitatives telles que l'analyse de corrélation ?

Il y a plusieurs réponses, nous y reviendrons en détail dans la section qui vient. Nous dirons simplement que cette partie a toute sa place dans ce support dans la mesure où, puisque les modalités sont dénombrables, **il nous est possible de représenter l'association des variables par un tableau de contingence**. Sauf, qu'à la différence des variables nominales, l'ordre des colonnes (lignes) est définie à l'avance. **Il n'est pas possible d'intervertir les colonnes (lignes) du tableau sans dénaturer l'information** que portent les données, et fausser les indicateurs que nous calculerons pour mesurer l'intensité de la liaison.

7.1 Caractérisation des associations

L'étude de l'association entre variables ordinales s'appuie toujours sur le tableau de contingence générique (Tableau 1.1). La principale différence est que les modalités de X (Y) sont ordonnées c.-à-d. $x_1 < x_2 < \ldots < x_C$ ($y_1 < y_2 < \ldots < y_L$). L'ordre des colonnes et des lignes est défini à l'avance. Dans le cas des variables nominales, l'interversion des colonnes (lignes) n'avait aucune incidence sur les indicateurs calculés.

L'objectif est d'évaluer si une augmentation de X est associée à une augmentation (resp. diminution) de Y, auquel cas, nous sommes en présence d'une association positive (resp. négative). Nous nous intéressons aux relations monotones, mais elles ne sont pas forcément linéaires. Dans cette section, nous

caractérisons les types de liens qui peuvent exister entre 2 variables ordinales à partir de différentes configurations du tableau de contingence (Tableau 7.1).

Remarque importante, nous considérerons que les variables ne jouent pas le même rôle, Y est la variable à expliquer (dépendante), X est la variable explicative (indépendante). Cela n'a aucune incidence sur les mesures symétriques. Les indicateurs asymétriques en revanche sauront en tenir compte.

Description	Exemples de tableaux de contingence
Monotonie stricte : lorsque X augmente, Y augmente également (décroît pour une liaison négative). On parle d'association "parfaite" lorsque que chaque valeur de X correspond une seule valeur de Y.	Y vs. X | x_1 | x_2 | x_3 : y_1 15 0 0 ; y_2 0 15 0 ; y_3 0 0 15 — ou — Y vs. X | x_1 | x_2 | x_3 | x_4 : y_1 15 0 0 0 ; y_2 0 15 0 0 ; y_3 0 0 0 15
Monotonie ordonnée : lorsque X augmente, Y augmente également. Mais il peut y avoir des ex-aequo : à une valeur de X correspond plusieurs valeurs de Y.	Y vs. X | x_1 | x_2 | x_3 : y_1 15 0 0 ; y_2 0 15 0 ; y_3 0 15 0 — ou — Y vs. X | x_1 | x_2 | x_3 | x_4 : y_1 15 0 0 0 ; y_2 0 0 0 15 ; y_3 0 0 0 15
Monotonie prédictive : à chaque valeur de X, on peut associer une et une seule valeur de Y. L'association reste monotone.	Y vs. X | x_1 | x_2 | x_3 : y_1 15 0 0 ; y_2 0 15 15 ; y_3 0 0 0 — ou — Y vs. X | x_1 | x_2 | x_3 | x_4 : y_1 15 15 0 0 ; y_2 0 0 15 15 ; y_3 0 0 0 0
Monotonie faible : lorsque X augmente, Y augmente également. Mais il y a des ex-aequo, tant sur X que sur Y.	Y vs. X | x_1 | x_2 | x_3 : y_1 15 0 0 ; y_2 15 0 0 ; y_3 15 15 15 — ou — Y vs. X | x_1 | x_2 | x_3 | x_4 : y_1 15 0 0 0 ; y_2 15 15 0 0 ; y_3 0 0 15 15
Relation non monotone : lorsque X augmente, Y augmente puis décroît. Les indicateurs pour variables ordinales ne détectent pas une association. Pourtant, en termes de prédiction, la relation est intéressante : à chaque valeur de X correspond une seule valeur de Y. Les mesures dédiées aux variables nominales, en descendant d'un cran dans le type d'information exploitée (ne pas tenir compte du caractère ordonné des variables), permettrait de mettre à jour cela.	Y vs. X | x_1 | x_2 | x_3 : y_1 15 0 15 ; y_2 0 0 0 ; y_3 0 15 0 — ou — Y vs. X | x_1 | x_2 | x_3 | x_4 : y_1 15 0 0 15 ; y_2 0 15 0 0 ; y_3 0 0 0 0

Tableau 7.1. Quelques associations caractéristiques entre variables ordinales

7.2 Variables ordinales et corrélation

Le coefficient de corrélation r de Pearson est destiné à mesurer la liaison entre 2 variables quantitatives. Un de ses mérites est d'être très largement diffusé et reconnu dans de nombreux domaines. Nous avons déjà exposé plus haut les formulations et tests associés à cet indicateur (Section 4.1).

Amplitude de l'écart entre les modalités

Puisque les modalités sont ordonnées, le premier réflexe serait de les coder $\{1, 2, 3, \cdots\}$ pour calculer le coefficient de corrélation pour évaluer l'association entre 2 variables ordinales. Cette approche, qui n'est pas dénuée de bon sens, comporte cependant certaines lacunes qu'il convient de circonscrire pour délimiter la portée des résultats.

En effet, en codant $\{1, 2, 3\}$ une variable ordinale comportant 3 modalités, {petit, moyen, grand} par exemple, nous introduisons implicitement plusieurs informations :

1. L'amplitude de l'écart entre les modalités est la même, c.-à-d. l'écart entre *petit* et *moyen* serait le même qu'entre *moyen* et *grand*. Peut-être que c'est vrai, mais peut être aussi que c'est totalement erroné. Tout dépend de l'analyse que l'on mène et du contexte de l'étude. En tous les cas, si nous adoptons ce codage, nous validons cette vision de l'écart entre les modalités.
2. Autre information implicite sur l'échelle cette fois-ci. Avec ce codage nous décidons que l'écart entre la troisième et la première modalité est 2 fois plus importante que l'écart entre les 2 premières. Encore une fois, peut-être est-ce justifié, peut-être que non. En tous les cas, ce type de codage avalise ce type d'information.

De fait, en l'absence d'informations sur l'amplitude et l'échelle des écarts, un codage $\{1, 25, 26\}$ est tout aussi justifié qu'un codage $\{1, 2, 3\}$. Pourtant, calculer le coefficient de corrélations sur les données codées différemment aboutit à des valeurs différentes, au risque de déboucher sur des conclusions contradictoires lors de l'inférence statistique.

Linéarité et monotonie de l'association

Autre point litigieux, corollaire au précédent, le coefficient de corrélation de Pearson sert avant tout à caractériser les associations linéaires. En choisissant judicieusement le codage des valeurs de X et Y, nous pouvons rendre linéaire le positionnement relatifs de observations. Mais il est évident que l'opération est périlleuse. Le tâtonnement pour trouver le codage "optimal" est assez improbable dans la pratique.

Nous pourrions dès lors nous tourner vers le coefficient de corrélation de rangs ρ de Spearman. Ce dernier s'affranchit de la linéarité en transformant les données en rangs. Il détecte les associations monotones. Il semble tout à fait approprié ici. Mais un autre problème apparaît. A cause du faible nombre de modalités des variables que nous manipulons, les ex-aequo sont très nombreux. Le coefficient de Spearman a du mal à caractériser correctement l'association entre les variables. Son utilisation n'est pas conseillée dans notre cadre.

Problème de l'hypothèse (sous-jacente) de normalité

Concernant le test de significativité, rappelons qu'il s'agit d'un test d'indépendance si, et seulement si, les variables suivent une distribution normale. Le test n'est pas symétrique. Si les variables sont indépendantes alors $r = 0$, quelle que soient leur distributions. La réciproque - c.-à-d. $r = 0$ implique l'indépendance - n'est vraie que si X et Y suivent une loi normale.

Ainsi, s'appuyer sur le test de significativité du coefficient de corrélation $H_0 : r = 0$ vs. $H_1 : r \neq 0$ pour caractériser l'indépendance entre les variables ordinales est assez hasardeux. De par la nature discrète des variables, le nombre de valeurs possibles est faible, l'hypothèse de normalité n'est, de près ou de loin, absolument pas crédible.

Avantage parfois à réunir en intervalles les données continues

Enfin, et c'est un argument fort qui milite en faveur des indicateurs que nous étudions dans ce chapitre, il peut être opportun de transformer une variable quantitative en une variable ordinale dans de nombreuses études. On parle de "discrétisation".

Prenons le cas de l'âge, une variable essentiellement quantitative. Selon l'étude que nous menons, il est plus approprié de la découper en intervalles. Si nous étudions les accidents au volant, il y a déjà 2 populations évidentes, les moins de 18 ans qui ne sont pas censés conduire seuls des véhicules puisqu'il n'ont pas le droit d'avoir le permis de conduire, et les plus de 18 ans qui peuvent l'avoir. Si nous étudions les comportements d'achats, nous pouvons découper différemment la même variable, en distinguant cette fois-ci les très jeunes, inactifs, les jeunes actifs qui effectuent essentiellement des petits boulots, les actifs, dans la force de l'âge, et les retraités qui ont un comportement différent.

Ainsi, la variable, originellement quantitative, est transformée pour mieux répondre au contexte et aux objectifs de l'étude. Il nous faut utiliser les indicateurs appropriés dans cette nouvelle situation.

