PAUL-MARIE BERNARD

ANALYSE DES TABLEAUX DE CONTINGENCE EN ÉPIDÉMIOLOGIE

CÉDÉROM INCLUS

Presses de l'Université du Québec

ANALYSE DES TABLEAUX DE CONTINGENCE EN ÉPIDÉMIOLOGIE

PRESSES DE L'UNIVERSITÉ DU QUÉBEC
Le Delta I, 2875, boulevard Laurier, bureau 450
Sainte-Foy (Québec) G1V 2M2
Téléphone : (418) 657-4399 · Télécopieur : (418) 657-2096
Courriel : puq@puq.uquebec.ca · Internet : www.puq.ca

Distribution :

CANADA et autres pays

DISTRIBUTION DE LIVRES UNIVERS S.E.N.C.
845, rue Marie-Victorin, Saint-Nicolas (Québec) G7A 3S8
Téléphone : (418) 831-7474 / 1-800-859-7474 • Télécopieur : (418) 831-4021

FRANCE

DISTRIBUTION DU NOUVEAU MONDE
30, rue Gay-Lussac, 75005 Paris, France
Téléphone : 33 1 43 54 49 02
Télécopieur : 33 1 43 54 39 15

SUISSE

SERVIDIS SA
5, rue des Chaudronniers, CH-1211 Genève 3, Suisse
Téléphone : 022 960 95 25
Télécopieur : 022 776 35 27

ANALYSE DES TABLEAUX DE CONTINGENCE EN ÉPIDÉMIOLOGIE

PAUL-MARIE BERNARD

2004

Presses de l'Université du Québec
Le Delta I, 2875, boul. Laurier, bur. 450
Sainte-Foy (Québec) Canada G1V 2M2

Catalogage avant publication de la Bibliothèque et Archives Canada

Bernard, Paul-Marie, 1942- .

Analyse des tableaux de contingence en épidémiologie

Comprend des réf. bibliogr. et un index.

ISBN 2-7605-1306-8

1. Tableaux de contingence. 2. Épidémiologie – Méthodes statistiques. 3. Épidémiologie – Méthodologie. 4. Biométrie. I. Titre.

RA652.2.M3B46 2004 614.4'072 C2004-940845-3

Nous reconnaissons l'aide financière du gouvernement du Canada par l'entremise du Programme d'aide au développement de l'industrie de l'édition (PADIÉ) pour nos activités d'édition.

Mise en pages : INFO 1000 MOTS INC.

Couverture : RICHARD HODGSON

1 2 3 4 5 6 7 8 9 PUQ 2004 9 8 7 6 5 4 3 2 **1**

Dépôt légal – 3e trimestre 2004
Bibliothèque nationale du Québec / Bibliothèque nationale du Canada
Imprimé au Canada

DÉDICACE ET REMERCIEMENTS

Je dédie ce volume d'abord à mon épouse, Louise, fidèle et amoureuse compagne de ma vie. Je veux la remercier pour l'indéfectible soutien qu'elle m'a apporté tout au long de la rédaction de cet ouvrage.

Je dédie aussi ce volume à toute ma petite famille :

à mes enfants, Nicolas, Geneviève et Pierre-Olivier ;

à ma belle-fille, Christine, et à mon gendre, Marc ;

à mes petits-enfants, Rébecca, Rachel, Étienne, Alice et Marianne.

Je dédie enfin ce volume à la mémoire d'un grand ami, Nérée Bujold, décédé accidentellement et prématurément en mars 2003.

Je veux remercier tous les étudiants et étudiantes qui, dans mes différents cours de biostatistique, ont stimulé ma réflexion par leur goût d'apprendre. Leurs questions, commentaires et encouragements m'ont été fort utiles pour la conception et la réalisation de ce volume. Je remercie plus particulièrement Arsène Ahoyo pour la lecture critique et soignée qu'il a faite de la première version de cet ouvrage.

Enfin, mes remerciements vont à l'équipe des Presses de l'Université du Québec pour leur accueil et leur aide à l'édition.

SIGLES ET SYMBOLES

Sigle	Symbole	Signification
ARC	arcsin $\sqrt{\ }$	Transformation du sinus de la racine carrée
DP	Δ	Différence de proportions ; Δ est la valeur théorique
DR		Différence des risques, synonyme de *DP*
DT	Δ	Différence de taux ; Δ est la valeur théorique
FA_1		Fraction attribuable chez les exposés
FA ou FA_t		Fraction attribuable de population ou totale
FP_1		Fraction prévenue chez les exposés
FP ou PR_t		Fraction prévenue de population ou totale
FV	–	Fonction de vraisemblance
MH	–	Mantel-Haenszel
RAC	$\sqrt{\ }$	Transformation racine carrée
RC	ψ	Rapport de cotes ; ψ est la valeur théorique
RP	ξ	Rapport de proportions ; ξ est la valeur théorique
RR		Risque relatif, synonyme de *RP*
RT	φ	Rapport de taux ; φ est la valeur théorique
RV		Rapport de vraisemblance
SMR (taux)	φ	*Standardized mortality ratio ;* φ est la valeur théorique
SMR (proportions)	ϕ	*Standardized mortality ratio ;* ϕ est la valeur théorique
	λ	Paramètre d'une loi de Poisson
	π	Proportion théorique (l'un des paramètres de la loi binomiale)
	τ	Taux théorique
	χ^2(assoc)	Khi-carré d'association
	χ^2(homog)	Khi-carré d'homogénéité
	χ^2(res)	Khi-carré résiduel
	χ^2(tend)	Khi-carré de tendance
	χ^2(total)	Khi-carré total

AVANT-PROPOS

Ce volume est le résultat de plusieurs années d'enseignement de la biostatistique dans le cadre de programmes de formation en recherche épidémiologique. Au cours des deux dernières décennies, le développement des méthodes d'analyse de données épidémiologiques a connu un essor important lié en grande partie à la puissance accrue des ordinateurs et aux raffinements apportés aux logiciels statistiques. Influencé à la fois par un contact assidu avec la recherche épidémiologique et par la progression constante de la performance informatique, j'ai voulu réunir dans un seul volume la description de différentes méthodes d'analyse de données épidémiologiques présentées sous forme de tableaux de contingence.

Certaines dimensions caractérisent la présentation du volume. D'abord, la présentation des méthodes se profile suivant les

mesures les plus utilisées en épidémiologie. De ce point de vue, le volume se trouve en lien important avec cette discipline et aussi avec un ouvrage déjà publié, en collaboration, sur le sujet : *Mesures statistiques en épidémiologie*[1]. Deux autres dimensions, plus statistiques celles-là, caractérisent aussi la présentation. Les méthodes statistiques présentées se concentrent essentiellement sur les deux grands outils d'analyse inférentielle : le test statistique et l'intervalle de confiance. Ces outils sont décrits, quand c'est possible, suivant trois approches de calcul : les méthodes exactes, les méthodes en approximation normale et les méthodes basées sur le maximum de vraisemblance. Enfin, une dernière dimension a marqué la présentation : la plupart des exemples numériques accompagnant la présentation des méthodes statistiques mettent à contribution le logiciel SAS, ce qui permet au lecteur de s'initier à ce type d'instrument devenu indispensable aux calculs statistiques.

Concrètement, ce volume se divise en 15 chapitres. Les deux premiers présentent les concepts nécessaires à l'inférence statistique. Dans le chapitre 1 sont décrites les lois de distribution les plus courantes en épidémiologie et, dans le 2, sont rappelés quelques grands concepts statistiques de base. L'intégration de ces notions facilite beaucoup la compréhension des différentes méthodes d'analyse présentées dans les chapitres suivants. Les méthodes d'analyse pour les mesures de fréquences de base sont présentées aux chapitres 3 pour les taux et 4 pour les proportions. Les analyses pertinentes aux mesures d'association et d'impact dans le contexte simplifié d'un tableau 2×2, le sont aux chapitres 5, 6 et 7. Les chapitres 9, 10 et 11 reprennent ces mêmes analyses, cette fois dans le contexte du contrôle d'une tierce variable, les techniques pertinentes à ce genre d'analyse ayant au préalable été exposées dans le chapitre 8. Les chapitres 12 et 13 traitent du problème de la comparaison, respectivement de plusieurs taux et de plusieurs proportions. Enfin, les deux derniers chapitres s'intéressent à l'analyse dans le contexte d'appariement, le 14 pour les études cas-témoins et le 15 pour les études de cohortes.

Avec le volume est fourni un cédérom contenant, pour l'ensemble des exercices numériques, les différents programmes informatiques utilisés, cités dans le texte à l'aide du sigle Pr suivi du numéro du chapitre et du rang du programme. Ce disque renferme aussi des indications sur l'utilisation de ces programmes, une présentation sommaire des principales fonctions ou procédures de SAS qui entrent en jeu, certains exercices supplémentaires et leurs solutions, quelques petites bases de

1. Bernard, P.-M, et C. Lapointe, *Mesures statistiques en épidémiologie*, Sainte-Foy, Presses de l'Université du Québec, 1987.

données. Un fichier « Lisez-moi » guide le lecteur dans l'utilisation des différentes composantes du disque. Ce disque n'est pas indispensable à la lecture ou à l'utilisation du manuel. Cependant, il s'avère un prolongement intéressant, voire utile au lecteur qui désire appliquer concrètement les différents outils d'analyse élaborés à l'aide du logiciel SAS.

Ce volume se veut à la fois conservateur et innovateur. Il est conservateur puisqu'il reprend la plupart des outils statistiques déjà connus, rafraîchis dans le contexte épidémiologique : le test du khi-carré de Pearson, le test de Mantel-Haenszel, les tests de tendance d'Armitage-Cochran et de Mantel, le test de McNemar, etc. Il est innovateur en ce qu'il met beaucoup d'emphase sur les méthodes de calcul exactes et du maximum de vraisemblance, qu'il propose certaines extensions d'approches déjà existantes, qu'il ose parfois de nouvelles approches d'utilisation plus simple que certaines approches conventionnelles. Ainsi, on présente de façon systématique les tests et intervalles de confiance par la méthode du rapport de vraisemblance, et quand c'est possible sur le plan informatique, les approches exactes correspondantes. On a étendu aux proportions la définition du *SMR* et les méthodes d'analyse qui lui sont liées ; on a étendu aux taux les techniques d'analyse d'une tendance sur les proportions ; on a étendu au rapport de proportions la définition du test de McNemar en analyse appariée. Enfin, on ose de nouvelles méthodes, de nature conditionnelle, pour l'analyse en approche exacte des mesures d'association dans un tableau simple, ou pour une analyse simplifiée des mesures d'impact. On propose aussi une méthode simple de calcul de la variance pour une mesure pondérée ayant été soumise à une transformation continue. Cette méthode, dans certaines conditions, permet de simplifier le calcul de la variance et donc de l'intervalle de confiance pour un grand nombre de mesures pondérées.

Ce volume se situe quelque part entre un ouvrage d'introduction et un ouvrage spécialisé. Le caractère multidimensionnel de sa structure le rend accessible et utile à un public varié, ayant déjà été initié aux concepts de base de la statistique et de l'épidémiologie. Sauf quelques rares exceptions, les méthodes sont présentées dans un langage tout à fait accessible, sans trop de démonstrations. Les présentations s'accompagnent systématiquement d'exemples numériques. Ce volume pourrait être utilisé pour un cours de base en biostatistique destiné aux étudiants du 2e cycle universitaire en épidémiologie ou en santé publique. Il pourrait aussi être utile à l'étudiant de 3e cycle, à l'assistant de recherche ou au chercheur en épidémiologie ou en santé publique, qui ont souvent besoin de reconnaître rapidement, pour une mesure donnée, la description d'un test ou d'un intervalle de confiance et les méthodes de calcul qu'il suppose. Sans verser dans la modélisation, ce volume présente de façon périphérique certaines

techniques qui lui sont reliées en vue d'arriver à des solutions simples. On peut dire que le contenu de ce volume précède la modélisation, mais aussi, en ouvrant les perspectives de cette voie, qu'il y conduit inéluctablement.

TABLE DES MATIÈRES

PARTIE 1

CONCEPTS DE BASE

CHAPITRE 1

PRINCIPALES LOIS DE DISTRIBUTION DES PROBABILITÉS

En recherche épidémiologique, clinique ou appliquée à la santé, la plupart des données ou mesures soumises aux analyses statistiques sont convenablement décrites par l'une des trois lois discrètes suivantes : la loi binomiale, la loi de Poisson et la loi hypergéométrique. D'ailleurs, les principales mesures des fréquences en épidémiologie correspondent assez naturellement à certains paramètres de ces lois : la proportion et la loi binomiale ; le taux et la loi de Poisson ; la cote et la loi hypergéométrique. Ces lois convergent toutes en approximation vers deux lois continues : la loi normale et la loi du khi-carré (χ^2). Cette dernière, en lien direct avec la loi normale, est très utilisée pour les tests d'hypothèse dans les tableaux de contingence.

Après le rappel de quelques concepts de base, nous allons décrire chacune de ces lois et les relations qui les unissent. Nous serons alors convenablement outillés pour définir les différents tests ou autres instruments statistiques propres aux analyses de données dans les tableaux de contingence.

1.1 CONCEPTS DE BASE

1.1.1 EXPÉRIENCE ALÉATOIRE, VARIABLE ALÉATOIRE ET PROBABILITÉ

L'expérience aléatoire est caractérisée par les conditions suivantes : 1) l'expérience peut générer divers résultats a priori bien définis et mutuellement exclusifs ; 2) on ne peut pas prévoir avec certitude le résultat de l'expérience, mais ce résultat est clairement identifiable ; 3) chaque résultat possible est caractérisé par son degré de vraisemblance, c'est-à-dire par sa probabilité.

Par exemple, au lancer d'un dé, le joueur ne sait prédire le résultat. Sera-ce un 4, un 6 ou une autre face ? Chaque face du dé est un résultat possible. Suite à l'expérience, la face observée résulte du seul choix du hasard, basé sur le degré de vraisemblance.

Si le dé est bien équilibré, la face 1 apparaîtra en moyenne une fois sur six, la face 2 en moyenne une fois sur six, la face 3 en moyenne une fois sur six, etc. On peut dire alors que la probabilité d'observer 1 est de 1/6, celle d'observer 2, également de 1/6, et ainsi de suite. On écrira $P(1) = 1/6$, $P(2) = 1/6$, etc. La seule certitude est que toute expérience conduira à un quelconque résultat ou, pour dire autrement, qu'aucune expérience ne conduira à aucun résultat. L'événement « un quelconque résultat » est un événement qui se produit à coup sûr. Cette certitude se traduit en la probabilité maximale de valeur 1. Si on désigne par Ω l'ensemble des événements possibles au lancer d'un dé, $\Omega = \{1,2,3,4,5,6\}$, alors on peut dire que l'événement Ω se réalise si un des événements qui le composent se réalise. L'événement Ω se réalise donc à coup sûr ; sa probabilité n'est rien d'autre que 1 : $P(\Omega) = 1$. L'événement « aucun résultat » ne se produit jamais. Cet événement « nul », désigné par $\varnothing$, traduit une impossibilité et a comme probabilité 0 : $P(\varnothing) = 0$. À tout autre résultat possible, mais caractérisé par l'incertitude, correspond une probabilité comprise entre 0 et 1, proportionnelle à son degré de vraisemblance. Si A désigne un tel événement, on a alors $0 \leq P(A) \leq 1$.

À une expérience aléatoire bien définie, on associe de façon formelle une variable dite aléatoire. Cette variable est une règle de correspondance entre l'ensemble des résultats possibles de l'expérience et des valeurs alphanumériques ou, le plus souvent, numériques. Si S désigne un résultat quelconque et R la règle de correspondance, alors $X = R(S)$ définit la variable aléatoire X. Par exemple, pour le lancer d'un dé, on pourrait définir la variable aléatoire X par $X = 1, 2, \ldots, 6$ pour les résultats possibles « la face du dé est 1 », « la face du dé est 2 », …, « la face du dé est 6 ». Pour le lancer d'une pièce de monnaie, on pourrait définir la variable aléatoire X telle que $X = 0$ pour « face » et $X = 1$ pour « pile ».

Suivant la nature même de l'expérience, la variable aléatoire est continue ou discrète. Elle est continue si ses valeurs peuvent être n'importe quelle quantité dans un certain intervalle numérique réel. Cela veut dire que toute valeur entre deux valeurs observables quelconques de la variable est théoriquement observable. Elle est discrète si ses valeurs sont en nombre fini (au plus dénombrable), assimilables à un sous-ensemble fini ou infini de l'ensemble des nombres naturels. Une variable discrète peut être de nature quantitative (valeur obtenue par dénombrement) ou qualitative (valeur correspondant à une caractéristique).

1.1.2 DISTRIBUTION DES PROBABILITÉS POUR UNE VARIABLE DISCRÈTE *X*

Considérons une variable aléatoire X discrète. Elle pourrait correspondre à l'expérience toute simple du lancer d'une pièce de monnaie (pile ou face) ou à celle du lancer d'un dé à six faces, vue précédemment. La distribution de probabilités de la variable X renvoie à la description des probabilités P des différentes valeurs de la variable X. Par exemple, pour l'expérience du lancer d'une pièce de monnaie bien équilibrée, la variable aléatoire X a pour distribution de probabilités pour (1/2, 1/2) ; pour l'expérience du dé bien équilibré, la distribution est (1/6, 1/6, …, 1/6).

TABLEAU 1.1

Résultat du lancer de la pièce de monnaie bien équilibrée	Valeur de X	Probabilité $P(X = x)$
Pile	$X = 1$	½
Face	$X = 0$	½

Pour cette première expérience, on écrira : $P(X = 1) = 1/2$ et $P(X = 0) = 1/2$.

On observe que $P(X = 1) + P(X = 0) = 1$.

TABLEAU 1.2

Résultat du lancer du dé bien équilibré	Valeur de X	Probabilité $P(X = x)$
1	$X = 1$	1/6
2	$X = 2$	1/6
…	…	…
6	$X = 6$	1/6

Pour cette deuxième expérience, on écrira : $P(X = 1) = 1/6$, $P(X = 2) = 1/6$, … $P(X = 6) = 1/6$.

On observe aussi que $P(X = 1) + P(X = 2) + \ldots + P(X = 6) = 1$.

Ainsi, il existe une règle ou fonction de probabilité (loi de probabilité) qui nous permet de calculer $P(X = x)$. La plupart du temps, cette fonction n'est pas aussi simple ou directe que celles qu'on a déduites dans les deux expériences décrites ci-dessus. Par exemple, pour une variable binomiale, on retrouve une expression comme : $P(X = x) = C_n^x p^x (1 - p)^{n-x}$. En général, la fonction qui décrit $P(X = x)$ est dite fonction de masse ou de probabilité de X.

On a évidemment la propriété suivante : $\sum_x P(X = x) = 1$.

Distribution des fréquences

Aussi beaux que puissent être ces calculs de probabilités, il se peut très bien que les données observées dans le cadre d'expériences sur la variable aléatoire X aient un comportement différent de celui que prévoit le modèle ou la loi. Il se peut aussi que le chercheur décide d'analyser les données pour savoir jusqu'à quel point le modèle probabiliste postulé correspond à la réalité.

Par exemple, on peut supposer que la pièce de monnaie est bien équilibrée : $P(X = 1) = 1/2$. Mais, après 100 lancers de cette pièce (dans une expérience tout à fait honnête), on pourrait très bien observer 46 faces et 54 piles. Est-ce que cette observation est compatible avec le modèle d'une pièce de monnaie bien équilibrée ? Les calculs conduisent à une valeur $P(\text{nombre de piles} \geq 54) = 0{,}2421$: qu'en pensez-vous ? Votre jugement changerait-il si vous observiez 540 piles dans une expérience comprenant 1000 lancers de cette même pièce de monnaie ? La probabilité $P(\text{nombre de piles} \geq 540)$ est alors de 0,0062.

Ces deux expériences, de 100 et de 1000 lancers, génèrent deux distributions de fréquences qui peuvent apparaître plus ou moins compatibles avec le modèle d'une pièce de monnaie bien équilibrée.

L'intérêt des modèles ou lois de distribution des probabilités se fonde sur la loi des grands nombres. En effet, si, après n essais d'une certaine expérience, on observe n_A fois l'événement (ou résultat) A, alors la fré-

quence relative $\frac{n_A}{n}$ de cet événement *A*, au cours des essais, tend vers la vraie valeur *p* de sa probabilité quand *n* tend vers l'infini. En d'autres termes, les fréquences relatives tendent vers des probabilités et, conséquemment, les distributions de fréquences tendent vers des distributions de probabilités lorsque la taille de l'échantillon devient très grande. On écrit : $\lim_{n\to\infty} \frac{n_A}{n} = \pi$.

MOYENNE ET VARIANCE D'UNE VARIABLE ALÉATOIRE DISCRÈTE QUANTITATIVE

Pour décrire une variable aléatoire *X*, on s'intéresse généralement à deux paramètres importants : la moyenne μ et la variance σ^2, qui permettent d'établir respectivement la position (ou tendance centrale) des valeurs de *X* et leur dispersion autour de μ. De façon générale, ces mesures se définissent et se calculent comme suit :

$$\mu = \sum_x xP_x$$

$$\sigma^2 = \sum_x (x-\mu)^2 P_x$$

où P_x équivaut à $P(X = x)$

Par exemple, pour la variable aléatoire *X* correspondant à l'expérience simple du lancer d'une pièce de monnaie bien équilibrée, on a :

$\mu = (0 \times ½) + (1 \times ½) = ½$

$\sigma^2 = [(0 - ½)^2 \times ½] + [(1 - ½)^2 \times ½] = ¼$

1.1.3 DISTRIBUTION OU FONCTION DE DENSITÉ POUR UNE VARIABLE *X* CONTINUE

Quand la variable aléatoire *X* considérée est une variable continue, comme le poids, la taille, la tension artérielle, la fonction de probabilité $P(X = x)$ est identiquement nulle. En d'autres termes, il est impossible d'observer pour une telle variable une valeur prescrite à l'avance. La probabilité n'a de sens que si on la traduit en terme d'intervalle. S'il est impossible d'observer une personne de taille 171,33333333… centimètres, il devient tout à fait concevable d'en observer une dont la taille se situera quelque part entre 171 et 172 centimètres : $P(171 \leq X \leq 172)$. Plus généralement, on s'intéressera non pas à la probabilité $P(X = x)$ mais à la probabilité que *X* se retrouve dans un voisinage de *x* :

$$P\left(x-\Delta x<X<x+\Delta x\right)=\frac{n_x}{n}$$

Si $\Delta x = 0$, nous retrouvons $P(X = x)$ ce qui ne solutionne pas notre problème. Nous allons le contourner de la façon suivante : on rapporte la probabilité à la taille Δx de l'intervalle, en faisant tendre ce Δx vers 0 sans toutefois l'atteindre. Ce rapport définit ce que nous appelons la densité de probabilité $f(x)$ de X :

$$\lim_{\Delta x\to 0}\frac{P\left(x-\Delta x<X<x+\Delta x\right)}{\Delta x}=\lim_{\substack{n\to\infty\\ \Delta x\to 0}}\left(\frac{n_x}{n}\right).\frac{1}{\Delta x}=f(x)$$

On parle ainsi, non plus de fonction de probabilités, mais de fonction de densité de probabilité. En fait, la fonction $f(x)$ de densité de X représente la probabilité que X se retrouve dans un voisinage infinitésimal de la valeur x, relativement à la dimension Δx de ce voisinage. Ainsi, suivant cette convention, la probabilité attachée à un voisinage infinitésimal de dimension dx entourant x est égale au produit de la densité $f(x)$ en x par la dimension dx du voisinage :

$$P(x-dx<X<x+dx)=f(x)dx$$

Et de façon générale, la probabilité $P(x_1 < X < x_2)$ peut alors être définie comme :

$$P(x_1<X<x_2)=\int_{x_1}^{x_2}f(x)dx$$

On a évidemment la propriété suivante : $\int_x f(x)dx=1$

Moyenne et variance d'une variable aléatoire continue

La moyenne μ et la variance σ^2 d'une variable aléatoire X continue d'une fonction de densité $f(x)$ sont définies comme :

$$\mu=\int_x xf(x)dx$$

$$\sigma^2=\int_x\left(x-\mu\right)^2 f(x)dx$$

La fonction de densité la plus célèbre est sans doute la fonction normale. Nous la présenterons dans une section ultérieure.

Histogramme et polygone de fréquences

Pour mieux visualiser le lien entre une fonction de densité $f(x)$ et une distribution de fréquences pour une variable aléatoire continue X, nous utilisons la représentation graphique de l'une et de l'autre. La distribution de fréquences d'une variable continue X est graphiquement représentée par l'histogramme ou le polygone de fréquences qui lui est rattaché.

L'histogramme est un mode de représentation graphique utile pour les distributions de fréquences d'une variable quantitative. Il est constitué d'une suite de rectangles contigus suivant les classes de la variable. Pour la classe x, la base B du rectangle correspond à l'intervalle Δx de la classe et la hauteur H à la fréquence relative rapportée à la base, de sorte que l'aire du rectangle mesure la fréquence relative de la classe x :

$$B = \Delta x \text{ et } H = \frac{n_x}{n} \times \frac{1}{\Delta x} \Rightarrow B \times H = \frac{n_x}{n}$$

Le polygone de fréquences est un graphique utile aussi pour décrire les distributions de fréquences d'une variable quantitative. Il peut être construit à partir d'un histogramme, comme une ligne brisée rejoignant les milieux des sommets des rectangles de l'histogramme. Si on augmente indéfiniment le nombre n d'observations tout en réduisant l'intervalle Δx des classes x, le polygone se transforme progressivement en une courbe lisse qui tient lieu de modèle de la distribution.

1.2 LOI BINOMIALE

1.2.1 ÉPREUVE OU ESSAI DE BERNOULLI

L'expérience aléatoire dont l'ensemble fondamental se réduit à deux événements : *succès, échec*, est appelée *essai de Bernoulli*. À cette expérience, on fait correspondre une variable aléatoire U telle que :

$U = 0$ pour un échec,

$U = 1$ pour un succès.

On désigne par π la probabilité que se produise un succès, et par $(1 - \pi)$ celle que se produise un échec : $P(U = 1) = \pi$ et $P(U = 0) = 1 - \pi$.

Comme exemple d'essai de Bernoulli, considérons l'expérience du lancer d'une pièce de monnaie. Deux événements sont alors possibles : *observer pile*, *observer face*. À cette expérience, on assigne une variable aléatoire *U* qui prendra la valeur 1 pour l'événement qui nous intéresse (disons *observer pile*) et 0 pour l'autre événement (*observer face*).

1.2.2 PROCESSUS DE BERNOULLI

L'expérience aléatoire constituée d'une suite d'essais de Bernoulli indépendants est dite *processus de Bernoulli* si les conditions suivantes sont satisfaites :

1. chaque essai ne peut conduire qu'aux deux mêmes résultats possibles : succès, échec ;
2. les essais sont indépendants, c'est-à-dire que la probabilité π_i d'obtenir un succès au i^e essai ne dépend pas des résultats des autres essais ;
3. à travers tous les essais *i*, les succès (donc les échecs) sont équiprobables, c'est-à-dire que tous les π_i sont égaux à une même valeur π.

L'expérience de *n* lancers d'une même pièce de monnaie, dans les mêmes conditions, peut être associée à un processus de Bernoulli :

1. chaque lancer ne peut conduire qu'à l'un des deux résultats possibles : pile ou face ;
2. le résultat d'un lancer n'est aucunement influencé par les résultats précédents ;
3. enfin, chaque essai conserve toujours la même probabilité π de l'événement « pile » ou $(1 - \pi)$ de l'événement « face ».

Au processus de Bernoulli, on peut associer certaines variables aléatoires issues de questions spécifiques.

- Combien faut-il d'essais pour obtenir *r* succès ? Le nombre *X* d'essais qu'il faut pour obtenir *r* succès est une variable aléatoire qui peut prendre toutes les valeurs entières $\geq r$. Cette variable obéit à une loi de Pascal (ou binomiale négative) :

  $$P(X = x) = C_{x-1}^{r-1} \pi^r (1-\pi)^{x-r}$$

- Combien faut-il d'essais pour obtenir 1 succès ? Cas particulier de la situation précédente où $r = 1$, ce nombre X d'essais est aussi une variable aléatoire. Elle obéit à une loi géométrique :
  $P(X = x) = \pi(1-\pi)^{x-1}$
- Pour un nombre fixé de n essais, combien va-t-on obtenir de succès ? Ce nombre X de succès est une variable aléatoire dont les valeurs se situent entre 0 et n : $0 \leq X \leq n$. Cette variable X obéit à une loi de distribution binomiale, notée Bin(π,n) où π désigne la probabilité d'un succès et n le nombre d'essais. On écrira $X \mapsto \text{Bin}(\pi,n)$.

Nous nous intéressons ici plus spécifiquement à cette dernière : la loi binomiale.

1.2.3 DESCRIPTION DE LA LOI BINOMIALE

Supposons que l'on veuille connaître la probabilité d'observer 2 piles dans 5 lancers d'une pièce de monnaie. On désigne par π_i la probabilité d'observer pile au i^e lancer et donc par $(1 - \pi_i)$ celle d'observer face à ce même lancer.

Pour résoudre ce problème, on fait d'abord quelques suppositions conformes aux conditions d'un processus de Bernoulli :

1. le lancer de la pièce conduit obligatoirement à l'observation de pile ($U = 1$) ou de face ($U = 0$) ; il n'existe aucune situation où le résultat puisse être ambigu ;
2. pour un essai i, la probabilité d'un résultat (pile ou face) n'est influencée par celui d'aucun autre essai ;
3. la probabilité d'observer pile, $P(U = 1)$, est bien déterminée et constante à travers les essais : $\pi_i = \pi$ pour tout i.

Alors on décrit de la façon suivante la solution du problème.

On représente tous les types d'expériences de 5 lancers qui conduisent à 2 piles. Pour chacune de ces expériences, on calcule la probabilité d'observer une telle répartition des résultats : 2 piles, 3 faces (tableau 1.3).