7.3 Les comparaisons par paires pour caractériser les associations ordinales

Paires concordantes et paires discordantes

Les mesures présentées dans ce chapitre s'appuient sur le principe des comparaisons par paires pour caractériser l'association entre variables nominales :

- On dit que 2 paires d'observations i et j sont concordantes si $(x_i > x_j)$ (resp. $(x_i < x_j)$) alors $(y_i > y_j)$ (resp. $(y_i < y_j)$).
- Elles sont discordantes si $(x_i > x_j)$ (resp. $(x_i < x_j)$) alors $(y_i < y_j)$ (resp. $(y_i > y_j)$)
- Enfin, il y a ex-aequo si $(x_i = x_j)$ ou $(y_i = y_j)$

Les indicateurs que nous étudierons reposent pour l'essentiel sur la confrontation du nombre P de paires concordantes et Q de paires discordantes.

Dénombrement des paires à partir d'un tableau de contingence

Pour calculer le nombre de paires concordantes et discordantes sur un échantillon de taille n, il faudrait a priori comparer chaque observation avec tous les autres, et ceci n fois : il y a n^2 comparaisons à faire, ce qui peut s'avérer très coûteux en temps de calcul si n est élevé.

Heureusement, il est possible de calculer les valeurs de P et Q à partir du tableau de contingence générique (Tableau 1.1) croisant les variables Y et X. Comme nous le faisions remarquer plus haut

(Section 7.1), à la différence du tableau pour les variables nominales, l'ordonnancement des lignes et des colonnes est défini par l'ordre des modalités. Il ne saurait y avoir de permutations des lignes ou des colonnes.

Le nombre de paires concordantes est égal à

$$P = \sum_l \sum_c n_{lc} C_{lc} \tag{7.1}$$

où C_{lc} est définie par la somme des éléments situés en haut à gauche ($h < l, k < c$) et en bas à droite ($h > l, k > c$) de la case de coordonnées (l, c), c.-à-d.

$$C_{lc} = \sum_{h<l} \sum_{k<c} n_{hk} + \sum_{h>l} \sum_{k>c} n_{hk} \tag{7.2}$$

De manière symétrique, le nombre de paires discordantes est égal à

$$Q = \sum_l \sum_c n_{lc} D_{lc} \tag{7.3}$$

où D_{lc} est définie par la somme des éléments situés en bas à gauche ($k > l, h < c$) et en haut à droite ($h < l, k > c$) de la case de coordonnées (l, c) c.-à-d-.

$$D_{lc} = \sum_{h>l} \sum_{k<c} n_{hk} + \sum_{h<l} \sum_{k>c} n_{hk} \tag{7.4}$$

Nous définirons également deux autres quantités :

$$D_X = n^2 - \sum_c n_{.c}^2 \tag{7.5}$$

D_X est le nombre de paires qui n'ont pas d'ex-aequo sur la variable X ; de même,

$$D_Y = n^2 - \sum_l n_{l.}^2 \tag{7.6}$$

est le nombre de paires qui ne présentent pas la même valeur sur la variable Y.

Prenons le cas de la variable X pour expliciter les calculs : si l'individu i porte la modalité x_1, il y a $n_{.1}$ observations qui portent la même valeur que lui (il doit se compter dedans) : il y a donc $n_{.1}^2$ paires d'observations qui ont la même valeur $X = x_1$. Il en est de même pour les autres modalités, comme il y a n^2 paires possibles, le nombre de paires n'ayant pas la même valeur sur X est bien $D_X = n^2 - \sum_c n_{.c}^2$

Rendons toujours à César ce qui lui appartient, les formules que nous présentons dans cette partie proviennent pour l'essentiel du site de documentation du logiciel SAS : http://v8doc.sas.com/sashtml/stat/chap28/sect20.htm.

7.4 Exemple "Amount vs. Duration" sur le fichier GERMAN-CREDIT

Dans ce chapitre, nous étudierons la relation entre la durée et le montant du crédit dans le fichier GERMAN CREDIT. L'idée à valider est relativement simple : un montant de crédit X élevé induit-il une durée de remboursement Y longue ?

Le premier réflexe serait d'étudier directement la liaison entre les deux variables quantitatives. Si nous calculons le coefficient de corrélation, nous obtenons la valeur $r = 0.6250$, très largement significative avec une p-value $p_c < 0.000001$. En traçant le graphique croisant les deux variables, nous constatons cependant que la situation est plus complexe qu'elle ne le semblait initialement : si la liaison paraît positive, les disparités deviennent fortes à mesure que le montant des emprunts sont élevés (Figure 7.1). Ceci montre, encore une fois, combien les graphiques sont précieux dans une phase exploratoire[1].

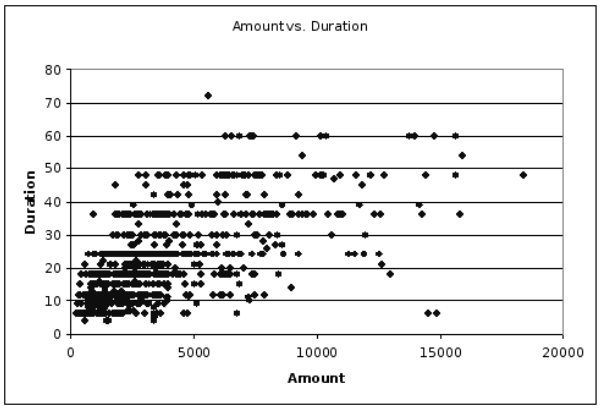


Fig. 7.1. CREDIT GERMAN - Croisement "montant du crédit" et "durée du crédit", données quantitatives

On pourrait s'appuyer sur des artifices pour stabiliser la variance, par exemple en passant aux logarithmes. Mais ici intervient un élément supplémentaire. Souvent les opérateurs raisonnent en termes de paliers sur les durées : crédit à 6 mois, crédit à 12 mois, etc. ; il pourrait en être de même concernant les montants. Il est donc décidé de découper les montants en 3 intervalles : moins de 1000, entre 1000 et 2000, et supérieur à 2000. De même concernant les durées, nous la subdivisons en 3 intervalles : moins de 1 an, entre 1 an et 2 ans, supérieur à 2 ans.

Si nous codons les variables Y et X en $\{1, 2, 3\}$, le coefficient de corrélation est égal à $r = 0.5063$. Il reste très significatif, un montant élevé semble bien conduire à une durée de remboursement plus longue. Ce constat s'accompagne bien évidemment des réserves d'usage quant à la pertinence du codage adopté.

Dans le cadre qui nous intéresse, nous construisons le tableau de contingence croisant, en ligne, la durée du crédit, avec en colonne, le montant (Figure 7.2). C'est à partir de ce tableau que nous illustrerons tous les indicateurs présentés dans ce chapitre. Il n'est plus question de construire le nuage de points, les

1. Sur l'art et la manière de réaliser des graphiques simples "intéressants" pour explorer les données, voir l'excellent ouvrage de Jacoby W., *Statistical Graphics for Univariate and Bivariate Data*, Sage University Paper 117, Quantitative Applications in the Social Sciences Series, 1997.

	disc_amount			
disc_duration	lo_1000	1000_2000	up_2000	Total
lo_1_year	93	161	105	359
1_2_years	22	146	243	411
up_2_years	1	9	220	230
Total	116	316	568	1000

Fig. 7.2. CREDIT GERMAN - Croisement "montant du crédit" et "durée du crédit", données discrétisées

observations sont superposées en 9 blocs de points, masquant toute information pertinente. Il faudrait dans ce cas associer à chaque coordonnée une artifice graphique, un cercle de surface plus ou moins grande par exemple, pour indiquer le nombre d'observations. On pourrait également construire un histogramme à 3 dimensions : la hauteur indiquerait alors l'effectif dans chaque case du tableau de contingence. Ce type de représentation est valable que les variables soient ordinales ou nominales d'ailleurs. Néanmoins, les effets de perspective rendent la lecture du graphique ardue (Figure 7.3).

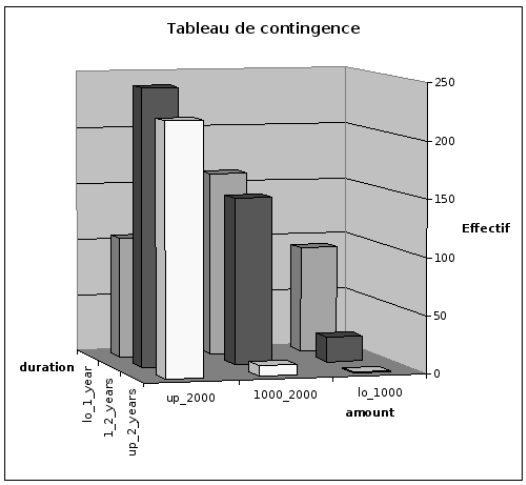


Fig. 7.3. CREDIT GERMAN - Croisement "montant du crédit" et "durée du crédit", données discrétisées - Histogramme 3D

Comptage paires concordantes et discordantes					
Case	n_{Ic}	C_{Ic}	n_{Ic} x C_{Ic}	D_{Ic}	n_{Ic} x D_{Ic}
(1,1)	93	618	57474	0	0
(1,2)	161	463	74543	23	3703
(1,3)	105	0	0	178	18690
(2,1)	22	229	5038	266	5852
(2,2)	146	313	45698	106	15476
(2,3)	243	254	61722	10	2430
(3,1)	1	0	0	655	655
(3,2)	9	115	1035	348	3132
(3,3)	220	422	92840	0	0
		P	338350	Q	49938

Fig. 7.4. CREDIT GERMAN - Croisement "montant du crédit" et "durée du crédit". Comptage des paires concordantes et discordantes

Nous détaillons le calcul des paires concordantes et discordantes (Figure 7.4). Au final, nous avons $P = 338350$ paires concordantes et $Q = 49938$ paires discordantes. Pour le calcul des paires concordantes, détaillons quelques cas pour préciser les idées. Les valeurs proviennent du tableau de contingence (Tableau 7.2) :

$$
\begin{aligned}
C_{11} &= 146 + 243 + 9 + 220 = 618 \\
C_{12} &= 243 + 220 = 463 \\
&\dots \\
C_{22} &= 93 + 220 = 313 \\
C_{23} &= 93 + 161 = 254 \\
&\dots \\
C_{33} &= 93 + 161 + 22 + 146 = 422
\end{aligned}
$$

8

Les mesures symétriques

8.1 γ de Goodman et Kruskal

8.1.1 Définition et estimation

Le γ de Goodman et Kruskal est défini par l'écart (normalisé) entre les probabilités que 2 observations prises au hasard soient concordants ou discordants (Siegel, 1988 ; page 292) :

$$\begin{aligned}\gamma &= \frac{P(\text{X et Y sont concordants}) - P(\text{X et Y sont discordants})}{1 - P(\text{ex-aequo sur X ou sur Y})} \\ &= \frac{P(\text{X et Y sont concordants}) - P(\text{X et Y sont discordants})}{P(\text{X et Y sont concordants}) + P(\text{X et Y sont discordants})}\end{aligned}$$

Par définition, $-1 \leq \gamma \leq 1$, il indique le surplus de paires concordantes sur les paires discordantes. La mesure est normalisée pour varier entre -1 et $+1$. Il qualifie comme parfaite une monotonie faible (voir tableau descriptif 7.1).

L'indicateur γ peut s'interpréter comme une mesure PRE (Proportional Reduction in Error). Il indique la réduction de l'erreur de prédiction de **l'ordre** des valeurs de 2 observations de Y si l'on se fonde sur l'ordre des valeurs prises par X.

Quoiqu'il en soit, lorsque $\gamma \approx 1$ (resp. -1), pour deux observations quelconques i et j, si $x_i > x_j$ alors il est presque certain que $y_i > y_j$ (resp. $y_i < y_j$). Lorsque les variables sont indépendantes, $\gamma = 0$. Gardons cependant à l'esprit que γ peut être nul dans d'autres circonstances, dès que le nombre de paires concordantes est égal au nombre de paires discordantes en fait.

Sur un échantillon de données, l'estimateur naturel de γ est

$$\hat{\gamma} = \frac{P - Q}{P + Q} \tag{8.1}$$

Remarque 24 (γ et $\mathbb{Q}$ de Yule). Le $\mathbb{Q}$ de Yule (Section 4.3) que nous avons étudié tantôt pour les variables binaires est un cas particulier du γ de Goodman et Kruskal.

Amount vs. Duration sur les données CREDIT

Dans notre exemple, l'estimateur est vite calculé :

$$\hat{\gamma} = \frac{338350 - 49938}{338350 + 49938} = 0.742779$$

Indubitablement, il semble exister un lien positif fort entre le montant du crédit et la durée de remboursement. Les outils de statistique inférentielle devront nous confirmer cela.

8.1.2 Intervalle de confiance

La construction de l'intervalle de variation au niveau $(1-\alpha)$ s'appuie sur la normalité asymptotique de la distribution de $\hat{\gamma}$. Il nous reste à produire une estimation de la variance pour compléter les calculs, elle s'écrit :

$$\sigma^2_{\hat{\gamma}} = \frac{16}{(P+Q)^4}\left[\sum_l \sum_c n_{lc}\left(Q \times C_{lc} - P \times D_{lc}\right)^2\right] \tag{8.2}$$

Nous en déduisons l'intervalle de confiance en introduisant le quantile d'ordre $(1-\alpha/2)$ de la loi normale centrée réduite $u_{1-\frac{\alpha}{2}}$:

$$\gamma \in \left[\hat{\gamma} \pm u_{1-\frac{\alpha}{2}} \times \sigma_{\hat{\gamma}}\right] \tag{8.3}$$

Amount vs. Duration sur les données CREDIT

Détaillons les calculs à l'aide d'une copie d'écran de notre feuille EXCEL (Figure 8.1). Les opérations sont consignées dans le tableau **A**. En partant de gauche à droite, nous construisons les colonnes destinées à produire une estimation de la variance asymptotique

P	338350
Q	49938
gamma	0.742779
sigma^2	0.000692
sigma	0.026313
u(0.95)	1.644853
Intervalle de confiance	
B.Basse	0.699498
B.Haute	0.786059

Calcul de la variance asymptotique (A)

Case	n_{lc}	C_{lc}	D_{lc}	Q x C_{lc}	P x D_{lc}	Sum_l Sum c
(1,1)	93	618	0	30861684	0	8.8577E+16
(1,2)	161	463	23	23121294	7782050	3.7882E+16
(1,3)	105	0	178	0	60226300	3.8086E+17
(2,1)	22	229	266	11435802	90001100	1.3580E+17
(2,2)	146	313	106	15630594	35865100	5.9778E+16
(2,3)	243	254	10	12684252	3383500	2.1020E+16
(3,1)	1	0	655	0	221619250	4.9115E+16
(3,2)	9	115	348	5742870	117745800	1.1290E+17
(3,3)	220	422	0	21073836	0	9.7703E+16
					Somme	9.8363E+17

sigma^2 (H0)	0.001265
sigma (H0)	0.035568
Test de significativité	
z	20.883351
p-value	0.000000

Calcul de la variance asymptotique sous H0 - Indépendance (B)

Case	n_{lc}	C_{lc}	D_{lc}	(C_{lc} - D_lc})^2	n_{lc} x ()^2
(1,1)	93	618	0	381924	35518932
(1,2)	161	463	23	193600	31169600
(1,3)	105	0	178	31684	3326820
(2,1)	22	229	266	1369	30118
(2,2)	146	313	106	42849	6255954
(2,3)	243	254	10	59536	14467248
(3,1)	1	0	655	429025	429025
(3,2)	9	115	348	54289	488601
(3,3)	220	422	0	178084	39178480
				Somme	130864778

Fig. 8.1. CREDIT GERMAN - Croisement "montant du crédit" et "durée du crédit". Calcul du γ de Goodman et Kruskal, intervalle de confiance et test de significativité

— Nous observons tout d'abord les coordonnées de cellules (l, c), les effectifs n_{lc}, les valeurs de C_{lc} et de D_{lc}

— Nous formons alors les quantités $Q \times D_{lc}$ et $P \times C_{lc}$

— Puis la double somme $\sum_l \sum_c n_{lc} \left(Q \times C_{lc} - P \times D_{lc}\right)^2 = 9.8363 \cdot 10^{17}$

— Reste alors à calculer la variance $\sigma^2_{\hat{\gamma}} = \frac{16}{(338350+49938)^4} \times 9.8363 \cdot 10^{17} = 0.000692$

— L'estimation de l'écart type est $\sqrt{0.000692} = 0.026313$.

Nous produisons enfin les bornes basses et hautes de l'intervalle de confiance au niveau 90%, avec le quantile $u_{0.95} = 1.644853$

$$bb = 0.742779 - 1.644853 \times 0.026313 = 0.699498$$
$$bh = 0.742779 + 1.644853 \times 0.026313 = 0.786059$$

Nous retenons essentiellement qu'il y a 95% de chances que la "vraie" valeur de γ soit supérieure à 0.699498. La durée du crédit est manifestement liée au montant du crédit. Ce qui n'est guère étonnant en fin de compte.

8.1.3 Test de significativité

Pour tester la significativité du coefficient $H_0 : \gamma = 0$ vs. $H_1 : \gamma \neq 0$, il nous faut produire une estimation de la variance en accord avec l'hypothèse nulle d'indépendance. Elle s'écrit :

$$\sigma^2_{\hat{\gamma}}(0) = \frac{4}{(P+Q)^2} \left[\sum_l \sum_c n_{lc} \left(C_{lc} - D_{lc}\right)^2 - \frac{1}{n}(P-Q)^2 \right] \tag{8.4}$$

Nous formons la statistique $z_\gamma = \frac{\hat{\gamma}}{\sigma_{\hat{\gamma}}(0)}$, la région critique du test pour un risque α devient

$$R.C. : |z_\gamma| > u_{1-\frac{\alpha}{2}}$$

Nous pouvons aussi fonder la décision sur la p-value du test.

Amount vs. Duration sur les données CREDIT

Nous reprenons notre feuille de calcul (Figure 8.1) pour nous concentrer sur le tableau **B** cette fois. Le détail des colonnes est le suivant :

— Nous observons tout d'abord les coordonnées de cellules (l, c), les effectifs n_{lc}, les valeurs de C_{lc} et de D_{lc}

— Nous formons la différence $(C_{lc} - D_{lc})^2$, puis la somme $\sum_l \sum_c n_{lc}(C_{lc} - D_{lc})^2 = 130864778$

— La variance est obtenue avec $\sigma^2_{\hat{\gamma}}(0) = \frac{4}{(338350+49938)^2} \times (130864778 - \frac{1}{1000}(338350 - 49938)^2) = 0.001265$

— Et l'écart type 0.035568

— La statistique réduite $z_\gamma = \frac{0.742779}{0.035568} = 20.883351$

La p-value du test est $p_c < 0.000001$. Le lien est très fortement significatif, ce qu'indiquait déjà l'intervalle de confiance.