TABLEAU 1.3

Type d'expérience	Résultat de *U* suivant l'ordre des lancers					Probabilité
	1er	2e	3e	4e	5e	
I	1	1	0	0	0	$\pi^2(1-\pi)^3$
II	1	0	1	0	0	$\pi^2(1-\pi)^3$
III	1	0	0	1	0	$\pi^2(1-\pi)^3$
IV	1	0	0	0	1	$\pi^2(1-\pi)^3$
V	0	1	1	0	0	$\pi^2(1-\pi)^3$
VI	0	1	0	1	0	$\pi^2(1-\pi)^3$
VII	0	1	0	0	1	$\pi^2(1-\pi)^3$
VIII	0	0	1	1	0	$\pi^2(1-\pi)^3$
IX	0	0	1	0	1	$\pi^2(1-\pi)^3$
X	0	0	0	1	1	$\pi^2(1-\pi)^3$

Ainsi, on remarque que les expériences conduisant à 2 piles peuvent se regrouper suivant 10 types d'événements dont la probabilité est toujours égale à $\pi^2(1-\pi)^3$.

On déduit alors que la probabilité $P(2 \text{ piles} \mid 5 \text{ lancers})$ d'observer 2 piles dans 5 lancers d'une pièce de monnaie est de $10\pi^2(1-\pi)^3$.

Le nombre 10 correspond au nombre d'expériences de 5 lancers conduisant à 2 piles, soit au nombre de sous-ensembles de 2 éléments qui peuvent être formés à partir d'un ensemble de 5 éléments :

$$C_5^2 = \frac{5!}{2!(5-2)!} = 10$$

Si $\pi = 0{,}40$ et $(1-\pi) = 0{,}60$, alors la valeur numérique de cette probabilité est de $10 \times (0{,}16) \times (0{,}216)$, soit 0,3456.

Nous pouvons aussi calculer les probabilités d'observer 0, 1, 3, 4 et 5 piles. Sans en détailler les calculs, nous présentons les résultats au tableau 1.4.

TABLEAU 1.4

Probabilité		Expression		Valeur
P(0 pile | 5 lancers)	=	$1\pi^0(1-\pi)^5$	=	0,07776
P(1 pile | 5 lancers)	=	$5\pi^1(1-\pi)^4$	=	0,25920
P(2 piles | 5 lancers)	=	$10\pi^2(1-\pi)^3$	=	0,34560
P(3 piles | 5 lancers)	=	$10\pi^3(1-\pi)^2$	=	0,23040
P(4 piles | 5 lancers)	=	$5\pi^4(1-\pi)^1$	=	0,07680
P(5 piles | 5 lancers)	=	$1\pi^5(1-\pi)^0$	=	0,01024
Total des probabilités	=	1	=	1,00

On remarque que la somme des probabilités couvrant toutes les éventualités est précisément égale à 1. On remarque aussi que chaque expression de ces différentes probabilités correspond à un terme du développement du binôme $[\pi + (1 - \pi)]^5$. Dans ce développement, l'exposant de π passe de 0 à 5 alors que celui de $(1 - \pi)$ passe de 5 à 0 ; les coefficients des différents termes sont 1, 5, 10, 10, 5 et 1 :

$$[\pi+(1-\pi)] = \pi^0(1-\pi)^5+5\pi^1(1-\pi)^4+10\pi^2(1-\pi)^3+ 10\pi^3(1-\pi)^2+5\pi^4(1-\pi)^1+\pi^5(1-\pi)^0$$

Généralisons le problème. Si X représente la variable correspondant au nombre de piles pour n lancers d'une même pièce de monnaie, alors la probabilité que cette variable X soit précisément égale à x (une valeur particulière de X) est donnée par : $P(X = x) = C_n^x \pi^x (1-\pi)^x$.

Les coefficients C_n^x peuvent être décrits à l'aide du célèbre triangle de Pascal :

n	Coefficients
0	1
1	1 1
2	1 2 1
3	1 3 3 1
4	1 4 6 4 1
5	1 5 10 10 5 1
6	1 6 15 20 15 6 1
7	1 7 21 35 35 21 7 1
…	…

1.2.4 PROPRIÉTÉS DE LA LOI BINOMIALE

Considérons la variable binomiale X telle que $X \mapsto \text{Bin}(\pi,n)$.

1. La distribution de X est caractérisée par les deux paramètres : π et n. Le premier, π, décrit la probabilité d'un succès dans l'expérience de base de Bernoulli et n le nombre d'essais de Bernoulli indépendants.
2. La moyenne $E(X)$ de cette distribution est $n\pi$: $\sum x_i P(x_i) = n\pi$.
3. La variance $V(X)$ de cette distribution est $n\pi(1-\pi)$: $\sum P(x_i)\left[x_i - E(X)\right]^2 = n\pi(1-\pi)$.
4. La proportion p ($= x/n$) calculée sur un échantillon de n essais de Bernoulli est une estimation de π. La moyenne ou la valeur attendue $E(p)$ et la variance $V(p)$ de p sont respectivement π et $\dfrac{\pi(1-\pi)}{n}$.
5. Puisque la variance $V(X)$ de la variable X est liée directement à la valeur de sa moyenne $E(X)$, il peut être intéressant de déterminer une transformation de X qui stabilise sa variance. L'arc sinus de la racine carrée de X est une telle transformation. : $\arcsin\sqrt{X}$. Dans le texte, on désignera par ARC cette transformation. On peut montrer que $V\left(\arcsin\sqrt{X}\right) = \dfrac{1}{4}$ si la valeur du sinus est traitée en radians. En conséquence, pour une proportion p, la variance de ARC(p) est donnée par $V\left(\arcsin\sqrt{p}\right) = \dfrac{1}{4n}$.

EXEMPLE 1.1

Au lancer d'une pièce de monnaie, la probabilité d'observer pile (un succès) est de 1/2. Si on lance 10 fois cette pièce de monnaie, quelle est la probabilité d'observer exactement 5 piles ($X = 5$) ?

Cette probabilité est calculée comme :

$$\begin{aligned} P(X = 5 \mid \pi = 0,5,\ \text{n} = 10) &= C_{10}^{5}\pi^5(1-\pi)^5 \\ &= \frac{10}{5!(10-5)!}(1/2)^5(1-1/2)^5 \\ &= 0,2461 \end{aligned}$$

On présente ici les probabilités correspondant aux différentes valeurs de X :

$P(X = 0) = 0{,}001$	$P(X = 6\ ) = 0{,}205$
$P(X = 1) = 0{,}010$	$P(X = 7\ ) = 0{,}117$
$P(X = 2) = 0{,}044$	$P(X = 8\ ) = 0{,}044$
$P(X = 3) = 0{,}117$	$P(X = 9\ ) = 0{,}010$
$P(X = 4) = 0{,}205$	$P(X = 10) = 0{,}001$
$P(X = 5) = 0{,}246$	

◆

EXEMPLE 1.2

Dans une population donnée, dix pour cent (10 %) des individus ont la caractéristique M (groupe sanguin B par exemple). On tire de cette population un échantillon de 100 personnes, sans remise. Quelle est la probabilité que l'échantillon contienne exactement 10 personnes ayant la caractéristique M ?

Pour résoudre ce problème dans le cadre de la loi binomiale, nous allons supposer que la population est grande, voire infinie, et que l'échantillon est négligeable par rapport à cette population. Ce contexte assure une certaine stabilité de la proportion π d'un tirage à l'autre. L'expérience s'ajuste alors assez bien à un processus de Bernoulli. Et la probabilité peut être calculée comme :

$$P(X = 10) = C_{100}^{10}(0{,}10)^{10}(0{,}90)^{90} = 0{,}1319$$

Si on s'intéresse à la probabilité d'avoir au plus 10 individus avec M, alors on calcule

$$P(X \leq 10) = \sum_{x=0}^{10} C_{100}^{x}(0{,}10)^{x}(0{,}90)^{100-x}$$

On a comme résultat :

$$\begin{aligned} P(X \leq 10) &= 0{,}00003 + 0{,}00030 + 0{,}00162 + 0{,}00589 + \\ &\quad 0{,}01587 + 0{,}03387 + 0{,}05958 + 0{,}08890 + \\ &\quad 0{,}11482 + 0{,}13042 + 0{,}13187 \\ &= 0{,}58316 \end{aligned}$$

◆

1.3 LOI DE POISSON

On parle de la loi de Poisson comme celle des événements rares. Comme on le verra, cette appellation est en partie justifiée bien que la fréquence des événements ne soit pas toujours faible. C'est une question de perspective. S'il est rare d'observer un cas incident de cancer du poumon dans une

population de 1000 personnes-années, il n'est pas rare d'y observer un cas incident de grippe. Et pourtant, ces deux phénomènes peuvent être décrits à l'aide de la loi de Poisson.

1.3.1 ESSAI DE POISSON

Considérons l'expérience élémentaire d'observer le nombre d'occurrences (succès) d'un certain événement dans un intervalle de temps très court *dt*, appelé *intervalle élémentaire*. Alors, cette expérience constitue un essai de Poisson si la probabilité que se produise un événement est faible et celle que se produisent plus d'un événement est négligeable. Cette probabilité d'un succès est proportionnelle à *dt* et est fonction d'une constante τ, appelée l'intensité de l'essai. Cette constante correspond au nombre moyen λ de succès rapporté à un intervalle élémentaire *dt*.

$$\tau = \frac{\lambda}{dt}$$

Si on désigne par *X* le nombre de succès pour une expérience élémentaire de Poisson, on peut se rendre compte qu'une fonction comme $P(X) = \frac{e^{-\lambda}\lambda^X}{X!}$ décrit très bien la probabilité du nombre de tels succès. En effet, si l'intervalle de temps *dt* est très court, la valeur λ est aussi très faible : $\lambda \approx 0$. Alors,

$$\begin{aligned} P(X=0) &= e^{-\lambda} &&\approx 1 \\ P(X=1) &= \lambda e^{-\lambda} &&\approx \lambda \\ P(X=2) &= \frac{\lambda^2 e^{-\lambda}}{2!} &&\approx 0 \end{aligned}$$

puisque $\frac{\lambda^2}{2!} \approx 0$. On écrira : $X \mapsto \text{Pois}(\lambda)$, où λ représente le paramètre unique de cette loi de distribution.

Comme exemple d'essai de Poisson, on peut considérer l'expérience d'observer un carrefour routier pendant une heure, pour y dénombrer les accidents d'automobiles. Vraisemblablement, pendant cette période, il ne se produira aucun accident. La probabilité qu'il s'en produise un est faible, qu'il s'en produise plusieurs négligeable.

1.3.2 PROCESSUS DE POISSON

Par analogie avec le processus de Bernoulli, on peut définir le processus de Poisson à partir d'une série d'essais élémentaires indépendants de Poisson.

Sur une période de temps définie, on a un processus de Poisson si les conditions suivantes sont remplies :

1. la probabilité que se produise un succès dans un intervalle de temps est faible, qu'il s'en produise plusieurs est négligeable ; ces probabilités sont proportionnelles à la longueur de l'intervalle ;
2. la réalisation des événements dans un intervalle de temps est indépendante de ce qui se passe dans les autres intervalles de temps ;
3. l'intensité de l'essai ou du processus est constante et égale à τ sur tous les intervalles élémentaires qui composent le processus.

Au processus de Poisson, on peut associer certaines variables aléatoires qui correspondent à des questions comme :

1. Quel temps (durée) T faut-il pour observer r événements ? Ce temps T, pouvant prendre toute valeur ≥ 0, est une variable aléatoire qui obéit à une loi gamma.
2. Quel temps T faut-il pour observer 1 événement ? Cas particulier de la situation précédente, cette durée T est aussi une variable aléatoire. Elle obéit à une loi exponentielle.
3. Pour un temps $T = t$ fixé, combien va-t-on observer d'événements ? Ce nombre X d'événements est la somme des nombres X_i d'événements sur chacun des essais élémentaires : $X = \sum X_i$. Cette variable X, dont les valeurs se situent entre 0 et ∞, obéit à une loi de distribution de Poisson, la plus connue des lois reliées au processus de Poisson.

1.3.3 PROCESSUS DE POISSON DANS LE CONTEXTE ÉPIDÉMIOLOGIQUE

Lorsqu'en épidémiologie on s'intéresse aux décès qui surviennent dans une population dynamique ouverte sur une période de temps déterminée, on se retrouve face à une expérience analogue à celle découlant d'un processus de Poisson. D'abord, en regard des décès qui peuvent s'y produire,

l'observation d'une personne-année cumulée dans cette population peut être considérée comme une expérience élémentaire de Poisson. La personne-année, correspondant à une surface élémentaire *ds* dans le plan, peut générer un ou plusieurs décès. La probabilité d'observer un décès sur *ds* est faible et celle d'en observer plusieurs est négligeable. L'intensité τ de ce processus correspond en approximation au taux de décès.

S'il est difficile, voire impossible, de s'imaginer qu'une personne observée pendant un an (365 jours) puisse de quelque façon être à risque de subir plus d'un décès, il en va autrement pour une personne-année cumulée à l'intérieur d'une population : la personne-année est plutôt considérée comme le résultat d'une multitude de personnes observées pour une courte période de temps, soit par exemple 365 personnes observées pendant une journée approximativement, ou 8766 personnes pendant une heure, ou 525 960 personnes pendant une minute, etc. Dans cette perspective, on comprend bien que, même si une personne ne peut générer qu'un décès, une personne-année, elle, peut en générer plusieurs.

Ainsi, l'observation d'une personne-année pour le décès peut être considérée comme une expérience élémentaire de Poisson d'intensité τ (= taux de décès) s'exerçant sur une surface unitaire *ds* (= 1 personne-année). Le paramètre de la distribution de Poisson en cause est λ (= τds). Ce paramètre λ correspond alors au nombre moyen de décès pour une personne-année.

Les personnes-temps cumulées dans une population dynamique ouverte, en situation d'équilibre quant au phénomène étudié, peuvent assez facilement être associées à un processus de Poisson. Si la personne-temps est telle qu'il soit rare d'y observer un événement et négligeable d'en observer plusieurs, alors l'observation de toutes les personnes-temps constitue en bonne approximation un processus de Poisson. La condition de régularité ou d'équilibre du phénomène assure jusqu'à un certain point l'indépendance entre ces expériences élémentaires (personnes-temps).

Il est plus difficile de faire un tel rapprochement avec les personnes-temps cumulées dans une cohorte (population dynamique fermée). On peut évoquer principalement deux raisons :

1. en cours d'observation, le nombre d'événements possibles par essai élémentaire décroît ;
2. en cours d'observation, la probabilité d'observer un événement est susceptible de changer sensiblement.

Ces deux conditions, marquées surtout dans les cohortes de faibles tailles, peuvent engendrer directement une dépendance entre les personnes-temps.

Par ailleurs, si les cohortes sont de taille suffisante et que l'intensité du phénomène est faible (τ petit), alors le rapprochement de ces personnes-temps avec un processus de Poisson pourrait être valable.

1.3.4 PROPRIÉTÉS DE LA LOI DE POISSON

Considérons la variable X de Poisson telle que $X \mapsto \text{Pois}(\lambda)$.

1. La distribution de X est caractérisée par le seul paramètre λ : $\lambda = \tau ds$.
2. La moyenne $E(X)$ de cette distribution est λ : $E(X) = \sum x_i P(X = x_i) = \lambda$.
3. La variance $V(X)$ de cette distribution est aussi λ : $V(X) = \sum P(x_i)\left[x_i - E(X)\right]^2 = \lambda$.
4. La somme de deux variables de Poisson indépendantes, respectivement de paramètres λ_1 et λ_2, est aussi une variable de Poisson de paramètre $(\lambda_1 + \lambda_2)$. Par extension, on peut dire que la somme de n variables de Poisson indépendantes, de paramètres λ_1, λ_2, λ_3,... λ_n, est aussi une variable de Poisson de paramètre $\lambda_1 + \lambda_2 + \lambda_3 + ... + \lambda_n$. Dans le cas particulier où l'on fait la somme de n variables indépendantes X_i de Poisson de même paramètre λ, on obtient une nouvelle variable X de Poisson de paramètre $\mu = n\lambda$. Dans ce contexte, on a aussi $E(X) = V(X) = \mu$.
5. Si l'expérience élémentaire de Poisson correspond à une personne-temps avec le paramètre $\lambda = \tau ds$, alors les n personnes-temps correspondent aussi à une expérience de Poisson avec le paramètre $\mu = n\lambda$. Le taux t d'événements observé ($t = \frac{x}{n}$) est une estimation de l'intensité τ du processus de Poisson. La moyenne ou valeur attendue de $E(t)$ et la variance $V(t)$ de $t = \frac{X}{n}$ sont respectivement τ et $\frac{\tau}{n}$.
6. Puisque la variance $V(X)$ de X est liée directement à la moyenne $E(X)$, il peut être intéressant de déterminer une transformation de X qui stabilise cette variance. La racine carrée de X est une telle transformation : $\sqrt{X}$. Dans le texte, on désignera par RAC cette

transformation. On peut montrer que $V\left(\sqrt{X}\right) = \frac{1}{4}$. En conséquence, pour un taux t, la variance de RAC(t) est donnée par : $V\left(\sqrt{t}\right) = \frac{1}{4n}$.

7. La loi de Poisson peut être considérée comme une loi limite de la loi binomiale. Lorsque la taille n de l'échantillon est large et π est faible, la loi de Poisson de paramètre $\mu = n\pi$ et la loi binomiale Bin(π,n) conduisent à des probabilités similaires. Dans ce cas, on peut remplacer la loi binomiale par la loi de Poisson, qui est d'un maniement plus facile.

EXEMPLE 1.3

Le taux de mortalité dans une population est de 0,007 par année, ou de 0,007 décès par personne-année. On veut calculer la probabilité d'observer exactement 0 décès, 1 décès et au moins 1 décès, dans un échantillon de 10 personnes-années : $P(X = 0)$, $P(X = 1)$ et $P(X \geq 1)$.

Pour le calcul, rappelons que $P(X = x) = \frac{e^{-\mu}\mu^x}{x!}$.

De l'énoncé, on peut tirer les valeurs suivantes :

$\lambda = \tau ds = (0{,}007 \text{ an}^{-1}) \times (1 \text{ an}) = 0{,}007$

$\mu = 0{,}007 \times 10 = 0{,}07$

X = le nombre de décès.

Donc :

$$P(X = 0) = \frac{e^{-0{,}07}\,0{,}07^0}{0!} \approx 0{,}9324$$

$$P(X = 1) = \frac{e^{-0{,}07}\,0{,}07^1}{1!} \approx 0{,}0653$$

$$P(X \geq 1) = 1 - P(X = 0) \approx 1 - 0{,}9324 = 0{,}0676$$

◆

EXEMPLE 1.4

Sur la base des données de l'exemple 1.3 et pour un échantillon de 1000 personnes-années, on veut calculer la probabilité d'observer exactement 0 décès, 1 décès et au moins 1 décès.

On fait alors $\mu = 0{,}007 \times 1000 = 7$.

Ainsi,

$$P(X=0)=\frac{e^{-7}\,7^{0}}{0!}\approx 0{,}0009$$

$$P(X=1)=\frac{e^{-7}\,7^{1}}{1!}\approx 0{,}0064$$

$$P(X\geq 1)=1-P(X=0)\approx 1-0{,}0009=0{,}9991$$

◆

EXEMPLE 1.5

La prévalence d'une malformation à la naissance est de 1 cas pour 1000 naissances. On s'intéresse alors à la probabilité d'observer 10 cas de malformation dans un échantillon de 5000 bébés naissants.

Cette probabilité devrait se calculer dans le cadre de la loi binomiale comme : $P(X=10\,|\,\pi=0{,}001,\ n=5000) = C_{5000}^{10}(0{,}001)^{10}(0{,}999)^{4990}$

On peut simplifier les calculs en utilisant la loi de Poisson de paramètre $\mu = 5$ ($\mu = n\pi$). On a alors,

$$P(X=10\,|\,\mu=5)=\frac{e^{-5}\,5^{10}}{10!}=0{,}0181$$

◆

1.4 LOI HYPERGÉOMÉTRIQUE

On considère une population finie de N individus dont M_1 possèdent une caractéristique alors que les M_0 ($= N - M_1$) autres ne l'ont pas. De cette population, on tire un échantillon, sans remise, de n individus. On suppose qu'à chaque tirage les individus composant la population résiduelle ont même chance d'être sélectionnés pour l'échantillon. On s'intéresse alors au nombre d'individus qui, dans l'échantillon, ont la caractéristique. Ce nombre X est une variable aléatoire dont le domaine de variation est fonction des tailles relatives de l'échantillon et des groupes M_1 et M_0. Ce nombre X ne peut pas être plus petit que le plus grand des deux nombres suivants : 0 et $n - M_0$. Par ailleurs, il ne peut pas être plus grand que le plus petit des deux nombres n et M_1. On écrit : $\max(0, n - M_0) \leq X \leq \min(n, M_1)$.

On peut représenter cette expérience d'échantillonnage à l'aide du tableau 1.5.

Alors, la probabilité que X soit précisément égale à x (une valeur particulière de X) est décrite par :

$$P(X=x) = \frac{C_{M_1}^{x}C_{M_0}^{n-x}}{C_N^n}$$

$$= \frac{M_1!M_0!n!(N-n)!}{x!(M_1-x)!(n-x)![M_0-(n-x)]!N!}$$

TABLEAU 1.5

Caractéristique	Échantillon	Population
Présente	x	M_1
Absente	$n - x$	M_0
Total	n	N

Nous allons concrétiser ce calcul à l'aide de l'exemple suivant.

EXEMPLE 1.6

Une urne contient 20 boules dont 12 sont blanches et les 8 autres noires. On tire un échantillon aléatoire de 5 boules. Quelle est la probabilité que cet échantillon contienne exactement 2 boules blanches (et donc 3 noires) ?

Nous décrivons le résultat de l'expérience dans le tableau 1.6.

TABLEAU 1.6

Couleur	Échantillon	Urne
Blanche	2	12
Noire	3	8
Total	5	20

Pour calculer cette probabilité, il nous faut établir le rapport suivant :

$$\frac{\text{nombre d'échantillons de 5 boules qui donnent le résultat désiré}}{\text{nombre total d'échantillons de 5 boules}}$$

Ce rapport correspond précisément à la probabilité recherchée. Décortiquons-le.

Le nombre d'échantillons de 5 boules qui peuvent être constitués à partir de l'ensemble de 20 boules est égal à autant de sous-ensembles distincts de 5 éléments que l'on peut constituer à partir d'un ensemble de 20 éléments. Ce nombre est donné par C_{20}^5 : $C_{20}^5 = \dfrac{20!}{5!(20-5)!}$

Le nombre d'échantillons de 5 boules qui donnent exactement 2 boules blanches et 3 noires à partir de cette urne, qui contient 12 boules blanches et 8 noires, est égal au produit des deux nombres suivants :

- le nombre de sous-ensembles distincts de 2 éléments que l'on peut former à partir d'un ensemble de 12 éléments : $C_{12}^2 = \dfrac{12!}{2!(12-2)!}$
- le nombre de sous-ensembles distincts de 3 éléments que l'on peut former à partir d'un ensemble de 8 éléments : $C_8^3 = \dfrac{8!}{3!(8-3)!}$.

La probabilité recherchée est alors donnée par

$$\begin{aligned} P(X=2) &= \frac{C_{12}^2 \times C_8^3}{C_{20}^5} \\ &= \frac{12!8!5!15!}{2!10!3!5!20!} \\ &\approx 0{,}2384 \end{aligned}$$

◆

Propriétés de la loi hypergéométrique

Considérons la variable hypergéométrique X telle que $X \mapsto \text{Hypr}(N,M_1,n)$

1. La distribution de la variable X est caractérisée par trois paramètres : N, M_1 et n, où $\frac{n}{N}$ représente la fraction d'échantillonnage f.
2. La moyenne $E(X)$ de cette distribution est $\frac{nM_1}{N}$.
3. La variance de $V(X)$ cette distribution est $\frac{nM_1M_0(N-n)}{N^2(N-1)}$.
4. Plus la fraction d'échantillonnage est faible, plus la loi hypergéométrique se rapproche d'une loi binomiale. La loi binomiale peut être considérée comme une loi limite de la loi hypergéométrique. Ainsi, quand la taille n de l'échantillon est faible par rapport à la taille N de la population, les lois hypergéométrique $\text{Hypr}(N,M_1,n)$ et binomiale $\text{Bin}(\pi,n)$ où $\pi = \frac{M_1}{N}$, conduisent à des probabilités similaires. Dans ce cas, on peut remplacer la loi hypergéométrique par la loi binomiale.

Exemple 1.7

Dans une clinique médicale, sur un total de 100 patients chez qui on a diagnostiqué la maladie *M*, 48 sont des hommes. Aux fins d'une étude, on décide d'examiner les dossiers de 50 patients parmi les 100. Si ces 50 dossiers sont tirés aléatoirement, quelle est la probabilité que l'échantillon comprenne les dossiers de 30 patients masculins ?

On peut figurer le problème à l'aide du tableau de fréquences suivant.

TABLEAU 1.7

Sexe	Échantillon	Population
Masculin	30	48
Féminin		52
Total	50	100

Remarquons d'abord que la variable X (nombre de dossiers masculins dans l'échantillon) peut varier entre 0 et 48 : $\max(0,-2) \leq X \leq \min(50,48)$.

La probabilité recherchée est donnée par la loi hypergéométrique :

$$P(X=30) = \frac{48!52!50!\,(100-50)!}{30!\,(48-30)!\,(50-30)!\,[52-(50-30)]!\,100!}$$
$$= 0{,}0091$$

L'approximation par une loi binomiale, où $\pi = 0{,}48$ et $n = 50$, donne comme probabilité :

$$P(X=30) = C_{50}^{30}(0{,}48)^{30}(0{,}52)^{20}$$
$$= 0{,}0270$$

Les probabilités sont sensiblement différentes. L'approximation binomiale n'est donc pas valide ici. On remarque que la fraction d'échantillonnage est grande : $f = 0{,}50$.

◆

EXEMPLE 1.8

Dans cette même clinique (voir l'exemple 1.7), sur l'ensemble des 5000 dossiers médicaux, 2400 sont ceux de patients masculins. Si on tire un échantillon aléatoire de 50 dossiers, quelle est la probabilité que 30 soient des dossiers de patients masculins ?

On peut figurer le problème à l'aide du tableau de fréquences suivant.

TABLEAU 1.8

Sexe	Échantillon	Population
Masculin	30	2400
Féminin	20	2600
Total	50	5000

La variable X (nombre de dossiers masculins dans l'échantillon) peut prendre toute valeur comprise entre 0 et 50 : $0 \leq X \leq 50$.

La probabilité recherchée est donnée par :

$$P(X=30) = \frac{2400!\,2600!50!\,(5000-50)!}{30!\,(2400-30)!\,(50-30)!\,[2600-(50-30)]!\,5000!}$$
$$= 0{,}02670$$

La valeur 0,0270 calculée par l'approximation binomiale est sensiblement la même que celle calculée par la loi hypergéométrique. On remarque d'ailleurs que la fraction d'échantillonnage est faible ($f = 0{,}01$).

◆

1.5 LOI NORMALE

Une variable aléatoire normale X, de moyenne μ et de variance σ^2, a comme fonction de densité :

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$

La distribution d'une variable normale X est entièrement caractérisée par les deux paramètres, μ et σ^2. On écrit $X \mapsto N(\mu,\sigma^2)$.

L'expression $\frac{X-\mu}{\sigma}$, que l'on retrouve dans la fonction de densité $f(X)$, définit une nouvelle variable normale Z, dite *centrée réduite* : X est centrée à la moyenne μ et de dispersion réduite par $\frac{1}{\sigma}$. Ainsi Z est normale, de moyenne 0 et de variance 1.

La loi normale est une loi omniprésente en statistique. Les données des tableaux de contingence n'y échappent pas. L'importance de cette loi relève surtout d'une propriété fondamentale énoncée sous forme de théorème : le théorème de la limite centrale.

Théorème de la limite centrale

Si X est une variable normale de moyenne m et de variance σ^2, *alors* $\overline{X}$ *est aussi une variable normale, de moyenne* μ *et de variance* $\frac{\sigma^2}{n}$.

Un énoncé encore plus fort et plus général

Soient n variables aléatoires indépendantes, identiquement distribuées, de moyenne μ *et de variance* σ^2 *:* $X_1, X_2, \ldots X_n$. *On forme la somme* $w = X_1 + X_2 + \ldots + X_n$, *dont la moyenne* $E(w)$ *et la variance* $\sigma^2(w)$ *sont respectivement données par* $n\mu$ *et* $n\sigma^2$. *Alors la variable-somme w, centrée réduite,*

$$Z_n = \frac{w - E(w)}{\sigma(w)}$$

converge vers une variable Z normale de moyenne 0 *et de variance* 1, *quand* $n \to \infty$.

La convergence de Z_n vers Z est assez rapide. On considère que déjà avec $n = 30$, l'écart entre la loi de distribution de Z_n et celle de Z (la loi normale) est faible, et ce, quelle que soit la distribution de la variable X au départ.

Une conséquence importante de ce théorème

Si X est une variable aléatoire (comme une binomiale, une Poisson, une hypergéométrique...) de moyenne $E(X)$ et de variance $V(X)$, observée sur un échantillon de taille n suffisamment grande, alors

$$\frac{X - E(X)}{\sqrt{V(X)}} \approx Z$$

Si m représente une mesure telle un taux t, une proportion p, une différence ou un rapport de taux (*DT* ou *RT*), une différence ou rapport de proportions (*DR* ou *RR*), un rapport de cotes (*RC*), ou encore une transformation appropriée de ces mesures, alors on peut inférer que

$$\frac{m - E(m)}{\sqrt{V(m)}} \approx Z$$

lorsque les échantillons sont suffisamment grands.

1.6 LOI DU KHI-CARRÉ

Si Z représente une variable normale centrée réduite ($\mu = 0$ et $\sigma^2 = 1$), alors Z^2 obéit à une loi du khi-carré (χ^2) à 1 degré de liberté.