8.2 τ_b de Kendall

8.2.1 Définition et estimation

Le τ_b de Kendall mesure l'association entre variables ordinales. Il est particulièrement indiqué pour les tableaux de contingence carrés, bien que son domaine d'action ne soit pas limité à ce type de tableau. A l'instar du coefficient γ de Goodman et Kruskal, il est défini par l'excès de paires concordantes sur les paires discordantes, à la différence que le facteur de normalisation est la moyenne géométrique du nombre de paires distinctes (non ex-aequos) sur X et sur Y.

Le coefficient τ_b est symétrique, il varie entre -1 (association négative) et $+1$ (association positive). Lorsque X et Y sont indépendants, $\tau_b = 0$. Il requiert la monotonie stricte (Tableau 7.1) pour atteindre la valeur $+1$ (ou -1 en cas d'association négative). Cependant, par rapport aux autres mesures, il pénalise moins les ex-aequo.

Sur un échantillon de taille n, il est estimé avec

$$\hat{\tau}_b = \frac{P - Q}{\sqrt{D_Y \cdot D_X}} \tag{8.5}$$

Amount vs. Duration sur les données CREDIT

Nous résumons tous les calculs concernant le τ_b dans une feuille EXCEl (Figure 8.2). Dans un premier temps, seul le tableau **A** nous intéresse pour construire l'estimation. Nous calculons :

- Nous avons $P = 338350$ paires concordantes et $Q = 49938$ paires discordantes.
- Dans le tableau **A**, nous avons les effectifs marginaux en ligne (resp. en colonne) $n_{l.}$ (resp. $n_{.c}$) et leurs carrés $n_{l.}^2$ (resp. $n_{.c}^2$).
- Le nombre d'ex-aequo sur Y est égal à $\sum_l n_{l.}^2 = 350702$ (resp. sur X, $\sum_c n_{.c} = 435936$).
- Le nombre de paires distinctes sur Y est ainsi égal à $D_Y = 1000^2 - 350702 = 649298$ (resp. sur X, $D_X = 1000^2 - 435936 = 564064$.

Nous disposons de tous les éléments pour estimer l'indicateur τ_b :

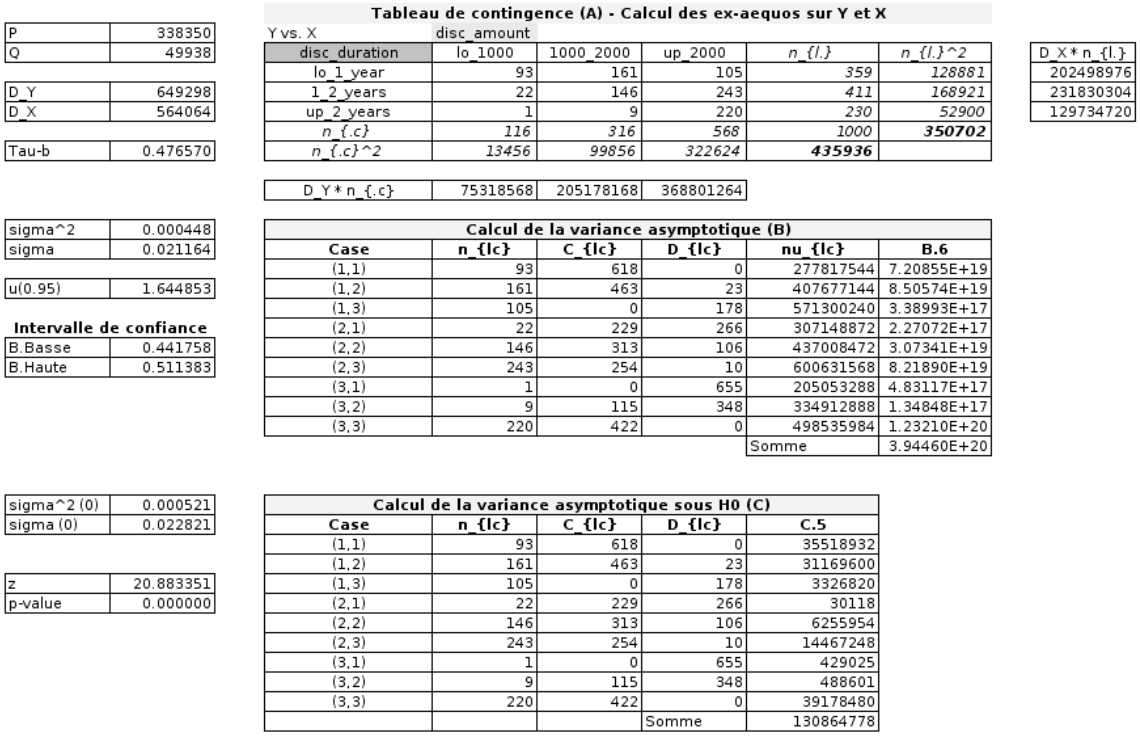
$$\hat{\tau}_b = \frac{338350 \cdot 49938}{\sqrt{649298 \cdot 564064}} = 0.476570$$

Même si τ_b et γ sont définis dans le même intervalle $[-1; +1]$, nous avions mentionné que le τ_b était plus exigeant, n'atteignant la valeur extrême qu'en cas de monotonie stricte. Cette caractéristique est confirmée sur cette exemple, pour la même situation, le τ_b est nettement inférieur à γ.

8.2.2 Intervalle de confiance

Pour produire un intervalle de variance de confiance $(1 - \alpha)$, il nous faut produire la formule de la variance asymptotique. Elle est particulièrement complexe :

$$\sigma^2_{\hat{\tau}_b} = \frac{1}{(D_Y D_X)^2} \left[\sum_l \sum_c n_{lc} \left(2\sqrt{D_Y D_X}(C_{lc} - D_{lc}) + \hat{\tau}_b \cdot \nu_{lc} \right)^2 - n^3 \hat{\tau}_b^2 (D_Y + D_X)^2 \right] \tag{8.6}$$



P	338350
Q	49938
D_Y	649298
D_X	564064
Tau-b	0.476570

Tableau de contingence (A) - Calcul des ex-aequos sur Y et X

Y vs. X	disc_amount				
disc_duration	lo_1000	1000_2000	up_2000	*n_{l.}*	*n_{l.}^2*
lo_1_year	93	161	105	*359*	*128881*
1_2_years	22	146	243	*411*	*168921*
up_2_years	1	9	220	*230*	*52900*
n_{.c}	*116*	*316*	*568*	*1000*	***350702***
n_{.c}^2	*13456*	*99856*	*322624*	***435936***	
D_Y * n_{.c}	75318568	205178168	368801264		

D_X * n_{l.}
202498976
231830304
129734720

sigma^2	0.000448
sigma	0.021164
u(0.95)	1.644853
Intervalle de confiance	
B.Basse	0.441758
B.Haute	0.511383

Calcul de la variance asymptotique (B)

Case	n_{lc}	C_{lc}	D_{lc}	nu_{lc}	B.6
(1,1)	93	618	0	277817544	7.20855E+19
(1,2)	161	463	23	407677144	8.50574E+19
(1,3)	105	0	178	571300240	3.38993E+17
(2,1)	22	229	266	307148872	2.27072E+17
(2,2)	146	313	106	437008472	3.07341E+19
(2,3)	243	254	10	600631568	8.21890E+19
(3,1)	1	0	655	205053288	4.83117E+17
(3,2)	9	115	348	334912888	1.34848E+17
(3,3)	220	422	0	498535984	1.23210E+20
				Somme	3.94460E+20

sigma^2 (0)	0.000521
sigma (0)	0.022821
z	20.883351
p-value	0.000000

Calcul de la variance asymptotique sous H0 (C)

Case	n_{lc}	C_{lc}	D_{lc}	C.5
(1,1)	93	618	0	35518932
(1,2)	161	463	23	31169600
(1,3)	105	0	178	3326820
(2,1)	22	229	266	30118
(2,2)	146	313	106	6255954
(2,3)	243	254	10	14467248
(3,1)	1	0	655	429025
(3,2)	9	115	348	488601
(3,3)	220	422	0	39178480
			Somme	130864778

Fig. 8.2. CREDIT GERMAN - Croisement "montant du crédit" et "durée du crédit". Calcul du τ_b de Kendall, intervalle de confiance et test de significativité

où $\nu_{lc} = n_{l.}D_X + n_{.c}D_Y$

Avec la normalité asymptotique, l'intervalle de variation vient naturellement

$$\tau_b \in \left[\hat{\tau}_b \pm u_{1-\frac{\alpha}{2}} \times \sigma_{\hat{\tau}_b}\right]$$

Amount vs. Duration sur les données CREDIT

Nous nous intéressons à une autre fraction de notre feuille EXCEL (Figure 8.2, tableau **B**) :

— Pour chaque cellule du tableau de contingence, nous calculons les quantités n_{lc}, C_{lc} et D_{lc}.

— Puis dans la 5-ème colonne, nous formons la quantité $\nu_{lc} = n_{l.}D_X + n_{.c}D_Y$

— Dans la 6-ème colonne (**B.6**), nous formons les quantités dans la double somme pour obtenir le total $3.94460 \cdot 10^{20}$.

— La variance asymptotique est égale à $\sigma^2_{\hat{\tau}_b} = \frac{1}{(649298 \cdot 564064)^2} \times [3.94460 \cdot 10^{20} - 1000^3 \cdot 0.476570^2 \cdot (649298 + 564064)^2] = 0.000448$

— Et l'écart type $\sqrt{0.000448} = 0.021164$

Nous pouvons maintenant obtenir la borne basse et haute de l'intervalle de confiance à 90%

$$b.basse = 0.476570 - 1.644853 \cdot 0.021164 = 0.441758$$

$$b.haute = 0.476570 + 1.644853 \cdot 0.021164 = 0.511383$$

La borne basse de l'intervalle de confiance est largement au delà de la valeur 0. Cela sera confirmé par le test de significativité dans la section suivante.

8.2.3 Test de significativité

Pour mettre en oeuvre le test de significativité, il nous faut une estimation de la variance sous l'hypothèse nulle, elle est (un peu) simplifiée

$$\sigma^2_{\hat{\tau}_b}(0) = 4\left[\frac{\sum_l \sum_c n_{lc}\left(C_{lc} - D_{lc}\right)^2 - \frac{1}{n}(P-Q)^2}{D_Y D_X}\right] \tag{8.7}$$

La statistique réduite $z_{\tau_b} = \frac{\hat{\tau}_b}{\sigma_{\hat{\tau}_b}(0)}$ permet de définir la région critique du test pour un risque α

$$R.C. : |z_{\tau_b}| > u_{1-\frac{\alpha}{2}}$$

Amount vs. Duration sur les données CREDIT

Intéressons-nous maintenant au tableau **C** de notre feuille EXCEL (Figure 8.2) :

- Pour chaque cellule du tableau de contingence, nous calculons toujours les quantités n_{lc}, C_{lc} et D_{lc}.
- Dans la 5-ème colonne **C.5**, nous calculons les quantités dans la double somme pour obtenir la somme totale 130864778
- La variance asymptotique sous l'hypothèse d'indépendance est obtenue avec $\sigma^2_{\hat{\tau}_b}(0) = 4\left[\frac{130864778 - \frac{1}{1000}(338350-49938)^2}{649298 \cdot 564064}\right] = 0.000521$
- Et l'écart type associé $\sqrt{0.000521} = 0.022821$
- La statistique du test réduite est égale à $z_{\tau_b} = \frac{0.476570}{0.022821} = 20.883351$

Au final, la p-value du test est $p_c < 0.000001$, indiquant, si besoin était au vu des calculs précédents, que la durée et le montant du crédit sont fortement liés.

8.3 τ_c de Kendall

8.3.1 Définition et estimation

Dans la littérature, le τ_b est souvent rattaché à l'analyse des tableaux 2×2, ou tout du moins carrés. A tort d'ailleurs puisqu'il peut s'appliquer à tous types de dimension, rien dans sa définition n'en restreint la portée. En revanche, il apparaît effectivement qu'elle ne peut mécaniquement pas atteindre la valeur $+1$ ou -1 lorsque le tableau est rectangulaire. Et subséquemment, on présente le τ_c comme une extension à des tableaux plus larges, de dimension quelconque.

Le τ_c de Kendall, également appelé τ_c de Kendall et Stuart, ou τ_c de Stuart, est estimé de la manière suivante :

$$\hat{\tau}_c = \frac{q(P-Q)}{n^2(q-1)} \tag{8.8}$$

où $q = \min(L;C)$ est le minimum du nombre de lignes et du nombre de colonnes du tableau de contingence.

Il s'agit toujours de mesurer l'écart entre le nombre de paires concordantes et le nombre de paires discordantes. Seule la normalisation est différente, et c'est son principal intérêt, il tient compte des dimensions du tableau de contingence. De fait, il est moins sensible au choix du nombre d'intervalles lorsque l'on discrétise les variables quantitatives.

On pourrait faire le parallèle avec l'indicateur ϕ et le V de Cramer pour les mesures dérivées du χ^2 (Voir section 2.4). La nouvelle mesure est une généralisation qui s'affranchit de la dimension de la table de contingence. Ce parallèle n'est pas complètement approprié. A la différence du τ_b, nous ne tenons pas compte des ex-aequo dans la seconde mesure τ_c.

L'indicateur τ_c est une mesure symétrique définie sur l'intervalle $[-1;+1]$. Il prend sa valeur maximum (en valeur absolue) en cas d'association stricte. A la différence du τ_b, il peut atteindre les valeurs $+1$ (ou -1) même si la table n'est pas carrée.

Amount vs. Duration sur les données CREDIT

A partir des valeurs calculées tout au long de ce chapitre, nous disposons de tous les éléments pour estimer τ_c :

$$\hat{\tau}_c = \frac{3(338350 - 49938)}{1000^2(3-1)} = 0.432618$$

8.3.2 Intervalle de confiance et test de significativité

Une fois n'est pas coutume, la formule de la variance est la même que l'on soit sous l'hypothèse d'indépendance ou non. Nous pouvons utiliser la même expression pour élaborer les intervalles de confiance et effectuer le test de significativité. Elle s'écrit :

$$\sigma^2_{\hat{\tau}_c} = \sigma^2_{\hat{\tau}_c}(0) = \frac{4q^2}{n^4(q-1)^2}\left[\sum_l \sum_c n_{lc}\left(C_{lc} - D_{lc}\right)^2 - \frac{1}{n}(P-Q)^2\right] \tag{8.9}$$

En nous appuyant sur la normalité asymptotique, nous pouvons préciser l'intervalle de variation au niveau $(1-\alpha)$

$$\tau_c \in \left[\hat{\tau}_c \pm u_{1-\frac{\alpha}{2}} \times \sigma_{\hat{\tau}_c}\right]$$

Après avoir formé la statistique réduite $z_{\tau_c} = \frac{\hat{\tau}_c}{\sigma_{\hat{\tau}_c}(0)}$, nous pouvons définir la région critique du test de significativité au risque α :

$$R.C. : |z_{\tau_c}| > u_{1-\frac{\alpha}{2}}$$

Amount vs. Duration sur les données CREDIT

P	338350
Q	49938
q	3
n	1000
tau-c	0.432618
sigma^2	0.000429
sigma	0.020716
u(0.95)	1.644853

Intervalle de confiance

B.Basse	0.398543
B.Haute	0.466693

Test de significativité

z	20.883351
p-value	0.000000

Calcul de la variance asymptotique (A)

Case	n_{Ic}	C_{Ic}	D_{Ic}	C.5
(1,1)	93	618	0	35518932
(1,2)	161	463	23	31169600
(1,3)	105	0	178	3326820
(2,1)	22	229	266	30118
(2,2)	146	313	106	6255954
(2,3)	243	254	10	14467248
(3,1)	1	0	655	429025
(3,2)	9	115	348	488601
(3,3)	220	422	0	39178480
			Somme	130864778

Fig. 8.3. CREDIT GERMAN - Croisement "montant du crédit" et "durée du crédit". Calcul du τ_c de Kendall, intervalle de confiance et test de significativité

Par rapport aux précédentes, notre feuille de calcul est largement simplifiée (Figure 8.3). La principale difficulté réside dans l'évaluation de la variance asymptotique, nous utilisons le tableau **A** pour obtenir les valeurs intermédiaires nécessaires au calcul de la variance, notamment la double somme. Nous avons ainsi

$$\sigma^2_{\hat{\tau}_c} = \frac{43^2}{1000^4(3-1)^2}\left[130864778 - \frac{1}{1000}(338350 - 49938)^2\right] = 0.000429$$

Avec l'écart type $\sigma_{\hat{\tau}_c} = \sqrt{0.000429} = 0.020716$, nous pouvons calculer les bornes de l'intervalle de confiance à 90%

$$b.basse = 0.432618 - 1.644853 \times 0.020716 = 0.398543$$
$$b.haute = 0.432618 + 1.644853 \times 0.020716 = 0.466693$$

Pour le test de significativité, nous formons $z_{\tau_c} = \frac{0.432618}{0.020716} = 20.883351$. Avec un p-value $p_c < 0.000001$, nous concluons que τ_c est significativement différent de 0, il y a un lien entre la durée du crédit et le montant demandé.

8.3.3 Quelle mesure privilégier ?

Toutes les mesures étudiées jusqu'ici ont pour point commun d'être symétriques. Si l'on transpose le tableau de contingence, l'indicateur prendrait la même valeur. En pratique, τ_b et τ_c sont très proches. Nous remarquons d'ailleurs que sur notre exemple, la statistique z coïncide pour tester la significativité du coefficient. S'il y a une différence, on conseille généralement d'adopter une attitude conservatrice en choisissant le coefficient plus faible.

Le γ de Goodman et Kruskal prend généralement des valeurs plus élevées. Elle est moins restrictive. Elle est surtout intéressante à cause de son interprétation en termes de différences de probabilités.

On peut difficilement donner une hiérarchie des mesures à utiliser dans la pratique. Finalement, elles ne diffèrent que par le mode de gestion des ex-aequo, elles ont un comportement très similaire (Tableau 8.1). Peut être nous bornerons nous à dire, comme tout le monde, de préférer τ_c à τ_b lorsque le tableau de contingence n'est pas carré.