Si $X \mapsto N(0,1)$ *alors* $Z^2 \mapsto \chi^2$

Ainsi, pour $\frac{X - E(X)}{\sqrt{V(X)}} = Z$, on a $\frac{[X - E(X)]^2}{V(X)} \approx \chi_1^2$

De même si $Z_1, Z_2, ..., Z_n$ sont n variables normales centrées réduites indépendantes, alors la somme $Z^2(n)$ des carrés de ces variables obéit à une loi du χ^2 avec n degrés de liberté :

$$Z^2(n) = Z_1^2 + Z_2^2 + ... + Z_n^2 \qquad \chi_n^2$$

La moyenne et la variance de $Z^2(n)$ sont respectivement égales à n et $2n$:

$$E[Z^2(n)] = n \text{ et } V[Z^2(n)] = 2n$$

On pourrait encore ici appliquer le théorème de la limite centrale à la variable $Z^2(n)$, comme suit :

$$\frac{Z^2(n) - E[Z^2(n)]}{\sqrt{V[Z^2(n)]}} = \frac{Z^2(n) - n}{\sqrt{2n}} \mapsto \mathrm{N}(0,1)$$

1.7 RELATIONS ENTRE LES LOIS DE DISTRIBUTION DANS UN TABLEAU 2 × 2

Entre les lois de Poisson, binomiale et hypergéométrique, il existe certaines relations déjà bien connues. Par exemple, n'a-t-on pas dit que la binomiale, dans certains cas, peut tendre vers une loi de Poisson ? Qu'une loi hypergéométrique peut parfois tendre vers une binomiale ? Il existe entre ces lois, par ailleurs, des relations moins bien connues que les précédentes ou moins citées dans les ouvrages statistiques de base, mais combien plus riches pour la définition d'outils d'analyse dans les tableaux de contingence. Nous allons présenter ces relations. Nous aurons ainsi l'occasion de décrire deux autres lois : la loi multinomiale et la loi hypergéométrique multiple.

1.7.1 LOI DE POISSON, LOI BINOMIALE

Soit une étude de cohorte où l'on s'intéresse à un facteur d'exposition X et à une maladie Y. L'apparition d'un cas de maladie est représentée par $Y = 1$, tant chez les exposés ($X = 1$) que chez les non-exposés ($X = 0$) au facteur X. Cette étude génère des données des personnes-années (tableau 1.9).

TABLEAU 1.9

	$X = 1$	$X = 0$	Total
$Y = 1$	a_1	a_0	a
Personnes-années	n_1	n_0	n

Considérons le nombre n_1 de personnes-années comme autant d'expériences élémentaires indépendantes de Poisson de même paramètre λ_1. À ce processus de Poisson, on associe la variable A_1 désignant le nombre de décès chez les exposés. De même, le nombre n_0 personnes-années correspond à autant d'expériences de Poisson de même paramètre λ_0, auquel on associe la variable A_0 pour désigner le nombre de décès chez les non-exposés. Alors, les variables A_1 et A_0 sont chacune des variables de Poisson dont les paramètres sont respectivement $n_1\lambda_1$ et $n_0\lambda_0$. Ainsi,

$$P(A_1 = a_1) = \frac{e^{-n_1\lambda_1}(n_1\lambda_1)^{a_1}}{a_1!} \text{ et } P(A_0 = a_0) = \frac{e^{-n_0\lambda_0}(n_0\lambda_0)^{a_0}}{a_0!}$$

Soit maintenant la variable $A = A_1 + A_0$ et supposons que A soit égale à la valeur fixe a. Sous ces contraintes, considérons $P(A_1 = a_1)$:

$$\begin{aligned} P(A_1 = a_1) &= \frac{P(A_1 = a_1 \,\&\, A_0 = a_0)}{P(A = a)} \\ &= \frac{\dfrac{e^{-n_1\lambda_1}(n_1\lambda_1)^{a_1}}{a_1!} \times \dfrac{e^{-n_0\lambda_0}(n_0\lambda_0)^{a_0}}{a_0!}}{\dfrac{e^{-(n_1\lambda_1+n_0\lambda_0)}(n_1\lambda_1+n_0\lambda_0)^{a}}{a!}} \\ &= \frac{e^{-(n_1\lambda_1+n_0\lambda_0)}(n_1\lambda_1)^{a_1}(n_0\lambda_0)^{a_0}\,a!}{e^{-(n_1\lambda_1+n_0\lambda_0)}(n_1\lambda_1+n_0\lambda_0)^{a}\,a_1!a_0!} \\ &= C_a^{a_1}\frac{(n_1\lambda_1)^{a_1}(n_0\lambda_0)^{a_0}}{(n_1\lambda_1+n_0\lambda_0)^{a}} \end{aligned}$$

Dans ces conditions, la variable A_1 obéit à une loi binomiale : $A_1 \mapsto \text{Bin}(\pi,a)$ où $\pi = \dfrac{n_1\lambda_1}{n_1\lambda_1 + n_0\lambda_0}$. Les paramètres a et π de cette distribution binomiale sont, pour l'un, déterminé par la condition $A = a$ et pour l'autre, par une fonction des paramètres des deux variables de Poisson. En posant $\varphi = \dfrac{\lambda_1}{\lambda_0}$, le paramètre π de la variable binomiale peut avantageusement être décrit comme une fonction du rapport φ des deux paramètres, λ_1 et λ_0, et donc des deux taux, τ_1 et τ_0 : $\pi = \dfrac{\varphi n_1}{\varphi n_1 + n_0}$. Dans cette écriture, on peut facilement établir que, sous l'hypothèse d'un rapport de taux $\varphi = 1$, la variable A_1 obéit à une distribution binomiale de paramètre $\pi = \dfrac{n_1}{n_1 + n_0}$ et $n = a$.

Pour une hypothèse quelconque sur φ, $A_1 \mapsto \text{Bin}(\pi,a)$ où $\pi = \dfrac{\varphi n_1}{\varphi n_1 + n_0}$.

Exemple 1.9

Considérons le tableau 1.10, qui décrit les données d'une étude de cohorte portant sur l'association entre le facteur *X* et la maladie *Y*. Les données sont en personnes-années.

Tableau 1.10

	$X = 1$	$X = 0$	Total
$Y = 1$	2	10	12
Personnes-années	3000	5000	8000

Au départ, suivant les conditions mêmes d'échantillonnage, les 8000 personnes-années de l'étude se partagent en 3000 exposées et 5000 non-exposées. Sous l'hypothèse qu'il n'y a pas d'association entre la maladie *Y* et le facteur *X*, l'observation de 2 cas exposés parmi les 12 cas recensés revient, par analogie, à observer 2 faces après 12 lancers d'une pièce de monnaie pour laquelle la probabilité π d'observer face est de 3/8. (3000/8000). On peut donc déduire la probabilité d'observer 2 cas exposés sous l'hypothèse qu'il n'y a pas d'association entre la maladie et le facteur *X* en utilisant la loi de distribution binomiale :

$$P(A_1 = 2) = C_{12}^2 (3/8)^2 (5/8)^{10} = 0{,}0844$$

◆

1.7.2 Loi de Poisson, loi multinomiale

Considérons *J* variables de Poisson A_j indépendantes de paramètre μ_j, $1 \leq j \leq J$. La somme des *J* variables définit une nouvelle variable de Poisson $A = \sum_j A_j$ de paramètre $\mu = \sum_j \mu_j$. Pour une valeur fixe de *A*, les variables A_j obéissent à une loi de distribution multinomiale.

Transposons ces propriétés des variables A_j au contexte d'une étude de cohorte avec des données de personnes-temps. On s'intéresse à un facteur d'exposition *X* qui comporte *J* catégories d'exposition et à une maladie Y (tableau 1.11).

Tableau 1.11

	Facteur d'exposition *X*				Total
	1	2	..	*J*	
$Y = 1$	a_1	a_2	..	a_J	a
Personnes-années	n_1	n_2	..	n_J	n

Pour chaque catégorie *j*, le taux $t_j = \dfrac{a_j}{n_j}$ est une estimation du taux théorique τ_j. La variable de Poisson A_j a comme valeur observée a_j et comme valeur attendue $\mu_j = \tau_j n_j$.

Sous la condition $\sum_j A_j = a$, il est facile de décrire cette loi multinomiale :

$$P(A_1 = a_1,\ A_2 = a_2,\ \ldots,\ A_J = a_J \mid A = a) \ =\ a! \times \prod_{j=1}^{J} \frac{\pi_j^{a_j}}{a_j!}$$

où pour un $j = j'$ donné mais quelconque, on a : $\pi_{j'} = \dfrac{\tau_{j'} n_{j'}}{\sum_{j=1}^{J} \tau_j n_j}$.

Évidemment $\sum_j \pi_j = 1$.

Sous l'hypothèse que les taux τ_j sont tous égaux (hypothèse nulle H_0), la loi multinomiale se réduit à :

$$\begin{aligned} P(A_1 = a_1,\ A_2 = a_2,\ \ldots,\ A_J = a_J \mid A = a) \ &=\ a! \times \prod_{j=1}^{J} \frac{\pi_j^{a_j}}{a_j!} \\ &=\ \frac{a!}{n^a} \times \prod_{j=1}^{J} \frac{\left(n_j\right)^{a_j}}{a_j!} \end{aligned}$$

sachant alors que $\pi_j = \dfrac{n_j}{n}$.

Quelques propriétés des variables multinomiales sous H_0

1. La moyenne de la variable A_j est donnée par $E(A_j) = \dfrac{a n_j}{n}$.
2. La variance de la variable A_j est donnée par
   $V(A_j) = \dfrac{a(n - n_j) n_j}{n^2}$.
3. La moyenne de la somme de deux variables, A_k et A_h, liées à une même distribution multinomiale, est la somme des moyennes de ces variables : $E(A_k + A_h) = E(A_k) + E(A_h) = \dfrac{a n_k}{n} + \dfrac{a n_h}{n}$.
4. La covariance entre deux variables A_k et A_h liées à une même distribution multinomiale est donnée par
   $Cov(A_k, A_h) = -\dfrac{a n_h n_k}{n^2}$.

5. La variance de la somme de deux variables, A_k et A_h, liées à une même distribution multinomiale, est la somme des variances de ces variables et de deux fois leur covariance :

$$V(A_k + A_h) = V(A_k) + V(A_h) + 2Cov(A_k, A_h)$$
$$= \frac{a(n-n_k)n_k}{n^2} + \frac{a(n-n_h)n_h}{n^2} - 2\frac{an_kn_h}{n^2}$$

1.7.3 LOI BINOMIALE, LOI HYPERGÉOMÉTRIQUE

TABLEAU 1.12

Groupe	$X = 1$	$X = 0$	Total
$Y = 1$	a_1	a_0	a
$Y = 0$	b_1	b_0	b
Total	n_1	n_0	n

Soit une étude de cohorte qui génère des données d'incidence cumulative (tableau 1.12). On y considère le facteur X et la maladie Y.

Les variables A_1 et A_0, décrivant le nombre de cas ($Y = 1$) respectivement chez les exposés ($X = 1$) et chez les non-exposés ($X = 0$), sont considérées comme deux variables binomiales indépendantes. Pour le tableau, les valeurs observées de A_1 et A_0 sont respectivement a_1 et a_0.

$$A_1 \mapsto \text{Bin}(\pi_1, n_1) \qquad P(A_1 = a_1) = C_{n_1}^{a_1} \pi_1^{a_1} (1-\pi_1)^{(n_1-a_1)}$$
$$A_1 \mapsto \text{Bin}(\pi_1, n_1) \qquad P(A_0 = a_0) = C_{n_0}^{a_0} \pi_0^{a_0} (1-\pi_0)^{(n_0-a_0)}$$

Soit la variable $A = A_1 + A_0$ et supposons $A = a$ fixe.

Sous ces contraintes, considérons $P(A_1 = a_1)$:

$$P(A_1 = a_1 \mid A = a) = \frac{P(A_1 = a_1 \text{ et } A_0 = a - a_1)}{P(A = a)}$$

$$= \frac{C_{n_1}^{a_1} \pi_1^{a_1} (1-\pi_1)^{(n_1-a_1)} \times C_{n_0}^{a_0} \pi_0^{a_0} (1-\pi_0)^{(n_0-a_0)}}{\sum_{x=\underline{A_1}}^{\overline{A_1}} C_{n_1}^{x} \pi_1^{x} (1-\pi_1)^{(n_1-x)} \times C_{n_0}^{a-x} \pi_0^{(a-x)} (1-\pi_0)^{[n_0-(a-x)]}}$$

où $\underline{A_1} = \max(0, n_1 - b)$ et $\overline{A_1} = \min(n_1, a)$.

Posons $\kappa_j = \frac{\pi_j}{1-\pi_j}$, où κ_j désigne la cote des risques respectivement chez les exposés ($j = 1$) et chez les non-exposés ($j = 0$). En faisant les substitutions nécessaires, on obtient l'expression :

$$P(A_1 = a_1 \mid A = a) = \frac{C_{n_1}^{a_1} C_{n_0}^{a_0} \kappa_1^{a_1} \kappa_0^{a_0}}{\sum_{x=\underline{A_1}}^{\overline{A_1}} C_{n_1}^{x} C_{n_0}^{a-x} \kappa_1^{x} \kappa_0^{a-x}}$$

Aussi, suivant le rapport de cotes $\psi = \frac{\kappa_1}{\kappa_0}$, l'expression devient

$$P(A_1 = a_1 \mid A = a) = \frac{C_{n_1}^{a_1} C_{n_0}^{a_0} \psi^{a_1}}{\sum_{x=\underline{A_1}}^{\overline{A_1}} C_{n_1}^{x} C_{n_0}^{a-x} \psi^{x}}$$

C'est la loi ***hypergéométrique non centrée***. Sous H_0, $\Psi = 1$ (c'est-à-dire $RC = 1$), l'expression se réduit à :

$$P(A_1 = a_1) = \frac{C_{n_1}^{a_1} \times C_{n_0}^{a_0}}{C_n^a} = \frac{n_1! n_0! a! b!}{a_1! a_0! b_1! b_0! n!}$$

TABLEAU 1.13

Groupe	$X = 1$	$X = 0$	Total
$Y = 1$	a_1	a_0	m_1
$Y = 0$	b_1	b_0	m_0
Total	n_1	n_0	N

C'est la loi hypergéométrique telle qu'elle fut présentée précédemment, que l'on dira loi hypergéométrique centrée.

Pour le schéma des études cas-témoins, la démonstration est analogue ; cependant le rôle des marges est inversé. Ainsi, considérons une étude cas-témoins (tableau 1.13).

Les variables A_1 et B_1 sont considérées comme deux variables binomiales indépendantes, $A_1 \mapsto \text{Bin}(\pi_1, m_1)$ et $B_1 \mapsto \text{Bin}(\pi_0, m_0)$, telles que :

$$P(A_1 = a_1) = C_{m_1}^{a_1} \pi_1^{a_1} (1-\pi_1)^{m_1 - a_1}$$

$$P(B_1 = b_1) = C_{m_0}^{b_1} \pi_0^{b_1} (1-\pi_0)^{m_0 - b_1}$$

La probabilité conditionnelle de A_1 prend alors la forme :

pour la loi hypergéométrique non centrée

$$P(A_1 = a_1 \mid n_1 = a_1 + b_1) = \frac{C_{m_1}^{a_1} C_{m_0}^{b_1} \psi^{a_1}}{\sum_{x=\underline{A_1}}^{\overline{A_1}} C_{m_1}^{x} C_{m_0}^{n_1 - x} \psi^{x}}$$

pour la loi hypergéométrique sous l'hypothèse nulle

$$P(A_1 = a_1 \mid n_1 = a_1 + b_1) = \frac{C_{m_1}^{a_1} C_{m_0}^{b_1}}{C_n^{n_1}}$$

On remarque que :

$$P(A_1 = a_1 \mid n_1 = a_1 + b_1) = \frac{C_{m_1}^{a_1} \times C_{m_0}^{b_1}}{C_n^{n_1}} = \frac{n_1! n_0! m_1! m_0!}{a_1! a_0! b_1! b_0! n!}$$

L'expression est identique à celle qu'on trouve dans le schéma d'analyse pour les études de cohortes, où a et b sont mis pour m_1 et m_0 respectivement.

Exemple 1.10

Considérons le tableau 1.14 qui décrit les données d'une étude de cas ($Y = 1$) et témoins ($Y = 0$) où on soupçonne le facteur X d'être un facteur de risque pour la maladie qui affecte les cas.

Tableau 1.14

Groupe	$X = 1$	$X = 0$	Total
$Y = 1$	2	10	12
$Y = 0$	3	5	8

Au départ, suivant les conditions mêmes d'échantillonnage, les 20 sujets de l'étude se partagent en 12 cas et 8 témoins. Sous l'hypothèse qu'il n'y a pas d'association entre la maladie et le facteur X, l'observation de 2 cas parmi les 5 sujets exposés revient, par analogie avec l'urne, à observer 2 cas dans un échantillon aléatoire de 5 individus tirés du groupe des 20 sujets. On peut donc déduire la probabilité d'observer 2 cas exposés sous l'hypothèse qu'il n'y a pas d'association entre la maladie et le facteur X en utilisant la loi de distribution hypergéométrique :

$$P(A_1 = 2 \mid n_1 = 5) = \frac{5! \times 15! \times 12! \times 8!}{2! \times 10! \times 3! \times 5! \times 20!} = 0{,}2384$$

◆

1.7.4 LOI BINOMIALE, LOI HYPERGÉOMÉTRIQUE MULTIPLE

Considérons maintenant J variables binomiales indépendantes A_j de paramètres π_j et N_j, $1 \leq j \leq J$. On peut concrétiser l'observation de telles variables dans une étude de cohorte de données de personnes portant sur la maladie Y et la variable d'exposition X comportant J catégories d'exposition (tableau 1.15).

TABLEAU 1.15

	Catégorie de X				Total
	1	2	...	J	
$Y = 1$	a_1	a_2	...	a_J	A
$Y = 0$	b_1	b_2	...	b_J	b
Total	n_1	n_2	...	n_J	n

Pour chaque catégorie, on mesure la proportion $p_j = \frac{a_j}{n_j}$ ou la cote $c_j = \frac{a_j}{b_j}$, qui sont des estimations de la proportion théorique π_j ou de la cote κ_j. On a $\kappa_j = \frac{\pi_j}{1 - \pi_j}$.

Sous la condition $\sum_j A_j = a$ (et donc celle $\sum_j B_j = b$), les variables A_j obéissent à une loi hypergéométrique multiple qui peut être décrite comme :

$$P(A_1 = a_1,\ A_2 = a_2,\ \ldots,\ A_J = a_J \mid \{\kappa_j\}, a) = \frac{\prod_{j=1}^{J} C_{n_j}^{a_j} \kappa^{a_j}}{\sum_{u \in R} \prod_{j=1}^{J} C_{n_j}^{a_{uj}} \kappa^{a_{uj}}}$$

où R désigne l'ensemble des réalisations u, $\{u_j\}$, de $\{A_j\}$ telles que $\sum_j u_j = a$.

Sous l'hypothèse nulle que les paramètres π_j sont tous égaux (et donc que les cotes κ_j sont toutes égales), la loi hypergéométrique multiple se réduit à :

$$P(A_1 = a_1,\ A_2 = a_2,\ \ldots,\ A_J = a_J \mid H_0) = \frac{\prod_{j=1}^{J} C_{n_j}^{a_j}}{C_n^a}$$

La variable A_j est une variable hypergéométrique J-dimensionnelle.

Quelques propriétés des variables hypergéométriques multiples sous H_0

1. La moyenne de la variable A_j est donnée par $E(A_j) = \dfrac{an_j}{n}$

2. La variance de la variable A_j est donnée par $V(A_j) = \dfrac{abn_j(n-n_j)}{n^2(n-1)}$

3. La moyenne de la somme de deux variables A_k et A_h liées à une même distribution hypergéométrique multiple est la somme des moyennes de ces variables :
$$E(A_k + A_h) = E(A_k) + E(A_h)$$
$$= \frac{an_k}{n} + \frac{an_h}{n}$$

4. La covariance entre deux variables A_k et A_h liées à une même distribution hypergéométrique multiple est donnée par
$$Cov(A_k, A_h) = -\frac{abn_k n_h}{n^2(n-1)}$$

5. La variance de la somme de deux variables A_k et A_h liées à une même distribution hypergéométrique multiple est la somme des variances de ces variables plus deux fois leur covariance :
$$V(A_k + A_h) = V(A_k) + V(A_h) + 2cov(A_k, A_h)$$
$$= \frac{abn_k(n-n_k)}{n^2(n-1)} + \frac{abn_h(n-n_h)}{n^2(n-1)} - 2\frac{abn_k n_h}{n^2(n-1)}$$

CHAPITRE

2

LE TEST STATISTIQUE ET L'INTERVALLE DE CONFIANCE

2.1 UNE HYPOTHÈSE, UNE ÉTUDE

Une hypothèse est un énoncé qui établit une relation entre deux ou plusieurs objets, deux ou plusieurs variables, et dont la véracité peut être vérifiée empiriquement. Souvent, cette relation est de type causal, dans le sens qu'une ou plusieurs variables, la cause présumée, influence les valeurs d'une autre variable, l'effet présumé. L'hypothèse, plus ou moins précise, appartient à l'horizon des connaissances acquises sur un sujet donné et en marque la frontière qu'il faut dépasser. L'hypothèse confirmée détermine un élément de la frontière. L'hypothèse générée invite à dépasser cette frontière, à pousser au-delà l'exploration des connaissances.

L'hypothèse peut naître du conflit entre deux théories, de la contradiction entre des faits et une théorie ou, tout simplement, de l'observation de faits nouveaux que la théorie ne saurait expliquer complètement.

En recherche empirique, le besoin de confirmer ou d'infirmer une hypothèse génère de nouvelles études, alors que l'exploration des données d'une étude peut donner naissance à de nouvelles hypothèses. Ainsi, tantôt l'hypothèse génère une étude, tantôt l'étude génère une hypothèse, suivant le contexte. L'étude sera dite *confirmatoire* si elle est destinée à confirmer ou à infirmer une hypothèse, *exploratoire* si elle sert à générer une hypothèse. Dans le contexte confirmatoire, l'hypothèse existe avant l'étude : l'hypothèse est dite *a priori* ; dans le contexte exploratoire, l'étude existe avant l'hypothèse : l'hypothèse est dite *a posteriori*. Certaines études sont davantage confirmatoires : les essais cliniques, les études d'évaluation ; d'autres exploratoires : les études descriptives. Par contre, la plupart des études confirmatoires peuvent être aussi exploratoires : confirmatoires quant aux hypothèses qu'elles doivent vérifier, mais exploratoires quant aux hypothèses qu'elles peuvent générer. Soulignons le fait que les données d'une étude (exploratoire) ne sauraient être utilisées pour confirmer l'hypothèse qu'elles ont générée. L'hypothèse générée par les données trouve son maximum de vraisemblance sur ces données. De nouvelles données sont requises pour juger de sa vraisemblance, sinon pour la confirmer.

Il peut arriver que les données d'une étude planifiée pour confirmer une hypothèse *A* puissent être utilisées pour confirmer une hypothèse *B* déjà énoncée ou reconnue. Dans ce cas, l'étude peut être considérée confirmatoire tant pour l'hypothèse *A* que pour l'hypothèse *B*.

Les termes confirmatoire et exploratoire permettent de qualifier les liens qui existent entre hypothèses et études, mais pas de traduire correctement tout le processus de la construction des connaissances sur un sujet donné. S'il suffit souvent d'une seule étude pour générer une hypothèse intéressante, en général il en faut plusieurs pour la confirmer, surtout dans le contexte des études d'observation. Cette situation est principalement due aux erreurs qui peuvent entacher la vérification d'une hypothèse. Une taille suffisante d'échantillon peut permettre de minimiser l'erreur aléatoire, mais ne peut en rien garantir l'absence d'erreur systématique : absence d'erreur de sélection, d'information ou de confusion, particulièrement. Ces erreurs, surtout présentes dans les études d'observation, exercent souvent une influence plus forte, mais aussi plus sournoise sur les résultats. Seule la répétition d'études confirmatoires, chacune arrivant à contrôler différents types de biais et présentant des résultats convergents,

permettra de confirmer ou d'infirmer l'hypothèse. N'a-t-il pas fallu plusieurs études confirmatoires pour établir le lien causal entre la cigarette et le cancer du poumon ?

La plausibilité d'une hypothèse repose sur plusieurs types de jugement : le jugement de validité, le jugement statistique et le jugement scientifique. Le premier type de jugement tient compte des erreurs systématiques et doit permettre de répondre à une question comme celle-ci : certains biais (de sélection, d'information ou de confusion) ne pourraient-ils pas expliquer la conformité ou la non-conformité des résultats avec l'hypothèse ? Le deuxième type de jugement tient essentiellement compte de l'erreur aléatoire et doit permettre de répondre à une question comme celle-ci : les simples fluctuations aléatoires ne pourraient-elles pas expliquer la non-conformité des résultats avec l'hypothèse ? Le troisième type de jugement tient compte des connaissances scientifiques sur le sujet et essaie de répondre à une question comme celle-ci : la non-conformité des résultats avec l'hypothèse est-elle explicable au plan scientifique par la non-congruence de l'hypothèse elle-même avec les connaissances ?

Ces différents types de jugement sont complémentaires, sans que l'un n'ait vraiment préséance sur les autres. Il pourrait très bien arriver qu'une association soit déclarée significative au plan statistique tout en étant entièrement explicable par l'existence d'un biais ou, encore, sans qu'elle n'ait vraiment d'intérêt au plan scientifique ou clinique. À l'inverse, on pourrait observer une association intéressante au plan de la validité, interprétable au plan de la causalité, mais non significative au plan statistique en raison du manque de puissance de l'étude. Le jugement statistique fait essentiellement appel aux lois des probabilités ; le jugement de validité implique des connaissances méthodologiques sur le plan de la qualité des stratégies et instruments d'observation ; le jugement scientifique repose sur les connaissances du sujet a priori.

Alors que les études exploratoires sont généralement de type descriptif, les études confirmatoires peuvent être conduites dans deux contextes différents : le contexte décisionnel et le contexte explicatif.

2.2 LE TEST STATISTIQUE

On peut considérer le test statistique comme une procédure qui permet soit de mesurer la vraisemblance d'une hypothèse, soit d'établir un choix entre deux hypothèses. La première attitude renvoie au contexte non décisionnel, exploratoire ou explicatif, où l'investigateur utilise l'information pour

mieux comprendre un phénomène ou tenter de l'expliquer. La seconde attitude relève d'un contexte plutôt décisionnel, où l'investigateur utilise l'information pour prendre une décision.

La construction d'un test statistique suppose donc au départ l'énoncé d'une hypothèse. Le plus souvent, l'hypothèse spécifiée est l'hypothèse nulle, notée H_0. Par hypothèse nulle, on ne veut pas forcément dire que le paramètre est nul, égal à zéro, mais plutôt que sa valeur correspond à l'affirmation d'une propriété que l'on veut contester, à la négation d'une association ou d'une relation que l'on veut démontrer. Cette contrepartie de l'hypothèse nulle est dite la contre-hypothèse, notée H_1 (l'hypothèse alternative, si on calque la terminologie anglaise). Par exemple, dans l'expérience du lancer d'une pièce de monnaie, partant de la supposition que la pièce est bien équilibrée, l'hypothèse nulle pourrait être que la probabilité d'observer pile est égale à ½ ($\pi = ½$) ; la contre-hypothèse serait alors que cette probabilité π diffère de ½. Dans une étude cas-témoins visant à identifier une association entre un facteur et une maladie, l'hypothèse nulle stipule la négation de cette association : l'hypothèse nulle correspond ainsi à un rapport de cotes égal à 1, la contre-hypothèse à un rapport de cotes différent de 1.

Ainsi, paradoxalement, ce n'est pas l'hypothèse nulle qu'il nous intéresse de démontrer, mais plutôt sa contre-hypothèse, stipulant l'existence d'une relation, d'une association ou d'une différence. À cette fin, on utilise une logique analogue à celle utilisée en justice : l'accusé est considéré innocent tant et aussi longtemps que les faits rapportés ne viennent pas en contradiction avec cette présumée innocence. L'hypothèse de la non-existence d'une relation est retenue tant et aussi longtemps que les faits observés ne viennent pas en contradiction avec cette hypothèse. Si on parle de la présomption d'*innocence* dans le langage judiciaire, on peut parler de la présomption d'*absence d'association* dans le langage statistique.

Dans d'autres circonstances et de façon plus générale, l'hypothèse à vérifier postule une valeur théorique : « la probabilité d'obtenir face est de 50 % », « le risque relatif est de 2 », « la différence entre les deux taux est de 5 pour 1000 », etc. Cette hypothèse est retenue si les faits observés lui sont compatibles ; autrement, elle est rejetée. Dans tous les cas, l'énoncé de l'hypothèse à vérifier se traduit dans une valeur numérique postulée pour un paramètre d'une loi de distribution à laquelle les données doivent se conformer : loi normale (la plus connue), loi du χ^2, loi binomiale, loi de Poisson, etc.

Enfin, au plan opérationnel, on peut distinguer deux types d'hypothèses à tester : les hypothèses unilatérales, les hypothèses bilatérales. L'hypothèse est dite unilatérale si son énoncé postule une direction bien

définie pour sa contre-hypothèse. Elle est dite unilatérale à droite si sa négation (la contre-hypothèse) stipule les valeurs à droite, à gauche si sa contre-hypothèse stipule les valeurs à gauche de son paramètre. Par exemple, l'hypothèse H_0 dont l'énoncé correspond à « la probabilité π d'observer pile n'est pas supérieure à ½ » s'oppose à la contre-hypothèse « la valeur π est supérieur à ½ ». Cette hypothèse H_0 est dite unilatérale à droite puisque son rejet (donc l'acceptation de H_1) postule les valeurs de π plus grandes que ½. On comprend bien qu'en inversant les relations de l'un et l'autre des énoncés précédents, on définit une hypothèse unilatérale à gauche. Par ailleurs, l'hypothèse sera dite bilatérale si sa contre-hypothèse stipule les valeurs de droite et de gauche. L'énoncé « que la probabilité π d'observer pile soit égale à ½ » décrit une hypothèse bilatérale.

L'hypothèse à vérifier ne saurait être reconnue vraie ou fausse ni a priori, ni a posteriori. Dans la pratique des tests statistiques, il n'existe que de la vraisemblance. Si, confrontée aux observations, une hypothèse est jugée vraisemblable, au-delà d'un certain seuil, elle sera rejetée dans le contexte décisionnel, en deçà de ce seuil, elle ne sera pas rejetée. Dans le contexte explicatif, sa vraisemblance pourra être jugée en des degrés divers : peu, assez ou très vraisemblable. Une des mesures les plus utilisées pour la vraisemblance, c'est la valeur-*p* (en anglais, *p-value*), que nous présentons dans la section suivante.

2.3 LA VALEUR-*P*

Si on admet que l'hypothèse vérifiée est vraie et que seule l'erreur aléatoire explique les fluctuations, la valeur-*p* mesure la probabilité, sous cette hypothèse, d'obtenir un résultat au moins aussi extrême que celui observé. La valeur-*p* obtenue est interprétée, suivant le contexte, à des fins décisionnelles ou comme mesure de vraisemblance.

Dans le cadre décisionnel, la valeur-*p* est comparée à un seuil de signification, désigné a priori. Si la valeur-*p* est inférieure à ce seuil, l'hypothèse est rejetée ; autrement elle ne l'est pas. Le rejet de l'hypothèse (nulle) conduit à l'acceptation de sa négation, la contre-hypothèse. Le résultat de la décision statistique s'ajoute aux autres critères, pratiques et scientifiques, sur lesquels se fonde la décision ultime dans le choix entre deux hypothèses. Par exemple, pour un seuil $\alpha = 0{,}05$, une valeur-*p* de 0,058 conduit au non-rejet de l'hypothèse nulle. Par ailleurs, les résultats convergents de plusieurs études confirmatoires sur une même hypothèse sont souvent plus intéressants, de signification plus forte, que ceux d'une seule étude conduisant à une valeur-*p* plus faible.

Dans le cadre explicatif ou exploratoire, la valeur-*p* se présente plutôt comme une mesure de la compatibilité entre les observations et l'hypothèse. De ce point de vue, l'hypothèse mise à l'épreuve est jugée sous l'aspect de sa vraisemblance. Par exemple, un test qui conduit à une valeur-*p* de 0,058 traduit une faible conformité des résultats avec l'hypothèse en cause, une faible vraisemblance de cette hypothèse.

2.4 DEUX FONCTIONS DE PARAMÈTRE

Dans le contexte des tests d'hypothèse, la valeur-*p* mesure la compatibilité des résultats avec l'hypothèse testée, le plus souvent l'hypothèse nulle. Mais, cette valeur-*p* ne dit rien de la compatibilité des résultats avec d'autres hypothèses qui pourraient être elles aussi considérées. Il peut être intéressant d'étendre la notion de valeur-*p* à celle de valeur-*p*-fonction (ou simplement *p*-fonction) : la valeur-*p* est fonction de l'hypothèse testée.

À ce concept de *p*-fonction se rattache celui de fonction de vraisemblance. Alors que, pour une hypothèse donnée, la *p*-fonction correspond à la probabilité d'observer des résultats aussi ou plus extrêmes que celui observé, la fonction de vraisemblance mesure la probabilité (ou la densité de probabilité) sous cette hypothèse, d'observer le résultat tel qu'il fut observé. Cette dernière fonction est assez régulièrement utilisée pour l'estimation des paramètres sous le critère dit « du maximum de vraisemblance ».

Nous allons donner une définition opérationnelle de ces deux fonctions. À cette fin, considérons une certaine variable X qui obéit à une loi de distribution de paramètre θ. On désigne par x et par $X(\theta)$ respectivement la valeur observée de X et sa valeur attendue.

2.4.1 LA *P*-FONCTION $P(\theta)$

La *p*-fonction $p(\theta)$ se définit comme :

$$p(\theta) = \begin{cases} P(X \geq x \mid \theta) \text{ si } x > X(\theta) \\ P(X \leq x \mid \theta) \text{ autrement} \end{cases}$$

Pour une valeur fixe de θ, la valeur $p(\theta)$ correspond alors à la valeur-*p* unilatérale à droite si $x > X(\theta)$ ou unilatérale à gauche si $x \leq X(\theta)$. On remarque alors que $p(\theta)$ définit une fonction, non de X, mais de θ ; le résultat x de X est une condition fixée par les données. Le plus souvent, la fonction $p(\theta)$ atteint son maximum lorsque θ est tel que $X(\theta) = x$.

2.4.2 LA FONCTION DE VRAISEMBLANCE $FV(\theta)$

La fonction de vraisemblance se définit comme $FV(\theta) = P(X = x \mid \theta)$.

Pour une valeur fixe de θ, la valeur $FV(\theta)$ décrit alors la probabilité d'observer le résultat x de X. Dans la description de cette fonction, le résultat x est une condition fixe et le paramètre θ est la variable indépendante. Le plus souvent, la fonction de vraisemblance $FV(\theta)$ atteint aussi son maximum lorsque θ est tel que $X(\theta) = x$.

EXEMPLE 2.1

Pour bien concrétiser les deux notions de *p*-fonction et de fonction de vraisemblance, et pour marquer leur différence, nous utilisons un exemple basé sur une variable binomiale X. La variable X décrit le nombre d'individus ayant le groupe sanguin O+ dans un échantillon de 200 individus prélevés au hasard d'une large population. Dans cet échantillon, on observe 60 individus O+. Alors, $X \mapsto \text{Bin}(\pi,200)$. On s'interroge sur le paramètre π.

La valeur de $p(\pi)$ correspond à la valeur-*p* unilatérale (à droite ou à gauche) calculée sous l'hypothèse π. Dans le tableau 2.1 sont décrites quelques valeurs de π et les valeurs de $p(\pi)$ et $FV(\pi)$ correspondantes. Si on suppose que $\pi = 0{,}20$, alors la valeur-*p*, donc celle de $p(\pi)$, correspondant à 60 succès parmi 200 essais sera de 0,00028. Par ailleurs, sous la même hypothèse, la probabilité $FV(\pi)$ d'observer précisément la valeur $X = 60$ est de 0,00022.

TABLEAU 2.1

π	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
$p(\pi)$	0,00000	0,00000	0,00028	0,04539	0,53481	0,07830	0,00213	0,00001	0,00000
$FV(\pi)$	0,00000	0,00000	0,00022	0,01708	0,06146	0,01993	0,00082	0,00001	0,00000

Pour la description graphique de $p(\pi)$ et de $FV(\pi)$, consultez respectivement les figures 2.1 et 2.2. (**PR2.1** et **PR2.2**)

◆

FIGURE 2.1

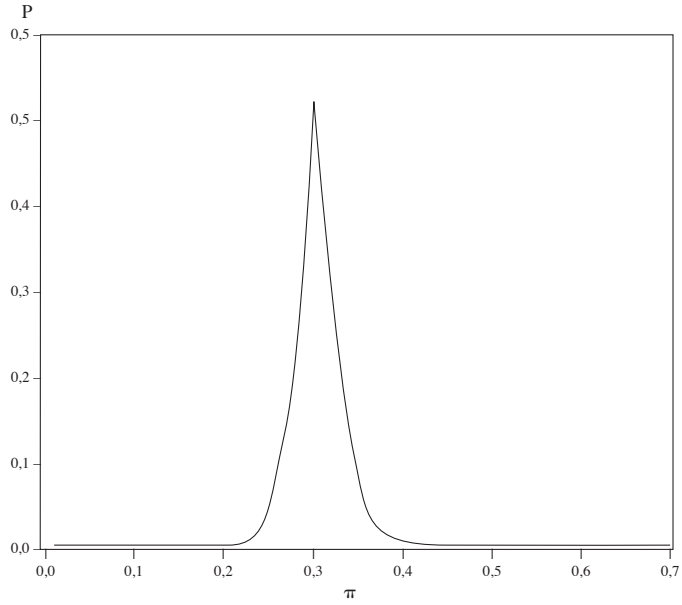
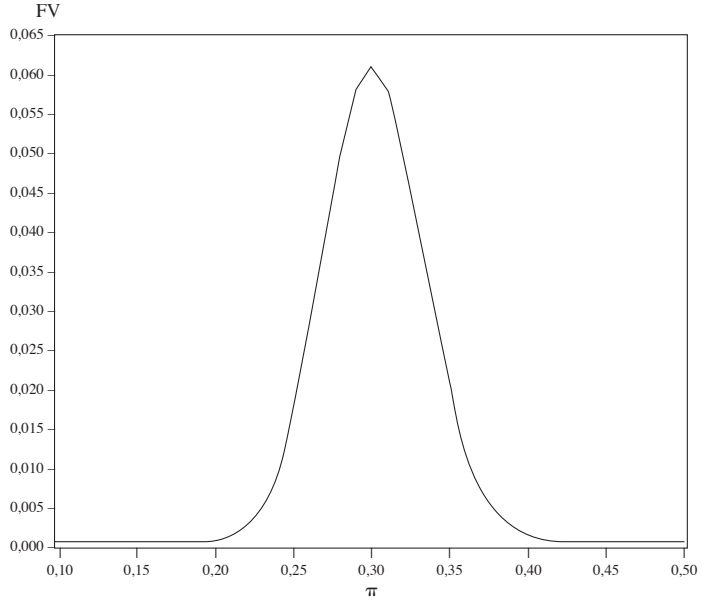


FIGURE 2.2



2.5 INTERVALLE DE CONFIANCE

Considérons la variable X qui obéit à une loi de distribution de paramètre θ, dont la valeur numérique est inconnue. Pour estimer ce paramètre, on utilise les observations faites à partir d'un échantillon. L'estimation retenue coïncide en général avec la valeur $\hat{\theta}$ qui maximise la *p*-fonction (section 2.4.1) ou la fonction de vraisemblance (section 2.4.2). Cette valeur $\hat{\theta}$ est dite l'estimation ponctuelle de θ. On rappelle que la valeur-*p* mesure la compatibilité entre les observations et l'hypothèse testée (section 2.3).

On définit l'intervalle de confiance de niveau $100(1 - \alpha)\%$ comme l'ensemble des valeurs de θ qui sont compatibles avec les données suivant un certain seuil α fixé. Par exemple, une valeur θ_1 sera dite compatible avec les données si elle conduit à une valeur-*p* supérieure à α. En termes opérationnels, sont compatibles au seuil α toutes les valeurs θ_g à gauche de $\hat{\theta}$ telles que $p(\theta_g) = P(X \geq x \mid \theta_g) > \alpha/2$, et toutes les valeurs θ_d à droite de $\hat{\theta}$ telles que $p(\theta_d) = P(X \leq x \mid \theta_d) > \alpha/2$. Les limites de cet ensemble ou intervalle de valeurs θ compatibles avec les données au seuil α sont dites les limites de confiance à $100(1 - \alpha)\%$ de θ. Les limites de confiance inférieure θ_{inf} et supérieure $\theta_{sup,}$ sont alors facilement décrites par les énoncés suivants :

> θ_{inf} *est la plus grande des valeurs de* θ_g *à gauche de* $\hat{\theta}$ *telles que* $p(\theta_g) \leq \alpha/2$
>
> θ_{sup} *est la plus petite des valeurs de* θ_d *à droite de* $\hat{\theta}$ *telles que* $p(\theta_d) \leq \alpha/2$

Comme on le voit, la définition des limites de confiance θ_{inf} et θ_{sup} est en lien direct avec $p(\theta)$.

2.6 VALEUR-*P* ET INTERVALLE DE CONFIANCE

Même considérée comme mesure de la vraisemblance d'une hypothèse, la valeur-*p* ne renseigne pas sur l'importance de la différence ou de la mesure d'association retrouvée dans l'étude. Par exemple, une petite valeur-*p* peut correspondre aussi bien à une faible qu'à une forte mesure d'association observée. Ainsi, une association forte observée sur de faibles échantillons ou une association faible observée sur de grands échantillons peuvent l'une et l'autre s'avérer statistiquement significatives, sans que les valeurs-*p* correspondantes ne traduisent cette différence de contextes.

Exemple 2.2

Soient deux études cas-témoins qui s'intéressent à l'association entre une maladie Y et un facteur d'exposition X. L'une comporte de larges effectifs (tableau 2.2) et l'autre de faibles effectifs (tableau 2.3).

Tableau 2.2

	$X = 1$	$X = 0$	Total
$Y = 1$	600	400	1000
$Y = 0$	500	500	1000
$RC = 1{,}5$, $\chi^2 = 20{,}20$, $p < 0{,}0001$			

Tableau 2.3

	$X = 1$	$X = 0$	Total
$Y = 1$	45	5	50
$Y = 0$	25	25	50
$RC = 9$, $\chi^2 = 19{,}05$, $p < 0{,}0001$			

Supposons que les tests statistiques présentent les mêmes valeurs de khi-carré et donc les mêmes valeurs-p, toutes deux inférieures à 0,001. Dans les deux cas, la valeur-p permet alors de conclure à une association fortement significative. Mais la première étude, conduite sur de larges échantillons, donne un RC de 1,5 et la seconde, sur de faibles échantillons, un RC de 9, avec des intervalles de confiance à 95 % respectivement de [1,26 ; 1,79] et [3,06 ; 26,44]. De chacun de ces intervalles, on conclut qu'il existe une association significative entre Y et X puisque les intervalles excluent la valeur 1 de non-association. Par ailleurs, la stabilité (ou la précision) de la mesure RC dans la première étude est plus grande que dans la seconde, les largeurs des intervalles de confiance à 95 % étant respectivement de 0,53 (1,79 – 1,26) et de 23,38 (26,44 – 3,06).

◆

De l'exemple précédent, on retient que l'intervalle de confiance est un outil statistique qui permet à la fois de présenter une estimation de la mesure et un jugement sur sa stabilité, tout en s'avérant analogue à un test statistique. Concrètement, il permet de répondre aux trois types de questions auxquelles l'analyse statistique essaie de répondre en général :

1. Quelle est la force de l'association ?
2. Cette association est-elle significative ?
3. Quelle est la stabilité de la mesure d'association ?

Quant au test statistique, il ne permet de répondre qu'à la deuxième question. À ce sujet, il est important de rappeler que le résultat d'un test statistique ne peut pas être pris comme une mesure d'association. Ainsi, dans les exemples traités précédemment, les rapports de cotes mesurent les forces ou intensités d'association entre le facteur *X* et la maladie *Y*, alors que les valeurs des χ^2 renvoient à la concordance entre les valeurs observées et les valeurs attendues sous l'hypothèse à l'étude.

2.7 CALCUL EXACT, CALCUL APPROXIMATIF ET CALCUL DE VRAISEMBLANCE

Que le cadre soit décisionnel ou pas, que les études soient exploratoires ou confirmatoires, les modalités de calcul de la valeur-*p* ou de l'intervalle de confiance demeurent les mêmes. Ces calculs s'exécutent le plus souvent suivant une procédure exacte, approximative ou de vraisemblance. Rappelons que la valeur-*p* d'un test correspond à la probabilité d'obtenir un résultat au moins aussi extrême que celui observé, en supposant que l'hypothèse à l'étude soit correcte. Un résultat est considéré aussi ou plus extrême que celui observé si son écart à la moyenne est au moins aussi grand que celui de la valeur observée.

Les calculs des tests ou des intervalles de confiance sont dits exacts s'ils sont exécutés directement à partir des fonctions de probabilité de base, sans passer par l'intermédiaire d'une loi d'approximation. Si les lois de base sont des lois discrètes, telles les lois binomiale, de Poisson et hypergéométrique, alors les calculs des probabilités sur des échantillons de grande taille peuvent être complexes, longs et fastidieux. On utilise alors des méthodes approximatives plus simples. La plus courante est l'approximation normale ; elle trouve sa justification dans le théorème de la limite centrale. On peut aussi mener les calculs au moyen de la fonction de vraisemblance. Par ailleurs, depuis et avec l'avènement des calculateurs rapides, les méthodes exactes prennent une place de plus en plus grande pour ces calculs jusqu'à un certain point, elles sont préférables dans leurs résultats aux méthodes approximatives ou de vraisemblance.

2.8 FORMULES GÉNÉRALES POUR LES TESTS

Pour une certaine variable X dont la loi de distribution comporte le paramètre θ, on s'intéresse à une valeur particulière θ_0 de ce paramètre. À partir d'une série d'observations faites sur X, on s'interroge sur la compatibilité de l'hypothèse θ_0 avec les données observées. Le test statistique

s'avère une procédure de calcul de la valeur-*p* sous l'hypothèse θ_0 spécifiée. Mathématiquement, on pourrait dire qu'il s'agit d'évaluer $p(\theta)$ au point $\theta = \theta_0$ (section 2.4.1).

2.8.1 FORMULES GÉNÉRALES POUR UN TEST EXACT

Lors d'un test exact, le calcul de la valeur-*p* peut se faire suivant certaines conventions qui traduisent la façon de traiter la probabilité de la valeur observée dans le calcul de la valeur-*p*. Les deux conventions généralement utilisées sont la convention intégrale et la convention mi-*p*. La première propose de compter la *valeur intégrale* de cette probabilité ; la deuxième propose de compter la *demi-valeur* de cette probabilité. Sans entrer dans l'épistémologie, nous présentons brièvement la logique qui sous-tend chacune de ces conventions et les conséquences pratiques qui en découlent.

CONVENTION DE LA VALEUR-*P* INTÉGRALE

Pour le calcul d'une valeur-*p* unilatérale, la convention intégrale propose d'effectuer la somme de la probabilité $P(X = x)$ de la valeur observée x et des probabilités de toutes les valeurs plus extrêmes que x. Ainsi calculée, la valeur-*p* est une véritable probabilité. Traditionnellement, cette valeur-*p* constitue un étalon pour les tests statistiques.

Cependant, cette convention s'adapte mal à la définition des tests bilatéraux. À la limite, elle peut conduire à des incohérences, comme par exemple une valeur-*p* supérieure à 1. Par ailleurs, pour obtenir une meilleure concordance entre cette valeur-*p* étalon et celles calculées par des méthodes approximatives, il faut appliquer à ces dernières une correction de continuité, la plus connue étant celle de Yates.

CONVENTION MI-*P*

La convention de la demi-valeur-*p* (qu'on appellera valeur « mi-*p* ») propose d'effectuer la somme de la demi-probabilité $\frac{1}{2}P(X = x)$ de la valeur observée x et des probabilités de toutes les valeurs plus extrêmes que x. Ainsi calculée, la valeur-*p* n'est pas une véritable probabilité. Cependant, elle a certains avantages qui nous la font préférer : elle rend les résultats des tests exacts davantage concordants avec ceux des tests approximatifs et du test du rapport de vraisemblance sans qu'on ait à appliquer des corrections pour ces derniers, et elle ne conduit en aucun cas à l'incohérence d'une probabilité supérieure à 1 pour les tests bilatéraux.

Dans la plupart des exemples numériques qui, dans cet ouvrage, accompagnent la présentation des tests et des intervalles de confiance exacts, c'est la convention mi-*p* qui sera utilisée même si, à l'occasion pour simplifier l'écriture, la valeur-*p* peut être présentée suivant la convention intégrale.

Test exact pour une hypothèse unilatérale

Sous l'hypothèse H_0 que le paramètre $\theta \leq \theta_0$, la valeur-*p* unilatérale à droite se définit simplement comme :

$$P(X \geq x \mid \theta_0) = \sum_{i=x}^{x_M} P(X = i \mid \theta_0) \text{ dans la convention intégrale,}$$

$$P(X \geq x \mid \theta) = 1/2\, P(X = x \mid \theta_0) + \sum_{i=x+1}^{x_M} P(X = i \mid \theta_0) \text{ dans la convention mi-}p.$$

De façon analogue, la valeur-*p* unilatérale à gauche se définit comme :

$$P(X \leq x \mid \theta_0) = \sum_{i=x_m}^{x} P(X = i \mid \theta_0) \text{ dans la convention intégrale}$$

$$P(X \leq x \mid \theta) = 1/2\, P(X = x \mid \theta_0) + \sum_{i=x_m}^{x-1} P(X = i \mid \theta_0) \text{ dans la convention mi-}p.$$

Les valeurs x_m et x_M représentent respectivement le minimum et le maximum de la variable X.

Pour une variable $X \mapsto \text{Pois}(\lambda)$, $x_m = 0$ et $x_M = \infty$

Pour une variable $X \mapsto \text{Bin}(\pi, n)$, $x_m = 0$ et $x_M = n$.

Pour une variable $X \mapsto \text{Hypr}(N, M_1, n, \Psi)$, $x_m = \max(0, n + M_1 - N)$ et $x_M = \min(n, M_1)$.

La valeur-*p* est unilatérale à droite ou à gauche suivant la latéralité de l'hypothèse testée.

Test exact pour une hypothèse bilatérale

Dans le cadre d'un test exact, le calcul de la valeur-*p* bilatérale n'est pas direct et relève le plus souvent d'attitudes. Mais, pour nous, la définition du test bilatéral exact se base sur la notion de valeurs extrêmes. Une valeur est au moins aussi extrême que la valeur observée si sa probabilité est au plus égale à celle de la valeur observée. Ainsi, pour une valeur x observée,

u est une valeur aussi ou plus extrême que x si $P(X = u|\theta_0) \leq P(X = x|\theta_0)$. La valeur-$p$ bilatérale est alors donnée par $p = \sum_u P(X = u \mid \theta_0)$ pour toutes les valeurs u de X (y compris x) également ou plus extrêmes que x. Dans la convention mi-p, la demi-probabilité est appliquée aux valeurs également extrêmes à x.

2.8.2 FORMULES GÉNÉRALES POUR UN TEST APPROXIMATIF NORMAL

Le test approximatif normal est généralement basé sur la comparaison entre la valeur θ_0 du paramètre suggéré par l'hypothèse à l'étude et son estimation $\hat{\theta}$ obtenue sur un échantillon de taille n (ou une série de n observations). On désigne par V_0 la variance des valeurs échantillonnales de θ, sous cette même hypothèse. Alors, suivant le théorème de la limite centrale (section 1.5), la variable centrée réduite $Z = \dfrac{\hat{\theta} - \theta_0}{\sqrt{V_0}}$ obéit en bonne approximation à une loi normale N(0,1). Assez souvent sous l'hypothèse nulle, soit directement ou par transformation de la variable, la valeur θ_0 du paramètre est réduite à 0. Dans ces conditions, Z se réduit à $\dfrac{\hat{\theta}}{\sqrt{V_0}}$.

La conduite du test se fait simplement en remplaçant dans l'expression de Z la valeur observée de $\hat{\theta}$. La valeur numérique z de Z obtenue est interprétée à partir de la loi de distribution normale.

Dans le contexte où la variance est stable, c'est-à-dire qu'elle ne varie pas avec l'hypothèse, ou peut remplacer en bonne approximation la variance V_0 par son estimation du maximum de vraisemblance V. La statistique considérée est alors $\dfrac{\hat{\theta}}{\sqrt{V}}$. On parle alors du test de Wald.

En raison du lien qui existe entre les deux lois, ce test de la loi normale peut très bien se convertir en un test du khi-carré : $X^2 \mapsto X^2_1$.

2.8.3 FORMULE GÉNÉRALE POUR UN TEST DU RAPPORT DE VRAISEMBLANCE

La vraisemblance d'une hypothèse sur θ correspond à la probabilité d'observer les données telles qu'elles se présentent, en supposant cette hypothèse correcte. La fonction $FV(\theta)$ qui décrit cette probabilité est la *fonction de vraisemblance*.

Le test du rapport de vraisemblance pour l'hypothèse $\theta = \theta_0$ est basé sur la comparaison des logarithmes de la fonction de vraisemblance évaluée à θ_0 et à θ_M, valeur de θ où la fonction atteint son maximum. Plus précisément, si on désigne $\log[FV(\theta_0)]$ et $\log[FV(\theta_M)]$ respectivement par $L(\theta_0)$ et L, alors le test du rapport de vraisemblance se présente comme :

$$\chi_1^2 = 2[L - L(\theta_0)]$$

De façon générale, la valeur $2[L - L(\theta)]$ correspond à la déviance entre la valeur du logarithme de la fonction de vraisemblance évaluée à θ et celle du logarithme de la fonction de vraisemblance évaluée en son maximum.

2.9 FORMULES GÉNÉRALES POUR LES INTERVALLES DE CONFIANCE

Le paramètre θ, inconnu, est estimé à partir d'un échantillon : $\hat{\theta}$. On s'intéresse alors à l'intervalle construit à partir de $\hat{\theta}$, regroupant l'ensemble des valeurs possibles de θ qui, suivant un niveau de signification α prescrit, sont compatibles avec les données observées sur l'échantillon. En pratique, cet intervalle a $100(1 - \alpha)$ % des chances de recouvrir la vraie valeur du paramètre θ, qui demeure toujours inconnu.

2.9.1 FORMULES GÉNÉRALES POUR UN INTERVALLE DE CONFIANCE EXACT

L'intervalle de confiance exact est celui dont les limites sont calculées à partir de la loi de distribution de base. Il suffit d'appliquer cette loi aux relations qui définissent les limites de confiance décrites à la section 2.5.

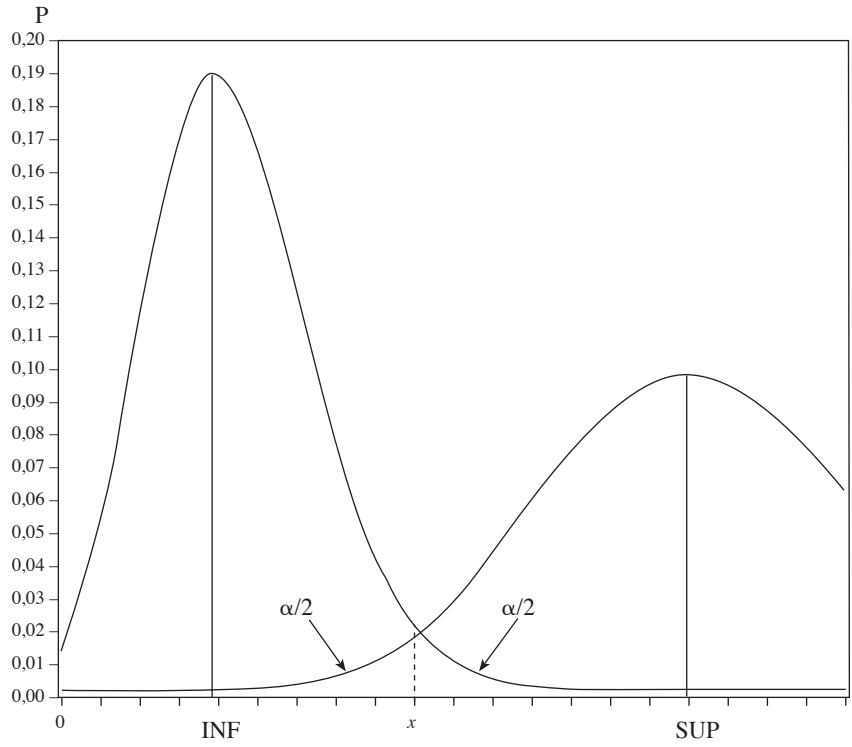
Si X désigne une variable obéissant à une loi de paramètre θ et si on a fait l'observation $X = x$, alors les limites inférieure θ_{inf} et supérieure θ_{sup} de l'intervalle de confiance de niveau $100(1 - \alpha)$ % pour θ sont respectivement calculées comme suit :

θ_{inf} = la plus grande des valeurs de θ_g telles que $P(X \geq x \mid \theta_g) \leq \alpha/2$

θ_{sup} = la plus petite des valeurs de θ_d telles que $P(X \leq x \mid \theta_d) \leq \alpha/2$

On peut consulter la figure 2.3 (**PR2.3**). Dans cet exemple, la première courbe est centrée sur INF et représente la distribution de X sous l'hypothèse $\theta = \theta_{inf}$. L'aire sous la portion de courbe supérieure à x est égale à $\alpha/2$. De même, la courbe centrée sur SUP représente la distribution de X sous l'hypothèse $\theta = \theta_{sup}$. L'aire sous la portion de courbe inférieure à x est aussi égale à $\alpha/2$.

FIGURE 2.3



2.9.2 FORMULE GÉNÉRALE POUR UN INTERVALLE DE CONFIANCE APPROXIMATIF

Les intervalles de confiance approximatifs sont eux aussi le plus souvent calculés dans le cadre de l'approximation normale. On peut la définir de façon analogue aux intervalles de confiance exacts.

Supposons que l'on s'intéresse à un paramètre (mesure) θ dont la valeur observée est $\hat{\theta}$. Alors, pour un niveau $100(1 - \alpha)\,\%$, les limites de confiance inférieure θ_{inf} et supérieure θ_{sup} sont définies comme précédemment.

En appliquant la transformation centrée réduite à la « variable » θ, les deux relations deviennent :

θ_{inf} = la plus grande des valeurs θ_g telles que

$$P\left(\frac{\theta - \theta_g}{\sqrt{V_g}} \geq \frac{\hat{\theta} - \theta_g}{\sqrt{V_g}}\right) \leq \alpha / 2$$

θ_{sup} = la plus petite des valeurs θ_d telles que

$$P\left(\frac{\theta - \theta_d}{\sqrt{V_d}} \leq \frac{\hat{\theta} - \theta_d}{\sqrt{V_d}}\right) \leq \alpha / 2$$

où V_g et V_d désignent les variances de θ respectivement sous les hypothèses $\theta = \theta_g$ et $\theta = \theta_d$.

En utilisant les propriétés d'une variable centrée réduite, de la première relation on peut dire que θ_{inf} (la plus grande valeur des θ_g) est telle que :

$$\frac{\hat{\theta} - \theta_{inf}}{\sqrt{V_{inf}}} = z_{\alpha/2} \text{ ce qui se traduit dans la relation : } \theta_{inf} = \hat{\theta} - z_{\alpha/2}\sqrt{V_{inf}}$$

où $V_{inf} = V(\theta_{inf})$.