Mesure	Estimation	I.Confiance à 90%	z-test	p-value
γ	0.742779	$[0.699498; 0.786059]$	20.883351	< 0.000001
τ_b	0.476570	$[0.441758; 0.511383]$	20.883351	< 0.000001
τ_c	0.432618	$[0.398543; 0.466693]$	20.883351	< 0.000001

Tableau 8.1. Comparaison des mesures ordinales pour l'association "Montant du crédit" et "Durée du crédit"

9

d de Sommers - Une mesure asymétrique

9.1 Définition et estimation

Souvent les analyses sont menées pour mettre à jour des causalités : on veut voir l'effet d'une variable sur une autre. Dans notre cas du crédit bancaire, nous voulons savoir si des montants élevés vont entraîner des durées de remboursement accrues. A l'instar des mesures pour les variables nominales où nous opposions les mesures symétriques dérivées du χ^2 aux mesures de type PRE asymétriques, nous introduisons une mesure asymétrique pour les variable ordinales : il s'agit du d de Sommers.

Le d de Sommers se définit comme une différence de probabilité, comme le γ de Goodman et Kruskal. Sauf que la normalisation s'effectue uniquement par rapport aux paires distinctes sur la variable indépendante X (Siegel, 1988 ; page 303) :

$$d = \frac{P(\text{Y et X sont concordants}) - P(\text{Y et X sont discordants})}{P(\text{paire distincte sur X})} \tag{9.1}$$

L'estimation naturelle sur un échantillon de données de taille n est

$$\hat{d} = \frac{P - Q}{D_X} \tag{9.2}$$

Il s'agit du surplus de paires concordantes. On peut le lire de la manière suivante : pour une paire d'observations prise au hasard qui ne sont pas des ex-aequo, le d de Sommers représente l'écart entre la probabilité conditionnelle d'avoir une paire concordante moins la probabilité d'avoir une paire discordante. Il atteint la valeur $+1$ pour association monotone prédictive stricte positive (voir tableau 7.1) c.-à-d. lorsque X augmente, Y augmente, et à chaque valeur de X (explicative, indépendante) correspond une et une seule valeur de Y (expliquée, dépendante). Dans notre tableau de contingence, dans chaque colonne, nous avons une seule ligne non-nulle. Les cases non-nulles s'échelonnent du coin en haut à gauche vers le coin en bas à droite.

A contrario, la valeur -1 correspond à une association négative, dans ce cas, quand X augmente, Y diminue. Lorsque les variables sont indépendantes, d est égal à 0.

Amount vs. Duration sur les données CREDIT

Grâce aux calculs précédents, nous disposons de toutes les informations pour estimer convenablement d, nous avons ainsi

$$\hat{d} = \frac{338350 - 49938}{564064} = 0.511311$$

Pour une paire d'individus (i, j) non ex-aequo sur X, lorsque $x_i > x_j$, l'écart entre la probabilité d'avoir $y_i > y_j$ et la probabilité d'avoir $y_i < y_j$ est (aux fluctuations d'échantillonnage près) égal à 0.511. La liaison semble positive et relativement conséquente. Pour en évaluer toute la portée, il faut passer par les techniques de statistique inférentielle.

9.2 Intervalle de confiance

Selon un schéma bien connu maintenant, l'intervalle de confiance au niveau $(1 - \alpha)$ s'appuie sur la normalité asymptotique. Pour compléter les calculs, nous devons produire une estimation de la variance asymptotique :

$$\sigma^2_{\hat{d}} = \frac{4}{D^4_X}\left[\sum_l \sum_c n_{lc}\left[D_X(C_{lc} - D_{lc}) - (P - Q)(n - n_{.c})\right]^2\right] \tag{9.3}$$

L'intervalle de confiance au niveau $(1 - \alpha)$ s'écrit

$$d \in \left[\hat{d} \pm u_{1-\frac{\alpha}{2}} \times \sigma_{\hat{d}}\right]$$

Amount vs. Duration sur les données CREDIT

La feuille de calcul est très similaire aux précédentes, nous retrouvons peu ou prou les quantités nécessaires à la production des estimations de la variance. Dans le cas du d de Sommers, nous l'organisons de la manière suivante (Figure 9.1) :

— Le tableau de contingence **A** nous sert surtout à obtenir les effectifs marginaux $n_{.c}$
— Dans le tableau **B**, nous formons tour à tour les quantités usuelles n_{lc}, C_{lc} et D_{lc}.
— Dans la colonne **B.5**, nous calculons $D_X(C_{lc} - D_{lc})$
— Dans **B.6**, $(P - Q)(n - n_{.c})$
— Dans **B.7**, nous formons la différence inscrite dans la double somme,
— Que nous pondérons avec les effectifs n_{lc} en **B.8**, Il ne rester plus qu'à additionner les valeurs pour obtenir la somme.

La variance et l'écart type asymptotiques sont obtenues avec

$$\sigma^2_{\hat{d}} = \frac{4}{564064^4}\left[1.3403 \cdot 10^{19}\right] = 0.000530$$
$$\sigma_{\hat{d}} = \sqrt{0.000530} = 0.023013$$

P	338350
Q	49938
D_X	564064
d	0.511311

Tableau de contingence (A)

	disc_amount		
disc_duration	lo_1000	1000_2000	up_2000
lo_1_year	93	161	105
1_2_years	22	146	243
up_2_years	1	9	220
n_{.c}	116	316	568

sigma^2	0.000530
sigma	0.023013
u(0.95)	1.644853

Intervalle de confiance

B.Basse	0.473457
B.Haute	0.549164

Calcul de la variance asymptotique (B)

Case	n_{lc}	C_{lc}	D_{lc}	B.5	B.6	B.7	B.8
(1,1)	93	618	0	348591552	254956208	8.7676E+15	8.1538E+17
(1,2)	161	463	23	248188160	197273808	2.5923E+15	4.1736E+17
(1,3)	105	0	178	-100403392	124593984	5.0624E+16	5.3155E+18
(2,1)	22	229	266	-20870368	254956208	7.6080E+16	1.6738E+18
(2,2)	146	313	106	116761248	197273808	6.4823E+15	9.4641E+17
(2,3)	243	254	10	137631616	124593984	1.6998E+14	4.1305E+16
(3,1)	1	0	655	-369461920	254956208	3.8990E+17	3.8990E+17
(3,2)	9	115	348	-131426912	197273808	1.0804E+17	9.7240E+17
(3,3)	220	422	0	238035008	124593984	1.2869E+16	2.8312E+18
						Somme	1.3403E+19

Test de significativité

sigma^2(0)	0.000599
sigma (0)	0.024484
z	20.883351
p-value	0.000000

Calcul de la variance asymptotique sous H0 (C)

Case	n_{lc}	C_{lc}	D_{lc}	C.5	C.6
(1,1)	93	618	0	381924	35518932
(1,2)	161	463	23	193600	31169600
(1,3)	105	0	178	31684	3326820
(2,1)	22	229	266	1369	30118
(2,2)	146	313	106	42849	6255954
(2,3)	243	254	10	59536	14467248
(3,1)	1	0	655	429025	429025
(3,2)	9	115	348	54289	488601
(3,3)	220	422	0	178084	39178480
				Somme	130864778

Fig. 9.1. CREDIT GERMAN - Croisement "montant du crédit" et "durée du crédit". Calcul du d de Sommers, intervalle de confiance et test de significativité

Il ne reste plus qu'à produire les bornes des intervalles de variation à 90%

$$b.basse = 0.511311 - 1.644853 \times 0.023013 = 0.473457$$
$$b.haute = 0.511311 + 1.644853 \times 0.023013 = 0.549164$$

9.3 Test de significativité

Pour tester la significativité, il faut produire une estimation de la variance en accord avec l'hypothèse d'indépendance. Elle s'écrit

$$\sigma^2_{\hat{d}}(0) = \frac{4}{D_X^2}\left[\sum_l \sum_c n_{lc}(C_{lc} - D_{lc})^2 - \frac{1}{n}(P-Q)^2\right] \tag{9.4}$$

Pour tester la significativité, nous formons la statistique réduite $z_{\hat{d}} = \frac{\hat{d}}{\sigma_{\hat{d}}(0)}$, et la région critique au risque α devient

$$R.C. : |z_{\hat{d}}| > u_{1-\frac{\alpha}{2}}$$

Amount vs. Duration sur les données CREDIT

Nous nous intéressons maintenant au tableau **C** de la feuille EXCEL (Figure 9.1) :

— Nous retrouvons classiquement les valeurs de n_{lc}, C_{lc} et D_{lc}

— En **C.5**, nous formons $(C_{lc} - D_{lc})^2$

— Puis en **C.6**, le produit dans la double somme. Le total final est 130864778

La variance et l'écart type sous l'hypothèse d'indépendance s'obtiennent avec

$$\sigma^2_{\hat{d}}(0) = \frac{4}{564064^2}\left[130864778 - \frac{1}{1000}(338350 - 49938)^2\right] = 0.000599$$
$$\sigma_{\hat{d}}(0) = \sqrt{0.000599} = 0.024484$$

La statistique réduite est

$$z_{\hat{d}} = \frac{0.511311}{0.024484} = 20.883351$$

La p-value du test est ainsi $p_c < 0.000001$, l'association est très largement significative.

9.4 Une version symétrique du d de Sommers

Il existe une version symétrique du d de Sommers, elle s'écrit

$$\hat{d}_s = \frac{P - Q}{\frac{1}{2}(D_Y + D_X)} \tag{9.5}$$

Son intérêt est limité par rapport aux mesures symétriques (τ_b notamment) présentées par ailleurs.