De même, de la seconde relation, on peut dire que θ_{sup} (la plus petite valeur des θ_d) est telle que :

$$\frac{\hat{\theta} - \theta_{sup}}{\sqrt{V_{sup}}} = -z_{\alpha/2} \text{ ce qui se traduit dans la relation : } \theta_{sup} = \hat{\theta} + z_{\alpha/2}\sqrt{V_{inf}}$$

où $V_{sup} = V(\theta_{sup})$.

Si on suppose que la variance de θ est indépendante de θ et constante, alors on posera : $V_{inf} = V_{sup} = V$. Ainsi, dans le cadre de l'approximation normale et sous la condition de la variance stable, l'intervalle de confiance à un niveau de $100(1 - \alpha)\,\%$ pour la mesure θ sera donné par : $IC : \theta \pm z_{\alpha/2}\sqrt{V}$.

Si la variance n'est pas constante, alors le calcul de l'intervalle devient théoriquement plus problématique. On peut utiliser certaines transformations de la mesure θ qui permettent de stabiliser la variance. Par exemple, la transformation RAC permet de stabiliser la variance d'une variable de Poisson et la transformation ARC, celle d'une proportion. En l'absence de telles transformations, l'utilisation de la variance observée peut être en général une option acceptable. On parle alors de la méthode de Wald.

Il existe aussi des méthodes itératives qui permettent de solutionner pour θ l'équation quadratique

$$\frac{\left[\hat{\theta}-\theta\right]^2}{V(\theta)}=\chi^2_{1,1-\alpha}$$

où $\hat{\theta}$ est la valeur observée et $100(1-\alpha)$ % le niveau de confiance visé. On suppose que $V(\theta)$ s'exprime en fonction de θ, au plus comme une fonction quadratique. Il suffit alors de retrouver les 2 zéros du polynôme quadratique $f(\theta)=\theta^2-\chi^2_{1,1-\alpha}V(\theta)-2\hat{\theta}\theta+\hat{\theta}^2$. Le plus petit zéro constitue la limite inférieure et le plus grand, la limite supérieure de l'intervalle.

Pour un niveau de confiance à 95 %, les limites inférieure θ_{inf} et supérieure θ_{sup} sont respectivement telles que :

$$\frac{\left(\hat{\theta}-\theta_{inf}\right)}{\sqrt{V_{inf}}}=1,96 \text{ et } \frac{\left(\hat{\theta}-\theta_{sup}\right)}{\sqrt{V_{sup}}}=1,96$$

2.9.3 FORMULE GÉNÉRALE DE L'INTERVALLE DE CONFIANCE PAR LE RAPPORT DE VRAISEMBLANCE

On veut calculer l'intervalle de confiance de la mesure θ. On rappelle que θ_M est l'estimateur du maximum de vraisemblance du paramètre θ. Les limites supérieure et inférieure de l'intervalle doivent satisfaire l'équation logarithmique $2[L-L(\theta)]-\chi^2_{1,1-\alpha}=0$, où L désigne la valeur de la fonction $L(\theta)$ en son maximum, c'est-à-dire $L = L(\theta_M)$, et $L(\theta)$ désigne la valeur de la fonction pour un θ quelconque. Il s'agit alors de solutionner pour θ l'équation de la forme : $2[L-L(\theta)]-\chi^2_{1,1-\alpha}=0$.

La plus petite et la plus grande valeur de θ qui satisfont cette équation sont alors respectivement la limite inférieure θ_{inf} et la limite supérieure θ_{sup} de l'intervalle. La résolution se fait généralement par procédures itératives.

2.10 ESTIMATION DE LA VARIANCE PAR LA MÉTHODE DELTA

L'estimation de la variance de certaines mesures peut parfois s'avérer difficile. On peut donner l'exemple des mesures qui s'expriment sous forme de rapport (rapport de proportions ou de taux), ou encore celui d'une transformation de telles mesures (logarithme d'un rapport). Pour estimer la variance de ces mesures qui posent problème, il existe une méthode assez fortement utilisée en biostatistique : la méthode delta, qui repose sur le développement en série de Taylor de la mesure transformée.

2.10.1 VARIANCE D'UNE MESURE SIMPLE

Supposons que l'on s'intéresse à une mesure m et à sa transformation $Y = \Phi(m)$. On suppose que la transformation Φ est continue et dérivable. Alors, le développement en séries de Taylor de Φ autour de la moyenne μ de m permet d'établir une relation simple entre la variance $V[\Phi(m)]$ de la transformation $\Phi(m)$, et celle $V(m)$ de la mesure m. Voici comment.

On considère le développement en séries de Taylor de $\Phi(m)$. En première approximation, ce développement conduit à l'expression suivante pour la fonction $\Phi(m)$:

$$\Phi(m) \cong \Phi(\mu) + \left.\frac{\partial \Phi(m)}{\partial m}\right|_{\mu} \times (m - \mu) = \Phi(\mu) + \Phi'(\mu)\,(m - \mu)$$

Dans cette expression, μ est la moyenne paramétrique, $\Phi(\mu)$ et $\Phi'(\mu)$ sont les valeurs de la fonction Φ et de sa dérivée évaluée à μ ; ces trois valeurs sont donc des constantes. On établit alors que :

$$V\left[\Phi(m)\right] = \left[\Phi'(\mu)\right]^2 V(m) \tag{2.1}$$

2.10.2 MESURE NÉCESSITANT UNE TRANSFORMATION LOGARITHMIQUE : Φ = LOG

Pour un rapport de taux, de proportions ou de cotes, on utilise fréquemment la transformation logarithmique. Dans ce cas, suivant la relation (2.1), la variance de la mesure transformée se présente comme :

$$V\left[\log(m)\right] = \left[\frac{1}{\mu}\right]^2 V(m)$$

Si μ n'est pas connu, ce qui est généralement le cas, on propose de le remplacer par m. La relation devient simplement :

$$V\left[\log(m)\right] = \left[\frac{1}{m}\right]^2 V(m)$$

2.10.3 VARIANCE D'UNE MESURE PONDÉRÉE EN ANALYSE STRATIFIÉE

On s'intéresse à une mesure m qui, en analyse stratifiée, est la somme des mesures spécifiques m_i, pondérées par les poids λ_i. Les poids λ_i sont tels que $0 < \lambda_i < 1$ et $\sum_i \lambda_i = 1$. On parlera de m comme d'une mesure pondérée ou résumée.

On suppose devoir appliquer à la mesure m une transformation Φ pour stabiliser sa variance ou pour contrer les problèmes d'asymétrie de sa distribution.

Alors, sous l'hypothèse de l'indépendance des strates et de la fixité des poids λ_i, on peut établir que :

$$V\left[\Phi(m)\right] \approx \sum_i \left\{ \frac{[\Phi'(m)]^2 \lambda_i^2 V[\Phi(m_i)]}{[\Phi'(m_i)]^2} \right\} \tag{2.2}$$

En effet, aux conditions posées et en faisant appel à la méthode delta, à partir de l'expression (2.1) on établit que

$$\begin{aligned} V\left[\Phi(m)\right] &= \left[\Phi'(m)\right]^2 V(m) \\ &= \left[\Phi'(m)\right]^2 V\left(\sum_i \lambda_i m_i\right) \\ &= \left[\Phi'(m)\right]^2 \sum_i \lambda_i^2 V(m_i) \end{aligned} \tag{2.3}$$

En appliquant cette fois la méthode delta à la variance $V(m_i)$, on peut l'exprimer comme une fonction de $V[\Phi(m_i)]$:

$$V(m_i) = V\left[\Phi(m_i)\right] \Big/ \left[\Phi'(m_i)\right]^2 \tag{2.4}$$

En remplaçant dans l'expression (2.3) la valeur de $V(m_i)$ décrite en (2.4), on obtient l'expression (2.2).

Finalement, dans l'hypothèse de l'uniformité des mesures spécifiques et en utilisant le fait que Φ est une transformation continue et dérivable, l'expression de la variance $V[\Phi(m)]$ se réduit à

$$V\left[\Phi(m)\right] = \sum_i \lambda_i^2 V\left[\Phi(m_i)\right] \tag{2.5}$$

puisque l'uniformité des m_i implique $m_i \approx m$ et le caractère continu et dérivable de Φ implique $\Phi'(m_i) \approx \Phi'(m)$.

2.10.4 ESTIMATION DE LA VARIANCE D'UNE MOYENNE DE PUISSANCE *K*

Soient n estimations d'une même mesure μ : m_1, m_2, … m_n, pondérées suivant λ_1, λ_2, …, λ_n. On définit la moyenne de puissance k de ces n estimations comme :

$$m^{[k]} = \left(\sum_i \lambda_i m_i^k\right)^{\frac{1}{k}}$$

En particulier, $m^{[k]}$ est la moyenne arithmétique si $k = 1$, la moyenne quadratique si $k = 2$ et la moyenne harmonique si $k = -1$.

La variance de $m^{[k]}$ peut être calculée de la façon suivante.

Si on pose $s_k = \sum_i \lambda_i m_i^k$, alors la transformation $\Phi = (\quad)^{\frac{1}{k}}$ appliquée à s_k donne $m^{[k]}$: $\Phi\left(s_k\right) = m^{[k]}$.

Ainsi, de l'expression (2.5), on a :

$$V(m^{[k]}) = V\left[\Phi\left(s_k\right)\right] = \sum_i \lambda_i^2 V\left[\Phi\left(m_i^k\right)\right] = \sum_i \lambda_i^2 V\left(m_i\right)$$

On remarque que la variance d'une moyenne de puissance k est indépendante de cette puissance k.

2.11 RELATION FONDAMENTALE ENTRE DEUX FONCTIONS DE DENSITÉ

Il peut s'avérer plus facile de calculer l'intervalle de confiance de $\Phi(m)$ que celui de m. Si Φ est une transformation monotone continue, croissante ou décroissante, alors, de l'intervalle de confiance sur $\Phi(m)$, on déduira celui sur m par simple transformation inverse $\Phi^{-1}[\Phi(m)]$.

Rappelons qu'une transformation Φ est monotone croissante si $X_1 \leq X_2$ implique $\Phi(X_1) \leq \Phi(X_2)$; elle est décroissante si $X_1 \leq X_2$ implique $\Phi(X_1) \geq \Phi(X_2)$.

Établissons d'abord la relation fondamentale suivante entre les fonctions de densité (ou fonctions de répartition) de deux variables aléatoires X et Y, dont l'une est une transformation monotone continue de l'autre : $Y = \Phi(X)$.

Désignons par $f(X)$ la fonction de densité de X et par $g(Y)$ celle de Y. Alors, si Φ est monotone croissante, on a $\int_a^b f(x)dx = \int_{\Phi(a)}^{\Phi(b)} g(y)dy$, ce qui se traduit en termes de probabilités par :

$$P\left[a \leq X \leq b\right] = P\left[\Phi(a) \leq Y \leq \Phi(b)\right] \tag{2.6}$$

Si Φ est monotone décroissante, on a $\int_a^b f(x)dx = \int_{\Phi(b)}^{\Phi(a)} g(y)dy$, ce qui se traduit en termes de probabilités par :

$$P\left[a \leq X \leq b\right] = P\left[\Phi(b) \leq Y \leq \Phi(a)\right] \tag{2.7}$$

De la relation (2.6) découle l'expression de l'intervalle de confiance de m pour une transformation Φ monotone croissante et de la relation (2.7) découle celle de l'intervalle de confiance de m pour une transformation Φ monotone décroissante

Φ monotone croissante : $m_{\text{inf}} = \Phi^{-1}[\Phi(m)_{\text{inf}}]$

$m_{\text{sup}} = \Phi^{-1}[\Phi(m)_{\text{sup}}]$

Φ monotone décroissante : $m_{\text{inf}} = \Phi^{-1}[\Phi(m)_{\text{sup}}]$

$m_{\text{sup}} = \Phi^{-1}[\Phi(m)_{\text{inf}}]$

Par exemple, à partir des limites de confiance d'une proportion p, on peut assez facilement déduire celles de sa proportion complémentaire q si on comprend que q est une transformation monotone décroissante de p : $q = \Phi(p) = 1 - p$.

De cette transformation dite complémentaire de p, on déduit que

$$q_{\text{inf}} = (1 - p_{\text{sup}}) \qquad (2,8)$$
$$q_{\text{sup}} = (1 - p_{\text{inf}})$$

Ainsi, en connaissant les limites de confiance de p, on connaît aussi celle de q et vice versa.

PARTIE

2

ANALYSE SIMPLE

CHAPITRE 3

LES TAUX

Pour une maladie donnée, le taux d'incidence (ou de décès) renvoie au nombre de cas incidents observés dans un groupe d'individus en tenant compte de la période de temps durant laquelle chaque individu a été considéré à risque. Nous présentons les tests statistiques permettant de comparer cette mesure à une valeur paramétrique et l'intervalle de confiance permettant de la décrire.

L'analyse d'un taux se fait dans le cadre de la loi de Poisson. Tant pour les tests que pour les intervalles de confiance, on s'intéresse alors à une variable X correspondant au nombre de succès pour n expériences indépendantes et identiques de Poisson, chacune de paramètre λ. La variable X est elle-même une variable de Poisson de paramètre $\mu = n\lambda$: $X \mapsto \text{Pois}(\mu)$.

À la suite des n expériences de Poisson, on observe $X = x$ et on mesure le taux $t = x/n$. Ce taux observé est une estimation du paramètre τ, l'intensité du processus de Poisson : on se rappelle que $\lambda = \tau ds$ (chapitre 1, section 1.3.4). (Dans le reste du texte, il nous arrivera régulièrement d'assimiler λ à τ, sachant que λ est le numérateur du taux τ mesuré sur l'expérience élémentaire ds de Poisson.)

3.1 TESTS SUR UN TAUX

On suppose $\tau = \tau_0$ (qui peut représenter aussi les hypothèses $\tau \leq \tau_0$ ou $\tau \geq \tau_0$). Sous cette hypothèse, la valeur attendue $E(X)$ et la variance $V(X)$ de X sont précisément égales à $n\tau_0$ ou μ_0. On désignera parfois cette valeur par X_0.

S'inrterroger sur la compatibilité entre le résultat $X = x$ et l'hypothèse $\tau = \tau_0$ revient à juger ce résultat dans le cadre des fluctuations de X autour de X_0. Pour ce jugement, on propose trois tests statistiques : un test exact, un test en approximation normale et le test du rapport de vraisemblance.

3.1.1 TEST EXACT SUR UN TAUX

Le test unilatéral à droite qui doit nous conduire à un jugement sur la compatibilité entre l'observation $X = x$ et l'hypothèse $\mu \leq \mu_0 = n\tau_0$, se décrit simplement comme :

$$\begin{aligned} p &= P(X \geq x \mid \mu_0) \\ &= \frac{e^{-\mu_0}\mu_0^x}{2x!} + \sum_{i=x+1}^{\infty} \frac{e^{-\mu_0}\mu_0^i}{i!} \\ &= 1 - \sum_{i=0}^{x-1} \frac{e^{-\mu_0}\mu_0^i}{i!} - \frac{e^{-\mu_0}\mu_0^x}{2x!} \end{aligned}$$

dans la convention mi-p.

Au besoin, le test unilatéral à gauche peut être défini de façon analogue, la sommation étant faite de 0 à x.

Le test bilatéral se décrit comme :

$$p = \sum_{u \neq x} \frac{e^{-\mu_0}\mu_0^u}{u!} + 1/2 \frac{e^{-\mu_0}\mu_0^x}{x!}$$

où u est toute observation sur X telle que $P(X = u) \leq P(X = x)$.

De façon analogue, on peut définir le test unilatéral à gauche :

$$p = P(X \leq x \mid \mu_0)$$
$$= \frac{e^{-\mu_0}\mu_0^x}{2x!} + \sum_{i=0}^{x-1} \frac{e^{-\mu_0}\mu_0^i}{i!}$$

3.1.2 TEST EN APPROXIMATION NORMALE SUR UN TAUX

Sous l'hypothèse $\tau = \tau_0$, la variable de Poisson X a comme valeur attendue et variance $n\tau_0$: $E(X) = X_0 = n\tau_0$ et $V(X) = X_0$.

Ainsi, la variable transformée réduite $\frac{X - E(X)}{\sqrt{V(X)}}$ obéit approximativement à une variable normale Z de moyenne 0 et d'écart-type 1 :

$$\frac{X - E(X)}{\sqrt{V(X)}} = \frac{X - X_0}{\sqrt{X_0}} \approx Z$$

Pour appliquer le test, il suffit alors de remplacer X par sa valeur observée : $X = x$.

$$\frac{x - X_0}{\sqrt{X_0}} = z$$

Pour la valeur z correspondante, on détermine à l'aide d'une table de la loi normale la probabilité $p = P(Z \geq z)$ pour un test unilatéral à droite, $P(Z \leq z)$ pour un test unilatéral à gauche ou $2P(Z \geq |z|)$ pour un test bilatéral.

Une variante du test est obtenue en appliquant l'approximation normale au taux t. Dans ce cas, $E(t) = \tau_0$ et $V(t) = \frac{\tau_0}{n}$. Par approximation normale, on a

$$\frac{t - \tau_0}{\sqrt{\frac{\tau_0}{n}}} \approx Z$$

Pour appliquer le test, il suffit alors de remplacer t par sa valeur observée : $t = x/n$.

Ce test est identique au précédent.

On peut aussi présenter ce test dans le cadre de la loi du khi-carré :

$$\frac{(t-\tau_0)^2}{\frac{\tau_0}{n}} = \frac{(O-A)^2}{A} = \chi_1^2$$

Dans la deuxième expression, la valeur O représente la valeur observée ($O = x$) et A la valeur attendue sous l'hypothèse τ_0 ($A = n\tau_0$).

Ce test en approximation normale n'est applicable que si $n\tau_0 \geq 5$.

3.1.3 TEST DU RAPPORT DE VRAISEMBLANCE POUR UN TAUX

Si X est une variable de paramètre μ, alors pour l'observation $X = x$, la fonction de vraisemblance se définit comme : $FV(\mu) = \frac{e^{-\mu}\mu^x}{x!}$, et son logarithme $L(\mu)$ comme $L(\mu) = -\mu + x\log(\mu) - \log(x!)$.

Le test du rapport de vraisemblance est alors simplement défini par la comparaison des valeurs $L(x)$ et $L(\mu_0)$ de la fonction $L(\mu)$ évaluée pour les hypothèses $\mu = \mu_0$ et $\mu = x$, cette dernière étant suggérée par les données. La statistique $2[L(x) - L(\mu_0)]$ générée par cette comparaison obéit à une loi du χ^2 avec un degré de liberté.

Pour la conduite du test, cette statistique se décrit simplement comme :

$$2[L(x) - L(\mu_0)] = 2\left[-x + x\log(x) - \log(x!) + \mu_0 - x\log(\mu_0) + \log(x!)\right]$$
$$= 2\left[O\log\left[\frac{O}{A}\right] - (O-A)\right]$$

où $O = x$, la valeur observée, et $A = \mu_0$, la valeur attendue suggérée par l'hypothèse.

Exemple 3.1

Dans une région donnée, on a observé 9 décès par la maladie *M* pour 12 000 personnes-années. Cette observation est-elle compatible avec un taux théorique de 5 décès par 10 000 personnes-années ?

Test exact

On a l'hypothèse $\tau_0 = \dfrac{5 \text{ décès}}{10\,000 \text{ personnes-années}} = \dfrac{0{,}0005 \text{ décès}}{1 \text{ personne-année}}$.

Puisque $ds = 1$ personne-année, on a $\lambda_0 = \tau_0 \times ds = 0{,}0005$ décès ou 0,0005.

L'hypothèse d'un taux égal à 5 décès par 10 000 personnes-années correspond à l'hypothèse d'un paramètre $\lambda_0 = 0{,}0005$ pour l'expérience élémentaire de Poisson. Pour 12 000 expériences de Poisson identiques et indépendantes, le paramètre μ_0 ou X_0 de la nouvelle variable X de Poisson sera : $\mu_0 = X_0 = 12\,000 \times 0{,}0005 = 6$. Dans ce contexte, on recherche la probabilité $P(X \geq 9 \mid X_0 = 6)$.

La valeur-*p* unilatérale à droite est donnée par :

$$
\begin{aligned}
p &= P(X \geq 9 \mid X_0 = 6) \\
&= 1/2 \frac{e^{-6} 6^9}{9!} + \sum_{x=10}^{\infty} \frac{e^{-6} 6^x}{x!} \\
&= 1/2 \frac{e^{-6} 6^9}{9!} + \left(1 - \sum_{x=0}^{9} \frac{e^{-6} 6^x}{x!} \right) \\
&= 0{,}11834
\end{aligned}
$$

Ces données sont relativement compatibles avec l'hypothèse. En effet, dans plus de 12 pour cent des cas, on pourrait observer au moins 9 décès de *M* parmi 12 000 personnes-années en supposant que le taux réel de décès de *M* dans la région soit de 5 décès pour 10 000 personnes-années.

Pour le test bilatéral, on observe que $P(X = 9) = 0{,}068838489$. De plus, la valeur $X = 2$ s'avère la plus grande valeur x de X, dans la partie latérale gauche de la distribution, telle que $P(X = x) \leq P(X = 9)$: $P(X = 2) = 0{,}0446175392$. Le test bilatéral pourra donc être calculé comme :

$$
\begin{aligned}
p &= P(X \leq 2 \mid X_0 = 6) + P(X \geq 9 \mid \mu = 6) \\
&= \sum_{i=0}^{2} \frac{e^{-6} 6^i}{i!} + 0{,}11834 \\
&= 0{,}18031
\end{aligned}
$$

Test en approximation normale

L'hypothèse testée est $\mu_0 = 12\,000 \times 0{,}0005 = 6$. La valeur observée de X est 9.

Le test unilatéral donne une valeur $z = \dfrac{9 - 6}{\sqrt{6}} = 1{,}2247$ qui correspond à la valeur-*p* de $P(Z \geq 1{,}2247) = 0{,}11034$. Cette valeur est assez près de celle obtenue par le test exact.

La valeur-*p* du test bilatéral est simplement donnée par $P(|Z| \geq 1{,}2247) = 2 \times 0{,}110344 = 0{,}22067$. Cette valeur-*p* bilatérale est sensiblement différente de celle obtenue par le test exact : 0,22069 versus 0,18031.

Dans le cadre de la loi du khi-carré, le test se présente comme

$$\chi^2_1 = \frac{(O-A)^2}{A} = \frac{(9-6)^2}{6} = 1{,}5$$

On remarque que $1{,}5 = 1{,}2247^2$.

Test du rapport de vraisemblance

La valeur observée de *X* est 9 ($O = 9$) et la valeur attendue sous l'hypothèse d'un taux $\tau_0 = 0{,}0005$ est 6 ($A = 6$).

Le test du rapport de vraisemblance est alors :

$$\begin{aligned}\chi^2_1 &= 2\left[O\log\left(\frac{O}{A}\right) - (O-A)\right] \\ &= 2\left[9\log\left(\frac{9}{6}\right) - (9-6)\right] \\ &= 1{,}2984\end{aligned}$$

La valeur-*p* correspondante est de 0,2545.

Dans le tableau 3.1, nous résumons les résultats des différents tests utilisés dans l'exemple.

Tableau 3.1

Méthode	Valeur du test	Valeur-*p* (bilatérale)	Programme SAS
Exacte	–	0,1803	PR3.1
Khi-carré	1,5	0,2207	PR3.2
RV	1,2984	0,2545	PR3.3

Ces résultats, quel que soit le test, reflètent une bonne compatibilité entre les observations et l'hypothèse d'un taux égal à 5 décès pour 10 000 personnes-années. Les deux tests approximatifs concordent assez bien.

◆

3.2 INTERVALLES DE CONFIANCE POUR UN TAUX

Pour n essais indépendants de Poisson, on observe $X = x$. Le taux $t = x/n$ est une estimation de τ. On veut déterminer l'intervalle de confiance à $100(1 - \alpha)\%$ de ce taux, où α désigne le risque d'erreur de première espèce.

3.2.1 INTERVALLE DE CONFIANCE EXACT POUR UN TAUX

Dans l'approche exacte, les limites de confiance inférieure τ_{inf} et supérieure τ_{sup} du taux sont décrites respectivement comme suit :

τ_{inf} est la plus grande des valeurs de τ telles que

$$\sum_{x=a}^{\infty} \frac{e^{-n\tau}(n\tau)^x}{x!} \leq \alpha / 2 \tag{3.1}$$

τ_{sup} est la plus petite des valeurs de τ telles que

$$\sum_{x=0}^{a} \frac{e^{-n\tau}(n\tau)^x}{x!} \leq \alpha / 2 \tag{3.2}$$

Par itération sur τ, on obtient assez facilement les limites supérieure et inférieure.

3.2.2 INTERVALLE DE CONFIANCE EN APPROXIMATION NORMALE POUR UN TAUX

Nous suggérons trois méthodes de calcul en approximation normale pour les limites de confiance d'un taux : la méthode simple, la méthode utilisant la transformation « racine-carrée » et la méthode quadratique.

MÉTHODE SIMPLE, SANS TRANSFORMATION

La variance du taux t mesuré sur un échantillon de taille n est simplement donnée par $V(t) = t/n$. Ainsi, par approximation normale, on obtient directement les limites de confiance suivantes :

$$\tau_{\text{inf}} = t - z_{\alpha/2}\sqrt{\frac{t}{n}}$$
$$\tau_{\text{sup}} = t + z_{\alpha/2}\sqrt{\frac{t}{n}}$$

Transformation RAC

Lorsque l'approximation normale fait défaut, principalement lorsque *nt* est inférieur à 5, le taux *t* peut être transformé à l'aide de la racine carrée : $t \to \sqrt{t}$. La variance de $\sqrt{t}$ ne dépend que de *n* : $V\left(\sqrt{t}\right) = \frac{1}{4n}$. La variance est stabilisée.

L'intervalle de confiance est alors calculé à partir de la racine carrée de *t*. Puis par transformation inverse, on obtient celui de τ :

$$\tau_{\text{inf}} = \left(\sqrt{t} - \frac{z_{\alpha/2}}{2\sqrt{n}}\right)^2$$
$$\tau_{\text{sup}} = \left(\sqrt{t} + \frac{z_{\alpha/2}}{2\sqrt{n}}\right)^2$$

Méthode quadratique

L'intervalle de confiance peut aussi être calculé sans condition sur la variance. Il s'agit de solutionner pour τ l'équation quadratique de la forme : $\frac{[(t-\tau)]^2}{\frac{\tau}{n}} = \chi^2_{1,1-\alpha}$. On rappelle que *x*, *n* (donc *t*) et $z_{a/2}$ sont des valeurs connues. Cela revient alors à résoudre pour τ l'équation quadratique suivante : $f(\tau) = n\tau^2 - (2x + \chi^2_{1,1-\alpha})\tau + xt = 0$

Les méthodes en approximation normale sont applicables en autant que *x* soit égal ou supérieur à 5.

3.2.3 INTERVALLE DE CONFIANCE PAR LA MÉTHODE DU RAPPORT DE VRAISEMBLANCE POUR UN TAUX

Pour le calcul de l'intervalle de confiance d'un taux τ par la méthode du rapport de vraisemblance, on considère le logarithme de la fonction de vraisemblance $L(\tau)$ de la variable X de Poisson. On rappelle que t est l'estimateur du maximum de vraisemblance du paramètre τ. Les limites supérieure et inférieure de l'intervalle doivent satisfaire l'équation logarithmique suivante : $2[L - L(\tau)] - \chi^2_{1,1-\alpha} = 0$, où L désigne la valeur de la fonction $L(\tau)$ en son maximum (c'est-à-dire lorsque $\tau = t$) et $L(\tau)$ désigne la valeur de la fonction pour la valeur τ. La plus petite et la plus grande valeur de τ qui satisfont cette équation correspondent alors aux limites recherchées : τ_{inf} et τ_{sup}.

La résolution peut se faire facilement par procédures itératives. On rappelle que la fonction de vraisemblance a la forme :
$L(\tau) = x\log(n\tau) - n\tau - \log(x!)$ et sa valeur maximale L :
$L = x\log(x) - x - \log(x!)$

EXEMPLE 3.2

Revenons à l'exemple 3.1. On a $n = 12\,000$ et $X = 9$. Le taux observé est de $\tau = 0{,}00075$.

INTERVALLE DE CONFIANCE EXACT AU NIVEAU 95 %

Alors, par itération avec une précision de 0,0000001, on obtient à partir des formules (3.1) et (3.2) les limites suivantes :

$$\tau_{inf} = 0{,}00036575$$

$$\tau_{sup} = 0{,}00137640$$

APPROXIMATION NORMALE

MÉTHODE SIMPLE

On veut calculer l'intervalle de confiance à 95 % en approximation normale. Dans ce cas, $z_{\alpha/2} = 1{,}96$.

En utilisant simplement l'approximation normale, on a :

$$\tau_{inf} = 0{,}00075 - 1{,}96\sqrt{\frac{0{,}00075}{12000}}$$
$$= 0{,}00026$$

$$\tau_{sup} = 0{,}00075 + 1{,}96\sqrt{\frac{0{,}00075}{12000}}$$
$$= 0{,}00124$$

Méthode basée sur la transformation RAC

La méthode utilisant la transformation racine-carrée nous conduit aux résultats suivants :

$$\sqrt{t} = \sqrt{0{,}00075} = 0{,}0274 \text{ et } V(S) = \frac{1}{4 \times 12\,000}$$

Ainsi,

$$\left(\sqrt{t}\right)_{\text{inf}} = 0{,}0274 - 1{,}96\sqrt{\frac{1}{4 \times 12\,000}} = 0{,}018454$$

et

$$\left(\sqrt{t}\right)_{\text{sup}} = 0{,}0274 + 1{,}96\sqrt{\frac{1}{4 \times 12\,000}} = 0{,}036346$$

On déduit alors les limites de confiance du taux τ :

$$\tau_{\text{inf}} = \left[0{,}018454\right]^2 = 0{,}00034$$
$$\tau_{\text{sup}} = \left[0{,}036346\right]^2 = 0{,}00132$$

Méthode quadratique

Pour la méthode quadratique, il faut résoudre l'équation suivante :

$$12\,000\tau^2 - (2 \times 9 + 3{,}84)\tau + 9 \times 0{,}00075 = 0$$

Les racines de cette équation quadratique correspondent aux limites recherchées :

$$\tau_{\text{inf}} = 0{,}00039459$$
$$\tau_{\text{sup}} = 0{,}00142553$$

Méthode du rapport de vraisemblance

En reprenant la description de la fonction de vraisemblance, l'équation de vraisemblance : $2[L - L(\tau)] - \chi^2_{1,1-\alpha} = 0$ peut s'écrire

$$[L - L(\tau)] - 0{,}5\chi^2_{1,1-\alpha} = x\log\left(\frac{x}{n\tau}\right) - (x - n\tau) - 0{,}5\chi^2_{1,1-\alpha}.$$

Puisque $x = 9$, $n = 12\,000$ et $\chi^2_{1,\,0{,}95} = 3{,}84$, alors l'équation logarithmique à résoudre se présente comme $9\log(\tau) - 12\,000\tau + 75{,}68 = 0$

Les deux racines de cette équation déterminent les limites de confiance de τ, au niveau 95 % :

$$\tau_{\text{inf}} = 0{,}0003574$$
$$\tau_{\text{sup}} = 0{,}0013535.$$

Si la substitution de ces valeurs dans l'équation à résoudre ne donne pas identiquement 0, cela n'est dû qu'aux arrondissements.

Dans le tableau 3.2, nous présentons les intervalles de confiance obtenus par les différentes méthodes de calcul appliquées aux données de l'exemple.

TABLEAU 3.2

Méthode	Intervalle de confiance	Programme
Calcul exact	[0,00037 ; 0,00138]	PR3.4
Approximation simple	[0,00026 ; 0,00124]	–
Transformation RAC	[0,00034 ; 0,00132]	–
Quadratique	[0,00039 ; 0,00143]	PR3.5
Méthode du RV	[0,00036 ; 0,00135]	PR3.6

On remarque que les résultats des méthodes exacte et du rapport de vraisemblance concordent assez bien. La méthode approximative simple est la moins concordante ; il vaut mieux l'éviter à cause de l'instabilité de la variance d'un taux.

◆

CHAPITRE 4

LES PROPORTIONS

En épidémiologie, la mesure des fréquences de base renvoie le plus souvent aux proportions. Pour une maladie donnée, l'incidence cumulative décrit, chez un groupe d'individus à risque, la proportion de ceux qui développent la maladie (ou qui décèdent) sur une période de temps déterminée. La prévalence mesure la proportion des individus qui sont affectés par la maladie à un moment donné dans le groupe à risque considéré.

Qu'il s'agisse d'une prévalence, d'une incidence cumulative ou de toute autre mesure de même nature, l'analyse d'une proportion se fait dans le cadre de la loi binomiale. On s'intéresse alors à une variable X générée par un processus de Bernoulli comportant n essais indépendants de même paramètre π. La variable X représente le nombre de cas ou succès observables sur ces n essais.

La variable X est une variable binomiale de paramètres π et n : $X \mapsto \text{Bin}(\pi,n)$. À la suite des n expériences de Bernoulli, on observe $X = x$ et on mesure la proportion $p = x/n$. Cette proportion observée est une estimation du paramètre π (chapitre 1, section 1.2.4).

Dans ce chapitre, nous présentons les tests statistiques permettant de comparer une proportion observée à une valeur paramétrique et l'intervalle de confiance permettant de la décrire.

4.1 TEST POUR UNE PROPORTION *P*

On suppose $\pi = \pi_0$. Est-ce que l'observation $X = x$ (ou $p = x/n$) est compatible avec l'hypothèse d'une proportion égale à π_0 ?

Pour répondre à cette question, on propose trois tests statistiques : un test exact, un test en approximation normale et le test du rapport de vraisemblance.

4.1.1 TEST EXACT POUR UNE PROPORTION

Le test exact unilatéral à droite, qui doit nous conduire à un jugement sur la compatibilité de l'observation $X = x$ avec l'hypothèse $\pi \leq \pi_{0,}$ se décrit simplement comme :

$$p = P(X \geq x \mid \pi_0) = 1/2\, C_n^x \pi_o^x (1-\pi_0)^{n-x} + \sum_{i=x+1}^{n} C_n^i\ \pi_0^i (1-\pi_0)^{n-i}$$

suivant la convention de la valeur mi-p.

Le test bilatéral, portant sur l'hypothèse $\pi = \pi_0$, se décrit comme :

$$\begin{aligned} p &= P(X \leq u \mid \pi_0) + P(X \geq x \mid \pi_0) \\ &= 1/2\, C_n^x \pi_0^x (1-\pi_0)^{n-x} + \sum_{i=x+1}^{n} C_n^i\ \pi_0^i (1-\pi_0)^{n-i} + \sum_{i=0}^{u} C_n^i\ \pi_0^i (1-\pi_0)^{n-i} \end{aligned}$$

où u est la plus grande des valeurs de X dans la partie latérale gauche de la distribution de X telle que $P(X = u) \leq P(X = x)$. Si u est tel que $P(X = u) = P(X = x)$, alors on applique la convention mi-p à $P(X = u)$. Cependant, une telle coïncidence arrive rarement sauf si $\pi = 0{,}5$. Dans ce cas de distribution symétrique, la valeur-p bilatérale s'obtient simplement en doublant la valeur-p unilatérale.

De façon analogue, on peut définir le test unilatéral à gauche :

$$p = P(X \leq x \mid \pi_0) = 1/2\, C_n^x \pi_o^x (1-\pi_0)^{n-x} + \sum_{i=0}^{x-1} C_n^i\ \pi_0^i (1-\pi_0)^{n-i}$$

4.1.2 TEST EN APPROXIMATION NORMALE POUR UNE PROPORTION

Sous l'hypothèse $\pi = \pi_0$, la variable binomiale X a comme valeur attendue $E(X) = n\pi_0$ et comme variance $V(X) = n\pi_0(1 - \pi_0)$. Ainsi, la variable centrée réduite $Z = \dfrac{X - E(X)}{\sqrt{V(X)}}$ obéit approximativement à une variable normale de moyenne 0 et d'écart-type 1. En remplaçant $E(X)$ et $V(X)$ par leurs valeurs respectives, on a :

$$\frac{X - E(X)}{\sqrt{V(X)}} = \frac{X - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}} \approx Z$$

Pour appliquer le test, il suffit alors de remplacer X par sa valeur observée : $X = x$.

$$\frac{x - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}} = z$$

Pour la valeur z correspondante, on détermine à l'aide d'une table de la loi normale la probabilité $p = P(Z \geq z)$ pour un test unilatéral à droite, $P(Z \leq z)$ pour un test unilatéral à gauche ou $2P(Z \geq |z|)$ pour un test bilatéral.

Une autre variante du test consiste à utiliser l'approximation normale pour une proportion. Si p désigne la proportion d'intérêt, alors $E(p) = \pi_0$ et $V(p) = \dfrac{\pi_0(1-\pi_0)}{n}$. Par approximation normale, on a

$$\frac{p - \pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}} \approx Z$$

Pour appliquer le test, il suffit alors de remplacer p par sa valeur observée : $p = x/n$.

$$\frac{\frac{x}{n} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = z$$

En se rappelant que le carré Z^2 d'une variable $Z \mapsto N(01)$ obéit à une loi du khi-carré à 1 degré de liberté, on peut présenter le test précédent dans la forme suivante :

$$\frac{(p-\pi_0)^2}{\frac{\pi_0(1-\pi_0)}{n}} = \sum \frac{(O-A)^2}{A} = \chi_1^2$$

Dans la deuxième expression, la sommation est faite sur les deux cellules, x et $n - x$. Pour chacune de ces cellules, O représente la valeur observée et A la valeur attendue sous l'hypothèse π_0.

Ce test en approximation normale n'est applicable que si $n\pi_0$ et $n(1 - \pi_0) \geq 5$.

4.1.3 TEST DU RAPPORT DE VRAISEMBLANCE POUR UNE PROPORTION

Ce test est simplement défini par le rapport de la fonction de vraisemblance de l'hypothèse théorique $\pi = \pi_0$ à celle de l'hypothèse $\pi = p$ suggérée par les données.

Pour une variable binomiale X de paramètre π et n, la fonction de vraisemblance se définit par $FV(\pi) = C_n^x \pi^x (1-\pi)^{(n-x)}$. Le log $L(\pi)$ de cette fonction correspond alors à $L(\pi) = x\log(\pi) + (n-x)\log(1-\pi) + \log C_n^x$.

La statistique $2[L(p) - L(\pi_0)]$, qui compare les logarithmes des deux fonctions de vraisemblance $FV(p)$ et $FV(\pi_0)$, sous l'hypothèse π_0, obéit à une loi du χ^2 avec un degré de liberté. Pour la conduite du test, cette statistique se décrit simplement comme :

$$2[L(p) - L(\pi_0)] = 2\left[x\log\left(\frac{p}{\pi_0}\right) + (n-x)\log\left(\frac{1-p}{1-\pi_0}\right)\right]$$

$$= 2\sum O\log\left(\frac{O}{A}\right)$$

Dans la deuxième expression, la sommation se fait sur les deux cellules, X et $n - X$. Pour chacune de ces cellules, O représente la valeur observée et A la valeur attendue sous l'hypothèse π_0.

Exemple 4.1

Dans une région donnée, sur 300 naissances, on observe 165 naissances masculines. Est-ce que cette observation est compatible avec l'hypothèse que la proportion de naissances masculines n'est pas supérieure à 0,51 ?

Test exact

L'hypothèse est $\pi \leq 0{,}51$. De plus, $n = 300$ et $X = 165$ (valeur observée). La valeur-p unilatérale à droite est donnée par :

$$\begin{aligned} p &= P(X \geq 165 \mid \pi = 0{,}51) \\ &= 1/2\, C_{300}^{165}(0{,}51)^{165}(1-0{,}51)^{135} + \sum_{i=166}^{300} C_{300}^{i}(0{,}51)^{i}(1-0{,}51)^{300-i} \\ &= 0{,}0831 \end{aligned}$$

Ces données sont relativement compatibles avec l'hypothèse. En effet, en supposant que la proportion de naissances masculines dans la population générale soit de 0,51, on pourra observer pour 300 naissances un nombre au moins aussi grand que 165 naissances masculines, dans plus de 8 pour cent des cas.

Déterminons maintenant la valeur-p bilatérale.

D'abord, on observe que $P(X = 165) = 0{,}0176689924$. Par ailleurs, dans la partie latérale gauche de la distribution, 141 s'avère la plus grande valeur u de X telle que $P(X = u) \leq P(X = 165)$: $P(X = 141) = 0{,}0176485585$. Le test bilatéral pourra donc être calculé comme :

$$\begin{aligned} p &= P(X \leq 141 \mid \pi = 0{,}51) &+ \quad & P(X \geq 165 \mid \pi = 0{,}51) \\ &= \sum_{i=0}^{141} C_{300}^{i}(0{,}51)^{i}(1-0{,}51)^{300-i} &+ \quad & P(X \geq 165 \mid \pi = 0{,}51) \\ &= 0{,}0920629706 &+ \quad & 0{,}0831310723 \\ &= 0{,}1751940428 \end{aligned}$$

Test en approximation normale

L'hypothèse est $\pi \leq 0{,}51$, $n = 300$ et $X = 165$ (valeur observée). La valeur-p unilatérale à droite est donnée par :

$$z = \frac{165 - 300 \times (0{,}51)}{\sqrt{300 \times (0{,}51)(1-0{,}51)}} = 1{,}3859$$

En utilisant la loi normale, on a : $P(Z \geq 1{,}3859) = 0{,}0828859937$

La valeur du test bilatéral est simplement donnée par :

$$2P(Z \geq 1{,}3859) = 2 \times 0{,}0828859937 = 0{,}1657719874$$

Le khi-carré, test de nature bilatéral, donne le résultat suivant :

$$\chi_1^2 = \sum \frac{(O-A)^2}{A} = \frac{(165-153)^2}{153} + \frac{(135-147)^2}{147} = 1{,}9208$$

On remarque que la valeur du χ^2 est égale à celle du z au carré : 1,9208 = $1{,}3859^2$, en tenant compte de l'arrondissement des valeurs.

Test du rapport de vraisemblance

Bien que l'hypothèse soit $\pi \leq 0{,}51$, le test du rapport de vraisemblance est de nature bilatéral. Le test appliqué aux données de l'exemple se présente comme suit :

$$\begin{aligned}\chi_1^2 &= 2\left[x\log\left(\frac{p}{\pi}\right) + (n-x)\left(\frac{1-p}{1-\pi}\right)\right] \\ &= 2\left[165\log\left(\frac{0{,}55}{0{,}51}\right) + (300-165)\log\left(\frac{1-0{,}55}{1-0{,}51}\right)\right] \\ &= 1{,}9248840758\end{aligned}$$

ou encore

$$\begin{aligned}\chi_1^2 &= 2\sum O\log\left(\frac{O}{A}\right) \\ &= 2\left[165\log\left(\frac{165}{0{,}51\times 300}\right) + 135\log\left(\frac{135}{0{,}49\times 300}\right)\right] \\ &= 1{,}9248840758\end{aligned}$$

Dans le tableau 4.1, nous présentons les résultats des différents tests pratiqués sur les données de l'exemple.

Tableau 4.1

Méthode	Valeur du khi-carré	Valeur *p* (bilatérale)	Programme SAS
Exacte	–	0,1752	PR4.1
Pearson	1,9208	0,1658	PR4.2
RV	1,9249	0,1653	PR4.3

De ces résultats, quel que soit le test, on conclut à une assez bonne compatibilité des observations avec l'hypothèse d'une proportion de naissances masculines égale à 0,51. On remarque, par ailleurs, une bonne concordance entre les résultats de ces différents tests.

◆

4.2 INTERVALLE DE CONFIANCE POUR UNE PROPORTION

On suppose que X est une variable binomiale telle que $X \mapsto Bin(\pi,n)$ où π est un paramètre à déterminer. Pour n essais, on observe $X = x$. La proportion $p = x/n$ est une estimation de π. On veut déterminer l'intervalle de confiance de π à un niveau $100(1 - \alpha)\,\%$ de confiance.

4.2.1 INTERVALLE DE CONFIANCE EXACT POUR UNE PROPORTION

Dans l'approche exacte, les limites de confiance inférieure π_{inf} et supérieure π_{sup} de π sont décrites respectivement comme suit :

π_{inf} est la valeur π qui satisfait l'équation suivante :

$$\frac{C_n^x \pi^x (1-\pi)^{n-x}}{2} + \sum_{i=x+1}^{n} C_n^i \pi^i (1-\pi)^{n-i} = \alpha/2 \tag{4.1}$$

π_{sup} est la valeur π qui satisfait l'équation suivante :

$$\frac{C_n^x \pi^x (1-\pi)^{n-x}}{2} + \sum_{i=0}^{x-1} C_n^i \pi^i (1-\pi)^{n-i} = \alpha/2 \tag{4.2}$$

La rapidité des calculateurs informatiques nous permet d'obtenir facilement et avec une bonne précision ces limites de confiance.

4.2.2 INTERVALLE DE CONFIANCE EN APPROXIMATION NORMALE POUR UNE PROPORTION

Approximation normale simple

Par approximation normale, les limites d'un intervalle de confiance pour une proportion se calculent simplement comme :

$$\pi_{inf} = p - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

$$\pi_{sup} = p + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

Cette méthode est applicable en autant que p ne s'éloigne pas trop de ½ et que les effectifs soient suffisants.

Transformation ARC

Lorsque l'approximation normale fait défaut, principalement lorsque la proportion p est faible ou forte sur des échantillons de petite taille, alors cette proportion p peut être transformée à l'aide de $\arcsin\sqrt{\ }$: $p \rightarrow \arcsin\sqrt{p}$. La variance de $\arcsin\sqrt{p}$ ne dépend plus que de n : $V\left(\arcsin\sqrt{p}\right) = \frac{1}{4n}$, lorsque la transformation est exprimée en radians. La variance est stabilisée.

L'intervalle de confiance est d'abord calculé sur la transformation de p. Puis, par transformation inverse, on obtient celui de π :

$$\pi_{\text{inf}} = \left[\sin\left(\arcsin\sqrt{p} - \frac{z_{\alpha/2}}{2\sqrt{n}}\right)\right]^2$$

$$\pi_{\text{sup}} = \left[\sin\left(\arcsin\sqrt{p} + \frac{z_{\alpha/2}}{2\sqrt{n}}\right)\right]^2$$

Méthode quadratique

L'intervalle de confiance peut aussi être calculé sans condition sur la variance. Il s'agit de solutionner pour π l'équation quadratique de la forme :

$\frac{[(p-\pi)]^2}{\frac{\pi(1-\pi)}{n}} = \chi^2_{1,1-\alpha}$. On rappelle que x, n (donc p) et $\chi^2_{1,1-\alpha}$ sont des valeurs connues. Cela revient alors à solutionner pour π l'équation quadratique suivante : $f(\pi) = (n + \chi^2_{1,1-\alpha})\pi^2 - (2x + \chi^2_{1,1-\alpha})\pi + xp = 0$

Les méthodes en approximation normale sont applicables en autant que x et $n - x$ soient égaux ou supérieurs à 5.

4.2.3 INTERVALLE DE CONFIANCE D'UNE PROPORTION PAR LE RAPPORT DE VRAISEMBLANCE

On veut calculer l'intervalle de confiance de la proportion π sachant que $p = x/n$. À cette fin, on considère le logarithme de la fonction de vraisemblance $L(\pi)$ de la variable binomiale X. La fonction $L(\pi)$ est décrite comme : $L(\pi) = x \log \pi + (n-x)\log(1-\pi) + \log C_n^x$. On rappelle que p est l'estimateur du maximum de vraisemblance du paramètre π ; c'est donc dire que $L(p)$ $(= L)$ est la valeur maximale de $L(\pi)$.

Les limites inférieure π_{inf} et supérieure π_{sup} de l'intervalle correspondent aux valeurs de π qui satisfont l'équation logarithmique $2[L - L(\pi)] - \chi^2_{1,1-\alpha} = 0$. Concrètement, cette équation se présente ici comme : $x \log\left(\frac{x}{n\pi}\right) + (n-x)\log\left(\frac{n-x}{n-n\pi}\right) - 0,5\chi^2_{1,1-\alpha} = 0$

La résolution de cette équation peut se faire facilement par procédures itératives.

EXEMPLE 4.2

Revenons à l'exemple 4.1. On veut déterminer l'intervalle de confiance à 95 % de π à partir des observations : $X = 165$ et $n = 300$. L'estimation p de π est 0,55.

INTERVALLE DE CONFIANCE EXACT

Par itération sur les formules (4.1) et (4.2) décrivant les limites de confiance exactes, on obtient les résultats suivants :

$\pi_{inf} = 0,4934$

$\pi_{sup} = 0,6057$

APPROXIMATION NORMALE

On veut calculer l'intervalle de confiance à 95 % en approximation normale. Dans ce cas, $z_{\alpha/2} = 1,96$.

Méthode simple

Par la méthode utilisant simplement l'approximation normale, on a :

$$\pi_{inf} = 0,55 - 1,96\sqrt{\frac{0,55 \times 0,45}{300}} = 0,4937$$

$$\pi_{sup} = 0,55 + 1,96\sqrt{\frac{0,55 \times 0,45}{300}} = 0,6063$$

Méthode utilisant la transformation ARC

On utilise la transformation en radians.

Alors, $\arcsin\sqrt{0,55} = 0,8355$.

Aussi, $V(\arcsin\sqrt{p}) = \dfrac{1}{4 \times 300}$

Alors :

$$\pi_{inf} = \left[\sin\left(0,8355 - \frac{1,96}{2\sqrt{300}}\right)\right]^2 = 0,4935$$

$$\pi_{sup} = \left[\sin\left(0,8355 + \frac{1,96}{2\sqrt{300}}\right)\right]^2 = 0,6059$$

Méthode quadratique

Pour la méthode quadratique, il faut résoudre l'équation suivante :

$$(n + z^2_{\alpha/2})\pi^2 - (2x + z^2_{\alpha/2})\pi + xp = (300 + 1,96^2)\pi^2 -$$
$$(2 \times 165 + 1,96^2)\pi + 165 \times 0,55 = 0$$

Les deux racines de cette équation quadratique correspondent aux limites de l'intervalle :

$$\pi_{inf} = 0,4934$$

$$\pi_{sup} = 0,6053$$

Méthode du rapport de vraisemblance

On veut calculer l'intervalle de confiance à 95 % en utilisant le maximum de vraisemblance. Dans ce cas, l'équation à résoudre pour π est

$$x\log\left(\frac{x}{n\pi}\right) + (n-x)\log\left(\frac{n-x}{n-n\pi}\right) - 0,5\chi^2_{1,1-\alpha} = 0$$

Pour $\alpha = 0,05$, on a $\chi^2_{1,(1-0,05)} = 3,84$.

Avec les valeurs numériques de x = 165 et n = 300, l'équation à résoudre devient $165\log(\pi) + 135\log(1-\pi) + 208{,}36 = 0$.

Les solutions trouvées sont :

$$\pi_{inf} = 0{,}4935$$

$$\pi_{sup} = 0{,}6057$$

Dans le tableau 4.2, nous présentons les intervalles de confiance obtenus par les différentes méthodes de calcul appliquées aux données de l'exemple.

TABLEAU 4.2

Méthode	Intervalle de confiance	Programme
Calcul exact	[0,4934 ; 0,6057]	PR4.4
Approximation simple	[0,4937 ; 0,6063]	PR4.5
Approximation ARC	[0,4935 ; 0,6059]	PR4.6
Quadratique	[0,4934 ; 0,6053]	PR4.7
Méthode du RV	[0,4935 ; 0,6057]	PR4.8 OU PR4.9

On remarque que les cinq méthodes donnent des résultats similaires. L'importance des effectifs et la bonne symétrie de la distribution ($\pi = 0{,}51$) favorisent cette concordance entre les méthodes.

Sur les échantillons plus faibles et pour un π fort (ou faible), les méthodes exactes et du RV demeurent concordantes, alors que les autres méthodes, particulièrement la méthode simple, présentent parfois d'importantes dissimilitudes. Dans ces situations, il vaut mieux éviter la méthode simple.

◆

PARTIE 3

ANALYSE DANS UN TABLEAU 2 × 2

CHAPITRE 5

LES TAUX DANS UN TABLEAU 2 × 2

Nous considérons les résultats d'une étude de cohortes portant sur l'association entre un facteur d'exposition X et une maladie Y. Cette étude a permis d'observer n_1 et n_0 personnes-temps (en l'occurrence des personnes-années) respectivement chez les sujets exposés et chez les sujets non exposés. Ces personnes-années ont généré respectivement a_1 et a_0 cas de la maladie. En raison de la nature dichotomique de l'événement Y (malade ou non malade) et du facteur d'exposition X (exposé ou non exposé) considérés, les données peuvent être disposées dans un tableau de contingence 2 × 2 (tableau 5.1).

TABLEAU 5.1

	$X = 1$	$X = 0$	Total
$Y = 1$	a_1	a_0	m_1
Total*	n_1	n_0	n

* Les données sont des personnes-temps.

Si t_1 et t_0 représentent respectivement les taux chez les exposés ($X = 1$) et chez les non-exposés ($X = 0$), alors $t_1 = \frac{a_1}{n_1}$ et $t_0 = \frac{a_0}{n_0}$. La différence (*DT*) et le rapport (*RT*) de ces mesures de base sont deux mesures d'association entre *X* et *Y*, simplement définies comme : $DT = t_1 - t_0$ et $RT = \frac{t_1}{t_0}$. (Nous laissons aux épidémiologistes le soin d'interpréter ces mesures d'association.)

Dans les sections suivantes, nous décrivons les tests statistiques et les intervalles de confiance pour chacune de ces mesures d'association. Tant pour les tests statistiques que pour les intervalles de confiance de ces deux mesures, nous présentons d'abord les approches exactes puis les approches approximatives. Ces dernières sont construites à partir de l'approximation normale et par le rapport de vraisemblance. Les tests statistiques seront essentiellement conduits sous l'hypothèse H_0. Dans une dernière section, nous ajoutons une présentation du SMR, mesure d'association basée sur la comparaison d'un taux observé à un taux théorique ou attendu.

5.1 QUELQUES RELATIONS DE BASE ENTRE LES MESURES

Considérons les données du tableau 5.1 recueillies pour étudier l'association entre *X* et *Y*. Les valeurs a_1 et a_0 décrivent les nombres de cas observés respectivement pour les n_1 personnes-temps à risque exposées au facteur *X* et n_0 personnes-temps à risque non exposées à ce facteur.

Nous rappelons que les variables A_1 et A_0, correspondant aux deux cellules du tableau, sont deux variables de Poisson indépendantes, respectivement de paramètres $n_1\tau_1$ et $n_0\tau_0$, où τ_1 et τ_0 sont les taux réels, généralement inconnus. Les taux estimés correspondants sont $t_1 = a_1/n_1$ pour τ_1 et $t_0 = a_0/n_0$ pour τ_0. On a aussi $t = m_1/n$.

Les mesures d'association les plus usitées dans ce contexte sont

- la différence Δ des taux, estimée par $DT = t_1 - t_0$,
- le rapport φ des taux, estimé par $RT = \frac{t_1}{t_0}$.

Sous la condition de m_1 fixe, A_1 est une variable binomiale : $A_1 \mapsto \text{Bin}(\pi, m_1)$.

Sous cette condition, on établit que

$$\pi = \frac{\varphi n_1}{\varphi n_1 + n_0} \tag{5.1}$$

On peut aussi montrer que π a la forme :

$$\pi = \frac{\Delta n_1 n_0 + m_1 n_1}{m_1 n} \tag{5.2}$$

Dans le contexte des marges fixes, il existe une correspondance biunivoque entre la variable A_1 et les mesures Δ et φ. En d'autres termes, à chaque valeur de A_1 ne correspond qu'une et une seule mesure Δ, qu'une et une seule mesure φ. Ces relations, essentiellement croissantes, sont décrites comme suit :

$$\varphi = \frac{A_1 n_0}{(m_1 - A_1) n_1} \tag{5.3}$$

$$\Delta = \frac{A_1 n - m_1 n_1}{n_1 n_0} \tag{5.4}$$

Leurs réciproques sont données par :

$$A_1 = m_1 \left(\frac{\varphi n_1}{\varphi n_1 + n_0} \right) \tag{5.5}$$

$$A_1 = \frac{\Delta n_1 n_0 + m_1 n_1}{n} \tag{5.6}$$

On peut aussi établir des relations biunivoques entre φ et Δ :

$$\varphi = \frac{m_1 + \Delta n_0}{m_1 - \Delta n_1} \tag{5.7}$$

$$\Delta = \frac{(\varphi - 1) m_1}{\varphi n_1 + n_0} \tag{5.8}$$

Ces différentes relations peuvent permettre de passer facilement d'une distribution à l'autre. Ainsi, toute hypothèse sur A_1 trouve une hypothèse correspondante sur φ et sur Δ. Mais aussi, toute hypothèse sur φ ou toute hypothèse sur Δ trouve une hypothèse correspondante sur l'autre mesure et sur la variable A_1.

De façon pratique, l'étude d'un *RT* sous une hypothèse φ revient à considérer la variable binomiale A_1 de paramètre π et m_1 où $\pi = \dfrac{\varphi n_1}{\varphi n_1 + n_0}$. De même, l'étude d'une différence *DT* entre deux taux sous une hypothèse Δ revient à considérer la variable binomiale de paramètres π et m_1, où cette fois $\pi = \dfrac{\Delta n_1 n_0 + m_1 n_1}{m_1 n}$.

De ces relations entre la variable binomiale et les mesures d'association, on peut dire que les tests exacts sous les hypothèses $\Delta = 0$ et $\varphi = 1$ sont équivalents. Il suffit alors de conduire le test exact sur $\varphi = 1$ pour tester aussi l'hypothèse $\Delta = 0$.

5.2 TESTS POUR LA COMPARAISON DE DEUX TAUX

Dans cette section, nous présentons les tests statistiques les plus courants pour la comparaison de deux taux en analyse simple : d'abord un test exact basé sur la distribution binomiale, puis quatre tests en approximation normale et enfin le test du rapport de vraisemblance. La plupart des tests présentés sont assez bien connus et souvent utilisés. Le test basé sur la transformation RAC fait exception. Il est présenté principalement pour son résultat qui peut servir à définir l'intervalle de confiance d'une différence entre deux taux dont les variances doivent être stabilisées.

Enfin, soulignons que les tests de Pearson et de Mantel-Haenszel pour la comparaison de deux taux sont identiques.

5.2.1 TEST EXACT POUR LA COMPARAISON DE DEUX TAUX

Le test exact est directement construit sur la variable binomiale A_1. Sous l'hypothèse nulle $\varphi = 1$ (ou $\Delta = 0$), la variable A_1 est une variable binomiale de paramètres $\pi = \dfrac{n_1}{n_1 + n_0}$ et m_1. Le test unilatéral à droite basé sur la convention mi-*p* est alors de la forme :

$$p = \frac{C_{m_1}^{a_1} \pi^{a_1} (1-\pi)^{m_1 - a_1}}{2} + \sum_{i=a_1+1}^{m_1} C_{m_1}^{i} \pi^{i} (1-\pi)^{(m_1 - i)}$$

De façon analogue, on peut définir le test unilatéral à gauche.

On rappelle aussi que le test bilatéral se définit simplement en ajoutant à la valeur-*p* unilatérale la somme des probabilités de toutes les valeurs de A_1 non considérées et qui sont également ou plus extrêmes que celle observée.