Amount vs. Duration sur les données CREDIT

A titre de curiosité, le coefficient d de Sommers dans le cas où nous inversons la causalité, la durée détermine le montant, est $\hat{d}_{X/Y} = 0.444$; le coefficient symétrique est $\hat{d}_s = 0.475$. Tous deux sont très fortement significatifs.

Cela relativise la pertinence des mesures asymétriques pour détecter les causalités. Il faut toujours être prudent lorsqu'on manipule des indicateurs. Néanmoins, nous remarquerons que dans les 2 cas, le coefficient est moins élevé que celui calculé dans le sens "le montant cause la durée" qui est égal à $\hat{d} = 0.511$.

10

Association ordinale pour les variables binaires (Mantel-Haenszel)

Lorsque les variables sont binaires, une particularité intéressante apparaît : le coefficient de corrélation est insensible au codage adopté. Que l'on code 0/1 ou 0/25, nous obtiendrons toujours la même valeur du coefficient de corrélation. De même, si nous codons différemment les variables X et Y (par exemple, respectivement 0/1 et 0/18), le coefficient r n'est en rien perturbé.

Cela ouvre des perspectives réjouissantes. Tout l'arsenal mis en place pour évaluer la corrélation peut être mis en oeuvre pour l'analyse des associations entre variables ordinales binaires, notamment le test de significativité basé sur le t de Student présenté plus haut dans ce support (Voir chapitre 4.1).

Un autre test est souvent cité dans la littérature, il s'agit du test du χ^2 de Mantel-Haenszel pour les associations linéaires [1]. Il permet de tester la significativité de la liaison. La statistique utilisée s'appuie sur la corrélation estimée r :

$$\chi^2_{MH} = (n-1) \times r^2 \tag{10.1}$$

Elle suit une loi du χ^2 à 1 degré de liberté sous l'hypothèse d'indépendance.

Remarque 25 (Association linéaire ?). Il ne faut pas se méprendre sur le caractère linéaire de l'association. Puisque X ne peut prendre que 2 valeurs possibles, de même pour Y, l'association mesurée ne peut être que linéaire. Ce n'est donc pas une contrainte de l'outil statistique mais plutôt une conséquence de la nature des variables étudiées.

Liaison X-Foreign Worker et Y-Telephone

Reprenons un exemple déjà largement utilisé précédemment. Les variables sont binaires. Nous avons reconstruit la feuille EXCEl "MANTEL-HAENSZEL CHI-SQ TEST" de manière à ce que le codage soit paramétrable (Figure 10.1 ; ici la seconde modalité est codée 2 pour Y, 7 pour X). Nous constatons aisément en modifiant le code de la seconde modalité, que ce soit pour la variable X ou Y, que le coefficient de corrélation reste strictement le même : $r = 0.107401$. La liaison est positive.

Nous formons la statistique du test de significativité

1. En anglais, Mantel-Haneszel chi-square, ou Mantel-Haenszel test for linear association, ou linear by linear association chi-square.

	codage	2	7
Ytelephone	Xforeign	Ytelephone	Xforeign
1	1	2	7
0	1	0	7
0	1	0	7
0	1	0	7
0	1	0	7
1	1	2	7
0	1	0	7
1	1	2	7
0	1	0	7
0	1	0	7
0	1	0	7
0	1	0	7
1	1	2	7
0	1	0	7
0	1	0	7

r	0.107401

M-H KHI-2	11.523420
p-value	0.000687

Fig. 10.1. CREDIT GERMAN - Association ordinale entre "Foreign worker" et "Telephone". Premières observations, codage paramétrable, coefficient de corrélation et test de Mantel-Haenszel

$$\chi^2_{MH} = (1000 - 1) \times 0.107401^2 = 11.523420$$

Et la p-value calculée est $p_c = 0.000867$. La liaison est significative. Bien entendu, il s'agit d'un exemple d'école, est-ce la possession ou la non-possession d'un téléphone à son nom qui est un avantage ?

Remarque 26 (Autre présentation). Puisque les variables sont binaires, il y a une relation directe entre le carré de la corrélation et le $\phi^2 = \frac{\chi^2}{n}$. L'indicateur peur s'écrire :

$$\begin{aligned} \chi^2_{MH} &= (n-1) \times \phi^2 \\ &= \frac{n-1}{n} \times \chi^2 \end{aligned}$$

11

Association partielle pour variables ordinales

TODO

Partie V

Annexes

A

Gestion des versions

Avec ce support, j'inaugure une nouvelle manière de travailler. Mes contraintes professionnelles m'interdisent d'élaborer en une seule fois ce gros chantier. Pourtant, j'estime que ce qui a été produit à ce jour (Version 1.0 – 10 Septembre 2007) peut être déjà utile aux étudiants et autres chercheurs intéressés par le sujet. Plutôt que de laisser le document en l'état et attendre une hypothétique plage de disponibilités assez large pour achever ce document, je préfère, comme pour les logiciels libres que je diffuse par ailleurs, introduire une gestion de versions. Un numéro sera donc attribué à chaque nouvelle version, elle permet de suivre dans le temps les sections, chapitres et parties rajoutés au document. Par ailleurs, la date de compilation, située dans la partie basse du texte, permet de suivre les modifications de détail (errata, modification de mise en forme, ajustement des figures, etc.).

Version 1.0

C'est la première version diffusée sur le web, mise en ligne le 10 septembre 2007.

Version 2.0

Le document comprend une nouvelle partie : "Analyse des associations pour les variables ordinales". Mise en ligne le 1er octobre 2007.

Dans la partie Annexes, une référence a été ajoutée. Il s'agit de la mise en oeuvre des mesures d'associations avec le logiciel TANAGRA (http://eric.univ-lyon2.fr/~ricco/tanagra/fr/tanagra.html, voir Annexe F).

Version 2.1

Avril 2020. Des coquilles ont été corrigées. La table de matière avec les liens apparaît maintenant dans les lecteurs de fichiers PDF et, surtout, des sections ont été ajoutées : 3.6, 4.4, 5.3.3] et . Le fichier Excel des exemples a été complété en conséquence.

B

Description des données GERMAN CREDIT

Ce jeu de données est disponible à différents endroits. Il a été utilisé dans de nombreuses évaluations tant les possibilités d'analyse sont nombreuses. Nous citerons principalement 2 (au cas où le premier venait à disparaître) URL permettant d'accéder aux données :

— http://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data) ;

— ftp://ftp.ics.uci.edu/pub/machine-learning-databases/statlog/german/

Pour notre part, le fichier au format EXCEL (**credit-german-dependance-qualitative.xlsx**) contient les données et certains traitements décrits dans ce support.

Voici le dictionnaire de données (en anglais) :

```
Description of the German credit dataset.

1. Title: German Credit data

2. Source Information

Professor Dr. Hans Hofmann
Institut f"ur Statistik und "Okonometrie
Universit"at Hamburg
FB Wirtschaftswissenschaften
Von-Melle-Park 5
2000 Hamburg 13

3. Number of Instances:  1000

Two datasets are provided.  the original dataset, in the form provided
by Prof. Hofmann, contains categorical/symbolic attributes and
is in the file "german.data".
```

```
For algorithms that need numerical attributes, Strathclyde University
produced the file "german.data-numeric".  This file has been edited
and several indicator variables added to make it suitable for
algorithms which cannot cope with categorical variables.   Several
attributes that are ordered categorical (such as attribute 17) have
been coded as integer.    This was the form used by StatLog.


6. Number of Attributes german: 20 (7 numerical, 13 categorical)
   Number of Attributes german.numer: 24 (24 numerical)


7.  Attribute description for german

Attribute 1:  (qualitative)
       Status of existing checking account
               A11 :      ... <    0 DM
       A12 : 0 <= ... <  200 DM
       A13 :      ... >= 200 DM /
     salary assignments for at least 1 year
               A14 : no checking account

Attribute 2:  (numerical)
      Duration in month

Attribute 3:  (qualitative)
      Credit history
      A30 : no credits taken/
    all credits paid back duly
              A31 : all credits at this bank paid back duly
      A32 : existing credits paid back duly till now
              A33 : delay in paying off in the past
      A34 : critical account/
    other credits existing (not at this bank)

Attribute 4:  (qualitative)
      Purpose
      A40 : car (new)
      A41 : car (used)
```

```
      A42 : furniture/equipment
      A43 : radio/television
      A44 : domestic appliances
      A45 : repairs
      A46 : education
      A47 : (vacation - does not exist?)
      A48 : retraining
      A49 : business
      A410 : others

Attribute 5:  (numerical)
      Credit amount

Attibute 6:  (qualitative)
      Savings account/bonds
      A61 :          ... <  100 DM
      A62 :   100 <= ... <  500 DM
      A63 :   500 <= ... < 1000 DM
      A64 :          .. >= 1000 DM
              A65 :   unknown/ no savings account

Attribute 7:  (qualitative)
      Present employment since
      A71 : unemployed
      A72 :       ... < 1 year
      A73 : 1  <= ... < 4 years
      A74 : 4  <= ... < 7 years
      A75 :       .. >= 7 years

Attribute 8:  (numerical)
      Installment rate in percentage of disposable income

Attribute 9:  (qualitative)
      Personal status and sex
      A91 : male   : divorced/separated
      A92 : female : divorced/separated/married
              A93 : male   : single
      A94 : male   : married/widowed
      A95 : female : single
```

```
Attribute 10: (qualitative)
      Other debtors / guarantors
      A101 : none
      A102 : co-applicant
      A103 : guarantor

Attribute 11: (numerical)
      Present residence since

Attribute 12: (qualitative)
      Property
      A121 : real estate
      A122 : if not A121 : building society savings agreement/
   life insurance
              A123 : if not A121/A122 : car or other, not in attribute 6
      A124 : unknown / no property

Attribute 13: (numerical)
      Age in years

Attribute 14: (qualitative)
      Other installment plans
      A141 : bank
      A142 : stores
      A143 : none

Attribute 15: (qualitative)
      Housing
      A151 : rent
      A152 : own
      A153 : for free

Attribute 16: (numerical)
              Number of existing credits at this bank

Attribute 17: (qualitative)
      Job
      A171 : unemployed/ unskilled  - non-resident
```

```
      A172 : unskilled - resident
      A173 : skilled employee / official
      A174 : management/ self-employed/
     highly qualified employee/ officer

Attribute 18: (numerical)
      Number of people being liable to provide maintenance for

Attribute 19: (qualitative)
      Telephone
      A191 : none
      A192 : yes, registered under the customers name

Attribute 20: (qualitative)
      foreign worker
      A201 : yes
      A202 : no

Last attribute : CLASS
      Good : the credit paid back duly
      Bad  : not
```

C

Description du classeur EXCEL - GERMAN CREDIT

Au delà des données, le classeur "credit-german-dependance-qualitative.xls" contient un certain nombre de feuilles de calcul que nous reprenons à plusieurs reprises dans le texte. Voici une énumération exhaustive de ces feuilles et les sujets qu'elles recouvrent.

1. **Données.** Cette feuille contient des données originelles.
2. **Housing vs. Job - Chi-2.** Tests et mesures fondées sur le χ^2, résidus standardisés, résidus ajustés, contributions.
3. **Housing vs. Job - Autres** Autres mesures d'association, information mutuelle.
4. **Lambda Goodman Kruskal.** Calcul du λ de Goodman et Kruskal.
5. **Tau Goodman Kruskal.** τ de Goodman et Kruskal pour les variables nominales.
6. **u de Theil** Coefficient d'incertitude de Theil.
7. **u symétrique** Coefficient d'incertitude symétrique.
8. **Table 2 x 2 - Tel x ForeignWker.** Analyse des tables 2×2. Coefficient de corrélation de Pearson, Correction de Yates, $\mathcal{Q}$ de Yule.
9. **Mc Nemar** Test de Mc Nemar. Comparaison de proportions pour échantillons appariés.
10. **Odds - Odds-ratio - RR.** Odds, odds-ratio et risque relatif dans les tableaux 2×2.
11. **Odds-ratio Tableau 2 x C.** Calcul des odds-ratios dans les tableaux à C colonnes.
12. **Kappa Cohen - 2 juges.** κ de Cohen. Mesure de concordance entre 2 juges, calculé sur un tableau de contingence.
13. **Kappa Fleiss.** κ de Fleiss. Concordance de m juges.
14. **Kappa généralisé.** κ de concordance, nombre de juges différents selon les individus.
15. **Var. Ordinales - Paires C et D.** Calcul des paires concordantes et discordantes pour le croisement des variables "Amount" et "Duration", toutes les deux discrétisées en 3 intervalles.
16. **Gamma de Goodman et Kruskal.** Calcul du γ de Goodman et Kruskal pour le croisement "Amount" vs. "Duration". Intervalle de confiance et test de significativité.
17. **Tau-b Kendall.** Calcul du τ_b de Kendall pour le croisement "Amount" vs. "Duration". Intervalle de confiance et test de significativité.

18. **Tau-c Kendall et Stuart.** Calcul du τ_c de Kendall pour le croisement "Amount" vs. "Duration". Intervalle de confiance et test de significativité.

19. **d de Sommers.** Calcul du d de Sommers pour le croisement "Amount" vs. "Duration". Intervalle de confiance et test de significativité.

20. **Mantel Haenszel chi-sq test.** Calcul de la statistique de Mantel-Haenszel pour le croisement des variables binaires "Telephon" et "Foreign". Statistique du χ^2 à partir du coefficient de corrélation.

D

Calculette en ligne (I) pour les tableaux 2×2

Une excellente référence Web recense un grand nombre de mesures utilisables pour les tableaux 2×2. Une très courte définition avec l'interprétation de la mesure est disponible (Vérifiée en Avril 2020).

La particularité de ce site est que nous disposons de surcroît d'une calculette qui nous permet de saisir les valeurs de notre tableau de contingence. Le programme fournit les indicateurs adéquats. Après vérification sur un tableur et comparaison avec les logiciels qui font foi, il apparaît que les calculs sont rigoureusement exacts.

Petit raffinement appréciable, le programme effectue les calculs dans les 2 sens du tableau "ligne x colonne" et "colonne x ligne". C'est important pour les mesures non-symétriques.

— (Courte) description des mesures – http://www.quantitativeskills.com/sisa/statistics/two2hlp.htm
— Calculette en ligne – http://www.quantitativeskills.com/sisa/statistics/twoby2.htm

E

Calculette en ligne (II) pour les tableaux 2×2

Une autre calculette, en français cette fois-ci, permet de réaliser les mêmes calculs. La seule différence est que le site est très tourné vers les études épidémiologiques, le vocabulaire utilisé est à l'avenant. Nous retrouvons néanmoins nos principaux repères et un lexique nous permet de mieux comprendre les valeurs produites.

Comme pour le site précédent, tous les calculs ont été plusieurs fois vérifiés. C'est réellement un travail remarquable.

Plusieurs calculettes sont disponibles (Vérifiées en Avril 2020) :

— Pour les associations "Facteur vs. Maladie", http://www.aly-abbara.com/utilitaires/statistiques/khi_carre_rr_odds_ratio_ic.html
— Pour l'évaluation de la valeur prédictive d'un test de diagnostic : http://www.aly-abbara.com/utilitaires/statistiques/sensibilite_specificite_vpp_vpn.html

F

Les mesures d'association dans le logiciel TANAGRA

La majorité des mesures décrites dans ce support ont été implémentées dans le logiciel *open source* TANAGRA (http://eric.univ-lyon2.fr/~ricco/tanagra/fr/tanagra.html, version 1.4.19). Plusieurs didacticiels décrivent la mise en ?uvre des techniques sur des fichiers exemples, accessibles en ligne.

— *Mesures d'association - Variables nominales*, mai 2007 ; http://tutoriels-data-mining.blogspot.com/2008/04/mesures-dassociation-variables.html ;

— *Mesures d'association - Variables ordinales*, mai 2007, http://tutoriels-data-mining.blogspot.com/2008/04/mesures-dassociation-variables_05.html

Littérature

Ouvrages

1. Agresti, A., *Categorical Data Analysis*, John Wiley and Sons, 1990.
2. Celeux, G., Nakache, J.P., *Analyse discriminante sur variables qualitatives*, Polytechnica, 1994.
3. Cohen, J., *Statistical Power Analysis for the Behavorial Sciences*, Second Edition, Routledge, 1988.
4. Croux, C., *Les coefficients d'association et les tests d'indépendance pour des variables qualitatives*, Chapitre 2 in "Modèles Statistiques pour données qualitatives", Droesbeke, Lejeune et Saporta, Éditeurs, TECHNIP, 2005.
5. Dagnelie P., *Statistique Théorique et Appliquée - Tome 2. Inférence Statistique à une et deux dimensions*, 2ème édition, 2006.
6. Howell, D., *Méthodes statistiques en sciences humaines*, De Boeck Université, 1998.
7. Millot G., *Comprendre et réaliser les tests statistiques à l'aide de R*, de boeck, 2008.
8. Rakotomalala R., *Analyse de corrélation - Étude des dépendances - Variables quantitatives*, mars 2005, http://eric.univ-lyon2.fr/~ricco/cours/ouvrages.html.
9. Rakotomalala R., *Comparaison de populations - Tests paramétriques*, juin 2013, http://eric.univ-lyon2.fr/~ricco/cours/ouvrages.html.
10. Rakotomalala R., *Pratique de la régression logistique - Régression logistique binaire et polytomique*, juin 2011, http://eric.univ-lyon2.fr/~ricco/cours/ouvrages.html.
11. Renaud, A., *Statistique épidémiologique*, Que Sais-Je ?, Presses Universitaires de France, 1986.
12. Saporta, G., *Probabilités, Statistique et Analyse des données*, Technip, 2006.
13. Scherrer B., *Biostatistique - Volume 1*, 2^{me} édition, Gaëtan Morin, 2007.
14. Siegel, S., Castellan Jr., J., *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill Inc., Second Edition, 1988.

Supports en ligne

15. Bonnardel, P., *Le coefficient Kappa*, http://kappa.chez-alice.fr/
16. Garson, D., *Measures of Association*, http://www2.chass.ncsu.edu/garson/PA765/association.htm, N/A en avril 2020.

17. Garson, D., *Nominal Association : Phi, Contingency coefficient, Tschuprow's T, Cramer's V, Lambda, Uncertainty Coefficient*, http://www2.chass.ncsu.edu/garson/PA765/assocnominal.htm, N/A en avril 2020.

18. SAS Institute Inc., Documentation SAS, *Measures of Association*, http://v8doc.sas.com/sashtml/stat/chap28/sect20.htm