5.2.2 TESTS EN APPROXIMATION NORMALE POUR LA COMPARAISON DE DEUX TAUX

Test de Mantel-Haenszel

Le test de Mantel-Haenszel est construit à partir de l'approximation normale appliquée à la variable binomiale A_1, sous l'hypothèse nulle $\varphi = 1$. Le test se formule alors à partir de la variable Z (ou Z^2) comme

$$Z = \frac{A_1 - E(A_1)}{\sqrt{V(A_1)}} \text{ ou } Z^2 = \frac{\left[A_1 - E(A_1)\right]^2}{V(A_1)}$$ qui obéit à une loi du χ^2 avec un degré de liberté.

Sous H_0, la valeur attendue $E(A_1)$ et la variance $V(A_1)$ de la variable binomiale A_1 correspondent respectivement à :

$$E(A_1) = \frac{m_1 n_1}{n} \text{ et } V(A_1) = \frac{m_1 n_1 n_0}{n^2}$$

En substituant ces valeurs de $E(A_1)$ et $V(A_1)$ dans l'expression de la variable Z et en appliquant le test aux données observées, on obtient :

$$\chi_1^2 = \frac{\left(a_1 - \frac{m_1 n_1}{n}\right)^2}{\frac{m_1 n_1 n_0}{n^2}}$$

$$= \frac{\left(a_1 n_0 - a_0 n_1\right)^2}{m_1 n_1 n_0}$$

Test de Pearson

Considérons la statistique *DT* qui renvoie à la différence entre les deux taux. Alors, la variable *Z* centrée réduite $\frac{DT - E(DT)}{\sqrt{V(DT)}}$ est, en bonne approximation, une variable normale N(0,1). Donc Z^2 obéit en bonne approximation à une loi du χ^2 avec un degré de liberté. Sous l'hypothèse nulle, on a $E(DT) = 0$ et $V(DT) = \frac{m_1}{n}\left(\frac{1}{n_1} + \frac{1}{n_0}\right)$. En substituant ces valeurs respectives de *E*(*DT*) et *V*(*DT*) dans l'expression de la variable *Z* et en appliquant le test aux données observées, on obtient :

$$\chi_1^2 = \frac{\left(\frac{a_1}{n_1} - \frac{a_0}{n_0}\right)^2}{\frac{m_1}{n}\left(\frac{1}{n_1} + \frac{1}{n_0}\right)} = \frac{\left(a_1 n_0 - a_0 n_1\right)^2}{m_1 n_1 n_0}$$

Ainsi, ce test s'avère identique au test de Mantel-Haenszel décrit précédemment.

Dans une forme plus usuelle, on peut décrire ce test comme : $\chi_1^2 = \sum \frac{(O - A)^2}{A}$, où la sommation se fait sur les deux cellules du tableau 5.1 ; pour chacune d'elles, *O* représente la valeur observée et *A* la valeur attendue sous l'hypothèse nulle.

Test sur les taux transformés par RAC

On considère la statistique $DT_T = \sqrt{t_1} - \sqrt{t_0}$, qui renvoie à la différence entre les deux taux transformés par la racine carrée.

La variable Z centrée réduite $\dfrac{DT_T - E(DT_T)}{\sqrt{V(DT_T)}}$ est, en bonne approximation, une variable normale N(0,1), donc Z^2 obéit en bonne approximation à une loi du χ^2 avec un degré de liberté. La variance de DT_T correspond simplement à : $V[DT_T] = \dfrac{1}{4}\left[\dfrac{1}{n_1} + \dfrac{1}{n_0}\right]$. Sous l'hypothèse nulle, on a $E(DT_T) = 0$. En substituant ces valeurs respectives de $E(DT_T)$ et $V(DT_T)$ dans l'expression de la variable Z et en appliquant le test aux données observées, on obtient :

$$\chi_1^2 = \frac{\left[\sqrt{t_1} - \sqrt{t_0}\right]^2}{\dfrac{1}{4}\left[\dfrac{1}{n_1} + \dfrac{1}{n_0}\right]}$$

$$= \frac{4n_1n_0\left[\sqrt{t_1} - \sqrt{t_0}\right]^2}{n}$$

$$= \frac{4\left[\sqrt{a_1n_0} - \sqrt{a_0n_1}\right]^2}{n}$$

On désigne par χ_T^2 la valeur du khi-carré de ce test.

Test sur log(*RT*)

Le test peut facilement se construire en utilisant la transformation logarithmique du rapport des taux : log *RT*.

Sous l'hypothèse nulle, on a $E(\log RT) = 0$ et par la méthode delta, $V[\log RT] = \left[\dfrac{1}{a_1} + \dfrac{1}{a_0}\right]$.

Alors, le test devient : $\chi_1^2 = \dfrac{(\log RT)^2}{V(\log RT)} = \dfrac{(\log RT)^2}{\left(\dfrac{1}{a_1} + \dfrac{1}{a_0}\right)}$

Le *RT* est celui mesuré sur les données. Il faut noter ici que la variance $V(\log RT)$ n'est pas calculée sous l'hypothèse nulle ; elle est celle du maximum de vraisemblance.

5.2.3 TEST DU RAPPORT DE VRAISEMBLANCE POUR LA COMPARAISON DE DEUX TAUX

Test du rapport de vraisemblance

Suivant les notations du tableau 5.1, les paramètres des deux variables de Poisson indépendantes, A_1 et A_0, sont décrits respectivement par $n_1\tau_1$ et $n_0\tau_0$. Alors, sous l'hypothèse nulle que les taux sont égaux ($\tau_1 = \tau_0 = \tau$), la fonction de vraisemblance FV_0 de ces données peut se décrire comme :

$$FV_0 = \left(\frac{e^{-n_1\tau}(n_1\tau)^{a_1}}{a_1!} \right) \times \left(\frac{e^{-n_0\tau}(n_0\tau)^{a_0}}{a_0!} \right)$$

L'hypothèse nulle se résume en un paramètre unique τ qui sera estimé par a/n.

Par ailleurs, sous une hypothèse H spécifiant des valeurs uniques, τ_1 et τ_0, pour chacun des deux taux, la fonction de vraisemblance FV_1 de ces données se présente comme :

$$FV_1 = \left(\frac{e^{-n_1\tau_1}(n_1\tau_1)^{a_1}}{a_1!} \right) \times \left(\frac{e^{-n_0\tau_0}(n_0\tau_0)^{a_0}}{a_0!} \right)$$

Les valeurs des deux paramètres sont fixées par les données elles-mêmes : $\tau_1 = \tau_1$, $\tau_0 = \tau_0$.

Le test du rapport de vraisemblance se construit ici par une comparaison des logarithmes des fonctions de vraisemblance : $-2\log\left(\frac{FV_0}{FV_1}\right)$.

Cette statistique obéit, en bonne approximation, à une loi du χ^2 avec un degré de liberté.

Traduit en formule, ce test se décrit comme :

$$\chi^2_{RV} = 2\left[a_1 \log\left(\frac{t_1}{t}\right) + a_0 \log\left(\frac{t_0}{t}\right) + [tn_1 - t_1n_1] + [tn_0 - t_0n_0] \right]$$

$$= 2\sum_{j=1}^{J}\left[O \log\left(\frac{O}{A}\right) \right]$$

puisque $\sum tn_j = \sum t_j n_j$. Pour la cellule j, O_j correspond à la valeur observée et A_j à la valeur attendue sous l'hypothèse nulle. La sommation se fait sur les deux cellules.

Exemple 5.1

Considérons les données du tableau 5.2 décrivant les résultats d'une étude portant sur l'association entre un facteur d'exposition X et la maladie Y. L'étude a pu être menée sur une population ouverte qui a généré 712 et 1535 personnes-années à risque respectivement chez les exposés et les non-exposés au facteur X. Alors, pour chacune des catégories de X, les rapports du nombre de cas de Y au nombre de personnes-années générées correspondent à des taux. On veut tester l'association entre X et Y. À cette fin, nous appliquons les différents tests présentés précédemment.

Tableau 5.2

	$X = 1$	$X = 0$	Total
Y (décès)	37	51	88
Personnes-années	712	1 535	2 247

Test exact

Le test unilatéral à droite, basé sur la convention mi-p, se calcule alors comme :

$$p = 1/2\, C_{88}^{37}\left(\frac{712}{2247}\right)^{37}\left(\frac{1535}{2247}\right)^{51} + \sum_{i=38}^{88} C_{88}^{i}\left(\frac{712}{2247}\right)^{i}\left(\frac{1535}{2247}\right)^{88-i}$$

La valeur-p unilatérale qui en découle est de 0,020744.

La valeur-p bilatérale correspond à 0,034121.

Test en approximation normale

Test de Mantel-Haenszel (ou de Pearson)

$$\begin{aligned} z^2 &= \sum \frac{(O-A)^2}{A} \\ &= \frac{\left(37 - \dfrac{(712\times 88)}{2247}\right)^2}{\dfrac{(712\times 88)}{2247}} + \frac{\left(51 - \dfrac{(1535\times 88)}{2247}\right)^2}{\dfrac{(1535\times 88)}{2247}} \\ &= \frac{(37-27,88)^2}{27,88} + \frac{(51-60,12)^2}{60,12} = 2,98 + 1,38 \\ &= 4,36 \end{aligned}$$

La valeur-p bilatérale correspondante est de 0,0368.

Test sur les taux transformés par RAC

Appliqué aux données du tableau 5.2, le test conduit aux résultats suivants :

$$\chi_1^2 = \frac{4\left[\sqrt{a_1 n_0} - \sqrt{a_0 n_1}\right]^2}{n}$$
$$= \frac{4\left[\sqrt{37 \times 1535} - \sqrt{51 \times 712}\right]^2}{2247} = 4,06$$

La valeur-*p* bilatérale correspondante est de 0,0439.

Test sur log(RT)

Appliqué aux données du tableau 5.2, on obtient :

$$z = \frac{(\log RT)^2}{\left(\frac{1}{a_1} + \frac{1}{a_0}\right)}$$
$$= \frac{(\log 1,5641)^2}{\left(\frac{1}{37} + \frac{1}{51}\right)}$$
$$= 4,2903$$

La valeur-*p* bilatérale correspondante est de 0,0383.

Test du rapport de vraisemblance

$$z^2 = 2\left[O_1 \log\left(\frac{O_1}{A_1}\right) + O_0 \log\left(\frac{O_0}{A_0}\right)\right]$$
$$= 2\left[37 \log\left(\frac{37}{27,88}\right) + 51 \log\left(\frac{51}{60,12}\right)\right]$$
$$= 2\left[10,47 + 8,39\right] = 4,16$$

La valeur-*p* bilatérale correspondante est de 0,0414.

Dans le tableau 5.3, nous présentons les résultats des différents tests appliqués aux données du tableau 5.2.

TABLEAU 5.3

Méthode	Khi-carré	Valeur-*p*		
		Unilatérale	Bilatérale	Programme
Exacte	–	0,0207	0,0341	PR5.1
MH ou Pearson	4,36	0,0184	0,0367	PR5.2
Transformation RAC	4,06	0,0219	0,0439	PR5.3
Basée sur log(*RT*)	4,29	0,0192	0,0383	PR5.4
RV	4,16	0,0207	0,0414	PR5.5

Avec une bonne concordance, tous ces tests concluent au peu de vraisemblance de l'hypothèse nulle.

◆

5.3 INTERVALLES DE CONFIANCE DES MESURES D'ASSOCIATION POUR DEUX TAUX

Dans cette section, nous présentons les intervalles de confiance des différentes mesures d'association déduites de la comparaison entre deux taux. Plus spécifiquement, nous présentons les intervalles de confiance pour le rapport φ et la différence Δ de deux taux. Pour chacune de ces mesures, nous présentons la méthode exacte de calcul, quelques méthodes en approximation normale et la méthode du rapport de vraisemblance.

5.3.1 INTERVALLE DE CONFIANCE POUR LE RAPPORT φ DE DEUX TAUX

MÉTHODE EXACTE

À partir des expressions 5.1 et 5.5, il est facile de déterminer l'intervalle de confiance exact de cette mesure. Il suffit alors de faire référence à la variable binomiale A_1 de paramètres π et m_1.

Les limites de l'intervalle de confiance de φ sont définies comme suit :

La limite inférieure φ_{inf} de l'intervalle est la valeur de φ qui satisfait l'équation

$$\sum_{i=a_1}^{m_1} C_{m_1}^{i} \left(\frac{\varphi n_1}{\varphi n_1 + n_0} \right)^{i} \left(\frac{n_0}{\varphi n_1 + n_0} \right)^{m_1 - i} = \alpha/2$$

De façon analogue, on définit la limite supérieure φ_{sup} comme la valeur de φ qui satisfait l'équation :

$$\sum_{i=0}^{a_1} C_{m_1}^{i} \left(\frac{\varphi n_1}{\varphi n_1 + n_0} \right)^{i} \left(\frac{n_0}{\varphi n_1 + n_0} \right)^{m_1 - i} = \alpha/2$$

Les formules ci-dessus sont présentées suivant la convention intégrale, bien que dans la pratique nous utilisions la convention mi-*p*. Enfin, on comprendra que de telles équations ne peuvent être résolues que par des méthodes itératives.

Valeurs de A1 correspondant aux limites exactes de φ

On désigne par A_{1inf} et A_{1sup} les valeurs de la variable A_1 qui correspondent respectivement à $E(A_1|\varphi_{inf})$ et $E(A_1|\varphi_{sup})$, pour les valeurs de φ calculées précédemment.

La valeur attendue $A_{1\,inf} = E\left(A_1 \mid \varphi_{inf}\right)$ correspond à la valeur de A_1 qui solutionne l'équation quadratique $\dfrac{A_1(n_0 - m_1 + A_1)}{(m_1 - A_1)(n_1 - A_1)} = \varphi_{inf}$. Concrètement, cette valeur est donnée par $A_{1\,inf} = \dfrac{-V - \sqrt{V^2 - 4UW}}{2U}$

où $U = \varphi_{inf} - 1$, et $V = -[\varphi_{inf}(n_1 + m_1) + (n_0 - m_1)]$.

La valeur attendue $A_{1\,sup}$ est obtenue de façon analogue à partir de la limite φ_{sup}.

Les deux valeurs, $A_{1\,inf}$ et $A_{1\,sup}$, ne sont pas intéressantes en soi. Mais, à l'aide de la relation (5.4), elles permettent de déduire facilement l'intervalle de confiance exact de la mesure Δ (différence de taux), comme on le verra dans une section suivante.

Méthodes en approximation normale

En approximation normale, deux méthodes sont disponibles : la méthode simple et la méthode basée sur le résultat d'un test.

Méthode simple

Le calcul de l'intervalle de confiance de φ se fait à l'aide de la transformation logarithmique de cette mesure. Les limites de l'intervalle de confiance du log φ sont d'abord calculées par approximation normale :

$\log(RT) \pm z_{\alpha/2}\sqrt{V(\log RT)}$.

Puis, les limites de confiance de φ sont déduites par transformation inverse : $RT \times \exp\left[\pm z_{\alpha/2}\sqrt{V(\log RT)}\right]$

où $V(\log\ RT)\ =\ \dfrac{1}{a_1}+\dfrac{1}{a_0}$

$$\varphi_{\text{inf}} = RT \times \exp\left(-z_{\alpha/2}\sqrt{(\log RT))}\right)$$
$$\varphi_{\text{sup}} = RT \times \exp\left(+z_{\alpha/2}\sqrt{(\log RT))}\right)$$

Méthode basée sur le résultat d'un test

En utilisant le résultat du test de Mantel-Haenszel, on peut déduire une estimation pour la variance *V*(log *RT*) décrite dans la méthode précédente. Voici comment.

Sous l'hypothèse nulle, la relation $\chi^2_{MH} \approx \dfrac{[\log RT]^2}{V(\log RT)}$ peut conduire à une bonne estimation de *V*(log *RT*) en autant que cette variance ne soit pas trop instable : $V(\log RT) \approx \dfrac{[\log RT]^2}{\chi^2_{MH}}$

Les limites de confiance de log φ correspondent alors à :

$$\log RT \pm z_{\alpha/2}\sqrt{\frac{[\log RT]^2}{\chi^2_{MH}}}$$

ou

$$\log RT\left(1 \pm z_{\alpha/2}\frac{1}{\sqrt{\chi^2_{MH}}}\right)$$

Par transformation inverse, on obtient donc les limites supérieure et inférieure de l'intervalle de confiance de φ à 100(1 – α) % :

$$\varphi_{\text{inf}} = RT^{\left[1 - \frac{z_{\alpha/2}}{\sqrt{\chi^2_{MH}}}\right]}$$

$$\varphi_{\text{sup}} = RT^{\left[1 + \frac{z_{\alpha/2}}{\sqrt{\chi^2_{MH}}}\right]}$$

Méthode du rapport de vraisemblance

On peut montrer que *RT* est l'estimation du maximum de vraisemblance du rapport φ des deux taux réels τ_1 et τ_0. À cette fin, rappelons que la variable A_1 du tableau 5.1, sous la condition de m_1 fixe, obéit à une loi binomiale de paramètre π et m_1, où $\pi = \dfrac{\varphi n_1}{\varphi n_1 + n_0}$.

La fonction de vraisemblance $L(\varphi)$ construite pour la variable binomiale A_1 atteint son maximum à $\varphi = RT$.

Dans le cadre de la régression binomiale, la fonction de vraisemblance $FV(\varphi)$ est de la forme :

$$\begin{aligned} FV(\varphi) &= C_{m_1}^{a_1} \pi_1^{a_1} (1-\pi_1)^{m_1 - a_1} \\ &= C_{m_1}^{a_1} \left(\frac{\varphi n_1}{\varphi n_1 + n_0} \right)^{a_1} \left(\frac{n_0}{\varphi n_1 + n_0} \right)^{m_1 - a_1} \end{aligned}$$

Les limites de l'intervalle de confiance de φ au niveau 100(1 – α) % correspondent alors aux deux valeurs de φ qui satisfont l'équation logarithmique

$$2\left[L - L(\varphi)\right] - \chi^2_{1,1-\alpha} = 0 \tag{5.9}$$

En utilisant les notations du tableau 5.1, on peut établir que

$$L(\varphi) = a_1 \log\left(\frac{\varphi n_1}{\varphi n_1 + n_0} \right) + a_0 \log\left(\frac{n_0}{\varphi n_1 + n_0} \right) + K$$

et que

$$L = a_1 \log\left(\frac{a_1}{m_1}\right) + a_0 \log\left(\frac{a_0}{m_1}\right) + K$$

où $K = \log C_{m_1}^{a_1}$, valeur indépendante de φ.

L'équation à solutionner par itération est alors de la forme :

$$2\left\{a_1 \log\left[\frac{a_1(\varphi n_1 + n_0)}{\varphi n_1}\right] + a_0 \log\left[\frac{a_0(\varphi n_1 + n_0)}{n_0}\right] - m_1 \log(m_1)\right\} - \chi^2_{1,1-\alpha} = 0$$

L'approche peut aussi être définie à partir de la régression de Poisson. Dans ce cadre, on peut montrer que le log(*RT*) est l'estimateur du maximum de vraisemblance de β du modèle linéaire $\log \tau(X) = \alpha + \beta X$, où τ(*X*) désigne le taux en fonction de *X*. La fonction de vraisemblance *L*(α,β) construite pour le modèle log τ(*X*) atteint son maximum lorsque α = log(t_0) et β = log(*RT*). On note que φ = e^{β}. Dans ce cadre, on peut facilement établir la fonction de vraisemblance comme suit :

$$\begin{aligned} FV(\alpha,\beta) &= \left[\frac{e^{-n_1\tau_1}(n_1\tau_1)^{a_1}}{a_1!}\right] \times \left[\frac{e^{-n_0\tau_0}(n_0\tau_0)^{a_0}}{a_0!}\right] \\ &= \left[\frac{e^{-n_1(e^{\alpha+\beta})}\left(n_1 e^{\alpha+\beta}\right)^{a_1}}{a_1!}\right] \times \left[\frac{e^{-n_0 e^{\alpha}}\left(n_0 e^{\alpha}\right)^{a_0}}{a_0!}\right] \end{aligned}$$

L'équation logarithmique (5.9) à résoudre est alors décrite à l'aide des valeurs :

$$L(\varphi) = -n_1\varphi e^{\alpha} - n_0 e^{\alpha} + (a_1 + a_0)\alpha + a_1 \log(\varphi) + K$$

et

$$L = -(a_1 + a_0) + a_1 \log(a_1 / n_1) + a_0 \log(a_0 / n_0) + K$$

où $K = a_1 \log a_1 + a_0 \log a_0 - \log a_1! - \log a_0!$, valeur indépendante de φ.

Pour une valeur déterminée de φ, la fonction *L*(φ) atteint son maximum si $e^{\alpha} = \dfrac{a_1 + a_0}{\varphi n_1 + n_0}$.

Ces éléments permettent d'induire assez facilement le processus itératif d'estimation, qui conduit à des résultats identiques à ceux de la méthode précédente.

EXEMPLE 5.2

Considérons les données du tableau 5.2. On veut calculer l'intervalle de confiance à 95 % de φ, dont la valeur estimée est de 1,56 : $RT = \frac{37 \times 1535}{51 \times 712} = 1,5641$

MÉTHODE EXACTE

Les limites de confiance inférieure et supérieure à 95 % du rapport de taux φ sont données par les valeurs de φ qui satisfont respectivement les relations suivantes :

$$\sum_{i=37}^{88} C_{88}^{i} \left(\frac{712\varphi}{712\varphi + 1535} \right)^{i} \left(\frac{1535}{712\varphi + 1535} \right)^{88-i} = 0,025$$

et

$$\sum_{i=0}^{37} C_{88}^{i} \left(\frac{712\varphi}{712\varphi + 1535} \right)^{i} \left(\frac{1535}{712\varphi + 1535} \right)^{88-i} = 0,025$$

Ces équations, qui ne peuvent être résolues que par itération, conduisent aux limites approximatives inférieure et supérieure suivantes :

$\varphi_{inf} = 1,0178$

$\varphi_{sup} = 2,3864$

Les valeurs $A_{1\,inf}$ et $A_{1\,sup}$ correspondantes sont calculées de la façon suivante.

– Pour $A_{1\,inf}$, on considère $\varphi_{inf} = 1,0178$ et les valeurs associées U, V et W.

U = 1,0178 –1 = 0,0178

V = –[1,0178(712 + 88) + (1535 – 88)] = –2261,24

W = 1,0178 × 712 × 88 = 63771,28

La valeur $A_{1\,inf}$ recherchée est donc

$$\begin{aligned} A_{1\,inf} &= \frac{-V - \sqrt{V^2 - 4UW}}{2U} \\ &= \frac{2261,24 - \sqrt{2261,24^2 - 4 \times 0,0178 \times 63771,28}}{2 \times 0,0178} \\ &= 28,2082 \end{aligned}$$

– Pour $A_{1\,sup}$, on considère $\varphi_{sup} = 2,3864$. Le calcul conduit alors à $A_{1\,sup} = 45,4040$.

Approximation normale

Méthode simple

$$\begin{aligned}\varphi_{inf} &= 1{,}5641 \times \exp\left[-1{,}96\sqrt{\frac{1}{37}+\frac{1}{51}}\right] \\ &= 1{,}0243 \\ \varphi_{sup} &= 1{,}5641 \times \exp\left[+1{,}96\sqrt{\frac{1}{37}+\frac{1}{51}}\right] \\ &= 2{,}3883\end{aligned}$$

Basée sur le résultat d'un test

La valeur du khi-carré de Mantel-Haenszel est de 4,36 (voir tableau 5.3). On déduit alors les limites de confiance suivantes :

$$\begin{aligned}\varphi_{inf} &= 1{,}5641^{\left[1-\frac{1{,}96}{\sqrt{4{,}36}}\right]} \\ &= 1{,}0278 \\ \varphi_{sup} &= 1{,}5641^{\left[1+\frac{1{,}96}{\sqrt{4{,}36}}\right]} \\ &= 2{,}3802\end{aligned}$$

Méthode du rapport de vraisemblance

La fonction de vraisemblance à solutionner par itération est :

$$2\left\{\begin{array}{l}37\times\log\left[\dfrac{37(\varphi\times 712+1535)}{\varphi\times 712}\right]+ \\ 51\times\log\left[\dfrac{51(\varphi\times 712+1535)}{1535}\right]-88\times\log(88)\end{array}\right\}-3{,}84=0$$

Les limites de confiance obtenues par itération sont :

$$\varphi_{inf} = 1{,}0178$$
$$\varphi_{sup} = 2{,}3807$$

Dans le tableau 5.4, on rappelle les intervalles de confiance obtenus par les différentes méthodes de calcul.

Tableau 5.4

Méthode	Intervalle de confiance à 95 %	Programme
Exacte	[1,0178 ; 2,3864]	PR5.6
Simple	[1,0243 ; 2,3882]	PR5.7
Résultat d'un test	[1,0279 ; 2,3799]	PR5.8
RV	[1,0178 ; 2,3807]	PR5.9

On remarque que la méthode exacte et celle du rapport de vraisemblance sont très concordantes.

◆

5.3.2 INTERVALLE DE CONFIANCE DE LA DIFFÉRENCE Δ ENTRE DEUX TAUX

MÉTHODE EXACTE

À partir des relations (5.2) et (5.6), il est facile de déterminer l'intervalle de confiance exact de cette mesure. Il suffit alors de revenir à la variable binomiale A_1 de paramètres π et m_1.

Les limites de l'intervalle de confiance de Δ sont définies comme suit :

- la limite inférieure Δ_{inf} est la valeur de Δ qui satisfait l'équation suivante :

$$\sum_{i=a_1}^{m_1} C_{m_1}^{i} \left(\frac{\Delta n_1 n_0 + m_1 n_1}{m_1 n} \right)^{i} \left(\frac{m_1 n_0 - \Delta n_1 n_0}{m_1 n} \right)^{m_1 - i} = \alpha/2$$

- la limite supérieure Δ_{sup} est la valeur de Δ qui satisfait l'équation :

$$\sum_{i=0}^{a_1} C_{m_1}^{i} \left(\frac{\Delta n_1 n_0 + m_1 n_1}{m_1 n} \right)^{i} \left(\frac{m_1 n_0 - \Delta n_1 n_0}{m_1 n} \right)^{m_1 - i} = \alpha/2$$

Les formules ci-dessus sont aussi présentées suivant la convention intégrale, bien que dans la pratique nous utilisions la convention mi-*p*.

On peut aussi déduire assez facilement les limites de confiance exactes de la différence Δ à partir de certaines relations décrites à la section 5.1 :

- par la relation (5.4) et utilisant les valeurs $A_{1\,inf}$ et $A_{1\,sup}$:

$$\Delta_{inf} = \frac{A_{1\,inf} n - m_1 n_1}{n_1 n_0}$$

$$\Delta_{sup} = \frac{A_{1\,sup} n - m_1 n_1}{n_1 n_0}$$

- par la relation (5.8), en utilisant les valeurs φ_{inf} et φ_{sup} :

$$\Delta_{inf} = \frac{(\varphi_{inf} - 1) m_1}{\varphi_{inf} n_1 + n_0}$$

$$\Delta_{sup} = \frac{(\varphi_{sup} - 1) m_1}{\varphi_{sup} n_1 + n_0}$$

Méthode en approximation normale

Méthode simple

La méthode est celle décrite par l'expression : $DT \pm z_{\alpha/2}\sqrt{V(DT)}$. Suivant les notations du tableau 5.1, la variance $V(DT)$ correspond à :

$$V(DT) = \left(\frac{t_1}{n_1} + \frac{t_0}{n_0}\right)$$
$$= \frac{a_1}{n_1^2} + \frac{a_0}{n_0^2}$$

Dans cette méthode, on suppose que la variance est stable. En général, cette supposition est relativement correcte. Mais parfois, elle peut s'avérer insatisfaite.

Méthode basée sur le résultat du test des taux transformés

La méthode est basée sur le résultat du test des taux transformés : χ_T^2 (voir la section 5.2.2). On suppose que le khi-carré sur la différence DT, dont la variance aurait été stabilisée, est approximativement égal à χ_T^2. Cette supposition permet d'obtenir une estimation $V_S(DT)$ de la variance stabilisée pour DT. Ainsi, sous H_0,

$$\frac{[DT]^2}{V_s(DT)} \approx \chi_T^2 \quad \Rightarrow \quad \mathrm{V_s}(DT) \approx \frac{[DT]^2}{\chi_T^2}$$

L'intervalle de confiance de DT pour laquelle la variance a été stabilisée se présente donc comme :

$$IC \quad : \quad DT \pm z_{\alpha/2}\sqrt{V_S(DT)}$$
$$: \quad \mathrm{DT}\left[1 \pm \mathrm{z}_{\alpha/2}/\chi_{\mathrm{T}}\right]$$

Méthode du rapport de vraisemblance

Établissons d'abord que DT est l'estimateur du maximum de vraisemblance de Δ. À cette fin, nous considérons le simple modèle $\tau(X) = \alpha + \beta X$, qui décrit le lien entre X et le taux. Le coefficient β de ce modèle représente identiquement le paramètre Δ. Il suffit alors de construire la fonction de vraisemblance à partir des données du tableau 5.1. Cette fonction se présente comme :

$$FV(\alpha,\beta) = \left(\frac{e^{-n_1\tau_1}(n_1\tau_1)^{a_1}}{a_1!} \right) \times \left(\frac{e^{-n_0\tau_0}(n_0\tau_0)^{a_0}}{a_0!} \right)$$
$$= \left(\frac{e^{-n_1(\alpha+\beta)}[n_1(\alpha+\beta)]^{a_1}}{a_1!} \right) \times \left(\frac{e^{-n_0\alpha}[n_0\alpha]^{a_0}}{a_0!} \right)$$

où $\tau_1 = \tau(1)$ et $\tau_0 = \tau(0)$.

Pour obtenir les estimateurs de α et β, on considère la fonction $L(\alpha,\beta)$: $L(\alpha,\beta) = -n_1(\alpha+\beta) + a_1 \log(\alpha+\beta) - n_0\alpha + a_0 \log\alpha + K$

où $K = a_1 \log a_1 + a_0 \log a_0 - \log a_1! - \log a_0!$, valeur indépendante de α et β.

Les estimateurs de α et β sont les valeurs qui annulent les dérivées de $L(\alpha,\beta)$:

$$\left\{ \begin{array}{l} \dfrac{\partial L(\alpha,\beta)}{\partial\alpha} = 0 \\ \dfrac{\partial L(\alpha,\beta)}{\partial\beta} = 0 \end{array} \right. \Rightarrow \left. \begin{array}{l} \alpha = t_0 \\ \beta = DT \end{array} \right\}$$

On remarque alors que $\alpha + \beta = t_1$.

Les limites de l'intervalle de confiance de Δ au niveau $100(1 - \alpha)$ % correspondent alors aux deux valeurs de Δ qui satisfont l'équation logarithmique $2\left[L - L(\beta)\right] - \chi^2_{1,1-\alpha} = 0$ où L correspond à la valeur maximale de la fonction du log $FV(\alpha,\beta)$ et $L(\beta)$ à la valeur maximale de la fonction du log $FV(\alpha,\beta)$ pour une valeur fixe de β.

Pour cette équation logarithmique, les valeurs L et $L(\beta)$ sont respectivement données par :

$$L = -n_1t_1 + a_1 \log t_1 - n_0t_0 + a_0 \log t_0 + K$$
$$L(\beta) = L(\alpha,\beta) = -n_1(\alpha+\beta) + a_1 \log(\alpha+\beta) - n_0\alpha + a_0 \log\alpha + K$$

où α est l'estimation du maximum de vraisemblance pour une valeur déterminée de β. La valeur α correspond, en définitive, à la racine positive de l'équation quadratique suivante :

$$n\alpha^2 + (n\beta - a_1 - a_0)\alpha - a_0\beta = 0$$

La valeur K est indépendante de α et de β.

Exemple 5.3

Pour les données du tableau 5.2, $DT = 0{,}0187$ ou 18,7 décès par 1000 personnes-années. On veut calculer l'intervalle de confiance à 95 % de cette mesure. À cette fin, nous allons appliquer les différentes méthodes de calcul présentées précédemment.

Méthode exacte

Les limites de confiance inférieure et supérieure de Δ correspondent alors aux valeurs de Δ qui satisfont respectivement les relations suivantes :

$$\sum_{i=37}^{88} C_{88}^{i} \left(\frac{\Delta \times 712 \times 1535 + 88 \times 712}{88 \times 2247} \right)^{i} \left(\frac{88 \times 1535 - \Delta \times 712 \times 1535}{88 \times 2247} \right)^{88-i} = 0{,}025$$

$$\sum_{i=0}^{37} C_{88}^{i} \left(\frac{\Delta \times 712 \times 1535 + 88 \times 712}{88 \times 2247} \right)^{i} \left(\frac{88 \times 1535 - \Delta \times 712 \times 1535}{88 \times 2247} \right)^{88-i} = 0{,}025$$

Pour les limites de confiance à 95 % de la différence Δ, les valeurs obtenues sont : $\Delta_{\text{inf}} = 0{,}000693$ et $\Delta_{\text{sup}} = 0{,}037723$

En utilisant la relation (5.4) et les valeurs de $A_{1\ \text{inf}}$ et $A_{1\ \text{sup}}$ calculées à l'exemple 5.2, on obtient :

$$\begin{aligned} \Delta_{\text{inf}} &= \frac{28{,}2045 \times 2247 - 88 \times 712}{712 \times 1535} \\ &= 0{,}000658 \\ \Delta_{\text{inf}} &= \frac{45{,}404 \times 2247 - 88 \times 712}{712 \times 1535} \\ &= 0{,}0360 \end{aligned}$$

En utilisant la relation (5.8) et les valeurs φ_{inf} (= 1,0178) et φ_{sup} (= 2,3864), calculées pour φ à l'exemple 5.2, on obtient :

$$\begin{aligned} \Delta_{\text{inf}} &= \frac{(1{,}0178 - 1) \times 88}{1{,}0178 \times 712 + 1535} \\ &= 0{,}000693 \\ \Delta_{\text{inf}} &= \frac{(2{,}3864 - 1) \times 88}{2{,}3864 \times 712 + 1535} \\ &= 0{,}0377 \end{aligned}$$

Les différences entre ces valeurs et celles calculées par la procédure exacte sont essentiellement dues aux arrondissements.

Approximation normale

Méthode simple

La variance de DT est estimée par :

$$V(DT) = \frac{37}{712^2} + \frac{51}{1535^2}$$
$$= 0,00009463$$

Les limites de l'intervalle de confiance à 95 % pour Δ sont alors données par :

$$\Delta_{inf} = 0,0187 - 1,96\sqrt{0,00009463}$$
$$= -0,0003247$$
$$\Delta_{sup} = 0,0187 + 1,96\sqrt{0,00009463}$$
$$= 0,037808$$

Méthode basée sur le résultat du test des taux transformés

Appliqué aux données du tableau 5.2, en rappelant la valeur du test χ^2_T calculé à l'exemple 5.1, on obtient pour l'intervalle de confiance à 95 % de la différence Δ (= 0,0187) les limites suivantes :

$$\Delta_{inf} = 0,0187\left[1 - \frac{1,96}{\sqrt{4,06}}\right] = 0,000513$$
$$\Delta_{sup} = 0,0187\left[1 + \frac{1,96}{\sqrt{4,06}}\right] = 0,036970$$

Méthode du rapport de vraisemblance

Cette méthode peut utiliser directement la procédure GENMOD de SAS.

Appliquée aux données du tableau 5.2, l'intervalle de confiance à 95 % de la différence Δ estimé par la méthode du maximum de vraisemblance correspond à : [0,000694 ; 0,039152].

Nous résumons dans le tableau 5.5 les résultats obtenus par les différentes méthodes de calcul.

Tableau 5.5

Méthode	*DT*	Intervalle de confiance à 95 %	Programme
Exacte	0,0187	[0,00069 ; 0,03772]	Pr5.10
Simple	0,0187	[-0,00032 ; 0,03781]	Pr5.11
Basée sur χ^2_T	0,0187	[0,00051 ; 0,03697]	Pr5.12
RV	0,0187	[0,00069 ; 0,03915]	Pr5.13

On remarque ici que l'intervalle de confiance par la méthode simple, [–0,00033 ; 0,03781], recouvre la valeur 0. Ses limites ne sont cohérentes ni avec les résultats des autres méthodes, ni même cohérentes avec les résultats des tests statistiques décrits précédemment, qui sont significatifs au niveau 5 % (tableau 5.3). L'instabilité des variances des taux, et donc de la différence de ces taux, peut expliquer en grande partie cette incohérence. L'approche de la variance stabilisée basée sur le résultat χ^2_T nous apparaît une méthode en approximation normale, plus cohérente et plus fiable.

◆

5.4 MESURE DU *SMR* POUR LES TAUX

Considérons le cas particulier où l'on veut comparer le taux t_1 d'un groupe G sous observation au taux τ_0 d'une population P de référence. On suppose que x décès ont été observés pour n_1 personnes-temps dans le groupe concerné. La comparaison du taux observé au taux de référence peut alors se concrétiser dans la comparaison entre le nombre x de cas observés dans le groupe G au nombre $X_0 = n_1\tau_0$ de cas attendus, sous l'hypothèse que le groupe est un échantillon aléatoire de la population de référence P. On aura reconnu le *SMR* dans le rapport $\frac{x}{X_0}$ ou $\frac{t_1}{\tau_0}$.

Si on désigne par τ_1 le vrai taux du groupe sous observation, alors la valeur attendue de X est donnée par $E(X) = n_1\tau_1$. Par ailleurs, la valeur vraie du *SMR*, désignée par ϕ, est donnée par $\phi = \frac{\tau_1}{\tau_0}$. Ainsi, dans le cas où le groupe G provient de la population de référence P, on a $\phi = 1$. Sinon $\phi \neq 1$. Concrètement, on ignore la valeur τ_1 et donc celle du vrai *SMR*.

Sous l'hypothèse nulle ($\phi = 1$ ou $\tau_1 = \tau_0$), la valeur attendue $E(X)$ est donnée par $E(X) = n_1\tau_1 = n_1\tau_0 = X_0$.

Sous une hypothèse quelconque sur ϕ, ou $\tau_1 = \phi\tau_0$, la valeur attendue $E(X)$ est donnée par $E(X) = n_1\tau_1 = n_1\phi\tau_0 = \phi[n_1\tau_0] = \phi X_0$.

Avec ces préalables, nous proposons ici quelques tests statistiques et intervalles de confiance parmi les plus usités pour porter un jugement sur le *SMR*.

5.4.1 TESTS STATISTIQUES POUR LE *SMR*

Nous suggérons trois tests pour le *SMR* : le test exact, deux tests en approximation normale et le test du rapport de vraisemblance. Certains de ces tests sont similaires à ceux décrits au chapitre 3, section 3.1.

Test exact

Le test exact est directement construit sur la variable X de Poisson. Sous l'hypothèse nulle, pour $X = x$, le test unilatéral à droite, dans la convention mi-p, est de la forme :

$$p = \frac{e^{-X_0}(X_0)^x}{2x!} + \sum_{i=x+1}^{\infty} \frac{e^{-X_0}(X_0)^i}{i!}$$
$$= 1 - \sum_{i=0}^{x-1} \frac{e^{-X_0}(X_0)^i}{i!} - \frac{e^{-X_0}(X_0)^x}{2x!}$$

Test en approximation normale

Test simple

À l'aide de l'approximation normale appliquée à la variable de Poisson X, il est facile de construire un test sous l'hypothèse nulle $\phi = 1$.

$$z = \frac{x - X_0}{\sqrt{X_0}}$$

ou encore

$$\chi_1^2 = \frac{(x - X_0)^2}{X_0}$$

Test sur log (SMR)

On peut facilement construire un test sur le logarithme du *SMR*.

Le *SMR* observé correspond à $\frac{x}{X_0}$. Sous l'hypothèse nulle $\phi = 1$, la statistique log (*SMR*) a comme valeur attendue 0. L'estimation de sa variance par le maximum de vraisemblance correspond simplement à $\frac{1}{x}$. Le test se présente alors simplement comme : $\chi_1^2 = x \left[\log(SMR)\right]^2$.

Test basé sur la transformation RAC

Rappelons que le SMR se définit comme : $SMR = \frac{X}{X_0}$ et que sous l'hypothèse nulle, la valeur attendue de X est X_0.

Le test est alors construit sur la comparaison de la valeur observée $\sqrt{X}$ à sa valeur attendue $\sqrt{X_0}$. La variance de $\sqrt{X}$ est égale simplement à $\frac{1}{4}$.

On construit alors la statistique $Z^2 = \frac{\left(\sqrt{X} - \sqrt{X_0}\right)^2}{\frac{1}{4}}$ qui obéit en bonne approximation à une loi du khi-carré avec un degré de liberté.

Le test prend alors la forme $\chi_1^2 = 4\left(\sqrt{x} - \sqrt{X_0}\right)^2$

Test du rapport de vraisemblance

Sous l'hypothèse nulle d'un *SMR* égal à 1 ($\phi = 1$ ou $\tau_1 = \tau_0$), le paramètre de la loi se réduit simplement à X_0 et la fonction de vraisemblance FV_0 peut s'écrire : $FV_0 = \frac{e^{-X_0}\left(X_0\right)^x}{x!}$.

Par ailleurs, la contre-hypothèse correspond à celle, parmi toutes, pour laquelle la fonction *FV* atteint son maximum. Puisque, dans ce cas, ϕ correspond au *SMR* mesuré sur les données, on a $FV_1 = \frac{e^{-x}x^x}{x!}$

Le test du rapport de vraisemblance se présente alors dans une forme déjà rencontrée : $-2\log\left(\frac{FV_0}{FV_1}\right) = 2\left[O\log\frac{O}{A} - (O - A)\right]$, où O représente la valeur observée ($O = x$) et A la valeur X_0 attendue sous l'hypothèse nulle.

Exemple 5.4

Dans un groupe de travailleurs qui ont cumulé 5000 personnes-années à risque, on a recensé 20 décès. Pour la même période, le taux de décès de la population de référence est de 0,002 par année. Le nombre de 20 décès recensés chez ce groupe de travailleurs correspond-il à un excès ou est-ce une donnée compatible avec l'hypothèse d'un *SMR* égal à 1 ?

Pour répondre à cette question, établissons d'abord que, sur ces données, x = 20 décès, $\tau_0 = 0{,}002$ an^{-1}, $n_1 = 5000$ personnes-années. Ainsi, $X_0 = 5000 \times 0{,}002 = 10$ décès, et $SMR = 20/10 = 2$.

Test exact

Puisque $X_0 = 10$, alors dans la convention mi-p le test unilatéral à droite se présente comme

$$p = 1 - \sum_{i=0}^{19} \frac{e^{-10}(10)^i}{i!} - \frac{e^{-10}(10)^{20}}{2 \times 20!}$$
$$= 0{,}0025213$$

L'hypothèse d'un excès se trouve donc tout à fait vraisemblable puisque celle d'un $SMR = 1$ (l'hypothèse nulle) l'est fort peu : valeur-p unilatérale = 0,00252. La valeur-p bilatérale est de 0,00302.

Approximation normale

Test simple

Le test donne :

$$\chi_1^2 = \frac{(20-10)^2}{10}$$
$$= 10$$

La valeur-p bilatérale correspondante est de 0,00157 et la valeur unilatérale de 0,00078.

Test basé sur log(*SMR*)

Le test donne :

$$\chi_1^2 = 20[\log 2]^2$$
$$= 9{,}61$$

La valeur-p bilatérale correspondante est de 0,00194 et la valeur unilatérale de 0,00097.

Test basé sur la transformation RAC

Le test donne :

$$\chi_1^2 = 4\left(\sqrt{20} - \sqrt{10}\right)^2$$
$$= 6{,}8629$$

La valeur-*p* bilatérale correspondante est de 0,008800 et la valeur unilatérale de 0,004400.

Test du rapport de vraisemblance

La valeur du test est calculée comme :

$$\chi_1^2 = 2\left[20\log\frac{20}{10} - (20-10)\right]$$
$$= 7,7259$$

La valeur-*p* bilatérale correspondante est de 0,00544 et unilatérale à droite de 0,00272.

Dans le tableau 5.6, nous présentons les résultats des différents tests appliqués aux données de l'exemple.

Tableau 5.6

Méthode	Khi-carré	Valeur-*p*		
		Unilatérale	Bilatérale	Programme
Exact	–	0,00252	0,00302	PR5.14
Normale simple	10,00	0,00078	0,00157	PR5.15
Basé sur log(*SMR*)	9,61	0,00097	0,00194	PR5.16
Transformation RAC	6,86	0,00440	0,00880	PR5.17
Rapport de vraisemblance	7,73	0,00272	0,00544	PR5.18

On remarque que les tests exacts et du rapport de vraisemblance sont assez concordants. Il en va de même pour les deux tests en approximation normale, qui apparaissent cependant un peu plus sensibles que les premiers. Mais, pour l'essentiel, ces tests sont tous concordants.

◆

5.4.2 INTERVALLE DE CONFIANCE DU *SMR*

L'intervalle de confiance d'un *SMR* se calcule de façon analogue à celui d'un taux (section 3.2, du chapitre 3).

Soit x décès observés pour les n_1 personnes-temps cumulées dans un groupe. On suppose que le taux de la population générale est de τ_0. L'estimation du *SMR* (ou de ϕ) sur les données est x/X_0. On veut déterminer l'intervalle de confiance de ϕ au niveau $100(1-\alpha)\%$.

Intervalle de confiance exact du *SMR*

Dans l'approche exacte, les limites de confiance inférieure ϕ_{inf} et supérieure ϕ_{sup} de ϕ se définissent de la façon suivante :

ϕ_{inf} correspond à la valeur de ϕ qui satisfait l'équation suivante :

$$\sum_{i=x}^{\infty} \frac{e^{-\phi X_0}(\phi X_0)^i}{i!} = \alpha/2$$

ϕ_{sup} correspond à la valeur de ϕ qui satisfait l'équation suivante :

$$\sum_{i=0}^{x} \frac{e^{-\phi X_0}(\phi X_0)^i}{i!} = \alpha/2 .$$

Par itération sur ϕ, et en adoptant la convention mi-*p*, on obtient assez facilement les limites supérieure et inférieure de l'intervalle.

Intervalle de confiance du *SMR* en approximation normale

La procédure basée sur l'approximation normale appliquée directement au *SMR* peut conduire à des incohérences. L'instabilité de la variance de cette mesure en est la cause : la variance du *SMR* est dépendante du *SMR* : $V(SMR) = \dfrac{\phi}{X_0}$.

Pour contrer ce problème, on propose soit la méthode basée sur le logarithme du *SMR*, soit celle basée sur la racine carrée du *SMR*.

Méthode basée sur le logarithme du SMR

On calcule d'abord l'intervalle de confiance sur la transformation logarithmique du *SMR* : $\log(SMR) \pm z_{\alpha/2}\sqrt{V\left[\log(SMR)\right]}$

Puisque $V\left[\log(SMR)\right] = \dfrac{1}{x}$, on a simplement :

$$\log(SMR) \pm z_{\alpha/2}\sqrt{\frac{1}{x}}$$

Par transformation inverse, on obtient les limites de confiance du *SMR.* :

$$\phi_{inf} = SMR \times \exp\left(-\frac{z_{\alpha/2}}{\sqrt{x}}\right)$$

$$\phi_{sup} = SMR \times \exp\left(+\frac{z_{\alpha/2}}{\sqrt{x}}\right)$$

Transformation RAC

On propose de travailler directement sur la racine carrée du *SMR*. On a :

$V\left(\sqrt{SMR}\right) = \frac{1}{4X_0}$, en se rappelant que le numérateur X du *SMR* est une variable de Poisson. L'intervalle de confiance est alors calculé sur la variable $\sqrt{SMR}$. Et par transformation inverse, on retrouve les limites de confiance du *SMR*.

Suivant cette méthode, les limites de confiance du *SMR* se décrient formellement comme :

$$\phi_{inf} = \left(\sqrt{SMR} - \frac{z_{\alpha/2}}{2\sqrt{X_0}}\right)^2$$

$$\phi_{sup} = \left(\sqrt{SMR} + \frac{z_{\alpha/2}}{2\sqrt{X_0}}\right)^2$$

Intervalle de confiance par la méthode du rapport de vraisemblance

On se rappelle que la fonction de vraisemblance pour le *SMR* (= ϕ) a la forme suivante :

$$FV(\phi) = \frac{e^{-\phi X_0}\left(\phi X_0\right)^x}{x!}$$

Ainsi, puisque $L(\phi) = x\log\left(\phi X_0\right) - \phi X_0 - \log x!$ et que $L = x\log(x) - x - \log x!$, l'équation $2\left[L - L(\phi)\right] - \chi^2_{1,1-\alpha} = 0$ devient $2\left\{(x\log x - x) - \left[x\log\left(\phi X_0\right) - \phi X_0\right]\right\} - \chi^2_{1,1-\alpha} = 0.$

Les deux valeurs de ϕ qui satisfont cette dernière équation correspondent aux limites de l'intervalle de confiance à $100(1 - \alpha)\,\%$ pour le *SMR*. La résolution de l'équation peut se faire assez facilement par itération. La procédure GENMOD de SAS permet aussi d'obtenir ces limites de confiance.

Exemple 5.5

Nous appliquons les différentes méthodes aux données de l'exemple 5.4, où le *SMR* = 2.

Méthode exacte

On obtient les limites de confiance suivantes :

$$\phi_{inf} = 1{,}25597$$

$$\phi_{sup} = 3{,}03396$$

Approximation normale

Méthode basée sur le logarithme du *SMR*

L'intervalle de confiance est obtenu par l'approximation normale appliquée à log (*SMR*).

On a :

$$\phi_{inf} = 2 \times \exp\left(-\frac{1{,}96}{\sqrt{20}}\right) = 1{,}2903$$

$$\phi_{sup} = 2 \times \exp\left(+\frac{1{,}96}{\sqrt{20}}\right) = 3{,}1000$$

SMR transformé par RAC

L'intervalle de confiance est obtenu par de l'approximation normale appliquée au *SMR* transformé par RAC.

On a :

$$\phi_{inf} = \left(\sqrt{2} - \frac{1{,}96}{2 \times \sqrt{10}}\right)^2 = 1{,}2195$$

$$\phi_{sup} = \left(\sqrt{2} + \frac{1{,}96}{2 \times \sqrt{10}}\right)^2 = 2{,}9726$$

Méthode du rapport de vraisemblance

Il faut solutionner l'équation logarithmique suivante :
$2\left\{(20\ \log 20 - 20) - \left[20\log(10\phi) - 10\phi\right]\right\} - 3{,}84 = 0$ ou plus simplement $\phi - 2\log\phi = 0{,}8057$.

Les deux solutions de cette équation correspondent aux limites de confiance recherchées :

$$\phi_{inf} = 1,24657$$
$$\phi_{sup} = 3,00898$$

Dans le tableau 5.7, nous présentons l'intervalle de confiance du *SMR* calculé par les différentes méthodes.

TABLEAU 5.7

Méthode de calcul	Intervalle de confiance à 95 % de ϕ	Programme
Exacte	[1,25597 ; 3,03396]	PR5.19
Transformation RAC	[1,21950 ; 2,97258]	PR5.20
Méthode du log	[1,29031 ; 3,10002]	PR5.21
Rapport de vraisemblance	[1,24657 ; 3,00898]	PR5.22

On observe une bonne concordance entre les quatre méthodes, surtout entre la méthode exacte et la méthode du rapport de vraisemblance.

◆

CHAPITRE 6

LES PROPORTIONS DANS UN TABLEAU 2 × 2

Nous considérons les résultats d'une étude portant sur l'association entre un facteur d'exposition *X* et une maladie *Y*. Les données de l'étude reposent sur l'observation de personnes. Elles peuvent être disposées dans un tableau de contingence 2 × 2 (tableau 6.1) où la distribution de l'exposition *X* est croisée avec celle de la maladie *Y*.

TABLEAU 6.1

	$X = 1$	$X = 0$	Total
$Y = 1$	a_1	a_0	m_1
$Y = 0$	b_1	b_0	m_0
Total	n_1	n_0	n

Si l'échantillonnage est pratiqué sur l'exposition ou sur la population totale, comme c'est le cas des études de cohortes et de certaines études de prévalence, alors p_1 et p_0 représentent respectivement les proportions (risque ou prévalence) chez les

exposés ($X = 1$) et chez les non-exposés ($X = 0$) : $p_1 = \frac{a_1}{n_1}$ et $p_0 = \frac{a_0}{n_0}$. La différence (*DP*), le rapport (*RP*), et le rapport de cotes (*RC*) de ces mesures de base sont les différentes mesures d'association les plus souvent considérées entre *X* et *Y*. Ces mesures d'association sont simplement définies comme :

$$DP = p_1 - p_0, \; RP = \frac{p_1}{p_0} \text{ et } \quad RC = \frac{p_1(1-p_0)}{p_0(1-p_1)} = \frac{a_1 b_0}{a_0 b_1}.$$

Si l'étude est de type cas-témoins, les proportions p_1 et p_0 sont d'une certaine façon artificielles et ininterprétables au plan épidémiologique. La seule mesure d'intérêt est le rapport de cotes *RC*.

Dans les sections suivantes, nous décrivons les tests statistiques et les intervalles de confiance pour chacune de ces mesures d'association. Tant pour les tests statistiques que pour les intervalles de confiance de ces mesures, nous présentons d'abord les approches exactes, puis les approches approximatives. Ces dernières sont construites à partir de l'approximation normale et par le rapport de vraisemblance. Les tests statistiques seront essentiellement conduits sous l'hypothèse H_0. Dans une dernière section, nous ajoutons une présentation du *SMR* défini pour les proportions, mesure d'association basée sur la comparaison d'une proportion observée à une proportion théorique ou attendue.

6.1 QUELQUES RELATIONS DE BASE ENTRE LES MESURES

Considérons les données du tableau 6.1, où a_1 et a_0 décrivent les nombres de cas observés respectivement pour les n_1 personnes exposées au facteur *X* et les n_0 personnes non exposées à ce facteur.

Nous rappelons que les variables A_1 et A_0, correspondant aux deux cellules du tableau, sont deux variables binomiales indépendantes respectivement de paramètres (π_1, n_1) et (π_0, n_0) où π_1 et π_0 sont les proportions réelles, généralement inconnues. Les proportions estimées correspondantes sont $p_1 = a_1/n_1$ pour π_1 et $p_0 = a_0/n_0$ pour π_0. On a aussi $p = m_1/n$.

Sous la condition de m_1 fixe, A_1 est une variable hypergéométrique : $A_1 \mapsto \text{Hypr}(n, n_1, m_1)$.

Nous nous intéressons alors à l'inférence statistique (tests ou intervalles de confiance) des mesures réelles d'association entre *X* et *Y*. Ces mesures d'association, désignées par Δ, ξ et ψ, correspondent respectivement aux mesures *DP*, *RP* et *RC* :

- la différence Δ des proportions, $\Delta = \pi_1 - \pi_0$,
- le rapport ξ des proportions, $\xi = \dfrac{\pi_1}{\pi_0}$,
- le rapport ψ de cotes des proportions, $\psi = \dfrac{\pi_1(1-\pi_0)}{\pi_0(1-\pi_1)}$.

On établit alors des relations similaires à celles présentées pour les taux à la section 5.1 du chapitre 5 :

$$\xi = \frac{A_1 n_0}{(m_1 - A_1)n_1} \tag{6.1}$$

$$\Delta = \frac{A_1 n - m_1 n_1}{n_1 n_0} \tag{6.2}$$

ou leurs relations réciproques :

$$A_1 = m_1\left(\frac{\xi n_1}{\xi n_1 + n_0}\right) \tag{6.3}$$

$$A_1 = \frac{\Delta n_1 n_0 + m_1 n_1}{n} \tag{6.4}$$

On peut aussi établir les relations suivantes

entre ψ et Δ :

$$\psi = \frac{(m_1 + \Delta n_0)(m_0 + \Delta n_1)}{(m_1 - \Delta n_1)(m_0 - \Delta n_0)} \tag{6.5}$$

entre ψ et ξ :

$$\psi = \frac{\xi[\xi n_1 - m_1 + n_0]}{[\xi(n_1 - m_1) + n_0]} \tag{6.6}$$

Ces relations nous permettent de passer facilement d'une distribution à l'autre. Ainsi, toute hypothèse sur ψ, donc sur A_1, trouve une hypothèse correspondante sur ξ et sur Δ. Et réciproquement.

De façon pratique, l'étude d'un *RP* sous l'hypothèse ξ revient à considérer la variable hypergéométrique A_1 de paramètres (n,n_1,m_1,ψ). De même, l'étude d'un *DP* sous l'hypothèse Δ revient à considérer cette même variable hypergéométrique.

De ces relations entre la variable hypergéométrique A_1 et les mesures d'association découle un premier corollaire. Les tests exacts sous les hypothèses $\Delta = 0$ et $\xi = 1$ sont équivalents à celui sous l'hypothèse $\psi = 1$. Il suffit donc de conduire le test exact sur $\psi = 1$ pour tester aussi les hypothèses $\Delta = 0$, $\xi = 1$.

Dans ce qui suit, nous présentons d'abord différents tests pour la comparaison de deux proportions observées ou plus précisément des tests sur l'association entre *X* et *Y*. Suit alors la présentation de certaines méthodes de calcul des intervalles de confiance pour les mesures Δ (*DP*), ξ (*RP*) et ψ (*RC*). Puis, dans une dernière section, nous présentons différents tests et méthodes de calcul pour les intervalles de confiance du *SMR* calculé sur des proportions.

6.2 TESTS POUR LA COMPARAISON DE DEUX PROPORTIONS

Dans cette section, nous présentons les tests statistiques les plus courants pour la comparaison de deux proportions en analyse simple : d'abord un test exact basé sur la distribution hypergéométrique (test de Fisher), puis quatre tests en approximation normale et, enfin, le test du rapport de vraisemblance. La plupart des tests présentés sont assez bien connus et souvent utilisés. Le test basé sur la transformation ARC fait exception. Il est présenté principalement pour son résultat qui peut être utilisé à la définition d'un intervalle de confiance d'une différence entre deux proportions dont les variances doivent être stabilisées.

6.2.1 TEST EXACT POUR LA COMPARAISON DE DEUX PROPORTIONS

Le test exact est directement construit sur la variable hypergéométrique A_1. Sous l'hypothèse nulle $\psi = 1$, le test unilatéral à droite basé sur la convention mi-*p* est alors de la forme :

$$p = \frac{C_{n_1}^{a_1} \times C_{n_0}^{(m_1 - a_1)}}{2C_n^{m_1}} + \sum_{i=a_1+1}^{\overline{A_1}} \frac{C_{n_1}^{i} \times C_{n_0}^{(m_1 - i)}}{C_n^{m_1}}$$

Remarquons que le test exact de Fisher est un test exact basé sur la convention de la valeur-*p* intégrale.

Le test bilatéral est simplement défini en ajoutant à la valeur-*p* unilatérale la somme des probabilités de toutes les valeurs de A_1 non considérées et qui sont également ou plus extrêmes que celle observée.

6.2.2 TESTS EN APPROXIMATION NORMALE POUR LA COMPARAISON DE DEUX PROPORTIONS

Quelle que soit la mesure considérée, *DP*, *RP* ou *RC*, les différents tests élaborés à partir de l'une ou l'autre de ces mesures d'association sont équivalents, sinon en résultats, du moins en nature. Nous présentons successivement les tests de Mantel-Haenszel, de Pearson, de Wald et celui basé sur la transformation ARC.

Test de Mantel-Haenszel

Le test de Mantel-Haenszel est construit à partir de l'approximation normale appliquée à la variable hypergéométrique A_1, sous l'hypothèse nulle $\psi = 1$. Le test se formule alors à partir de la statistique $Z^2 = \frac{\left[A_1 - E(A_1)\right]^2}{V(A_1)}$, qui obéit à une loi du χ^2 avec un degré de liberté.

Sous H_0, la valeur attendue $E(A_1)$ et la variance $V(A_1)$ de la variable hypergéométrique A_1 (tableau 6.1) correspondent respectivement à :

$$E(A_1) = \frac{m_1 n_1}{n} \text{ et } V(A_1) = \frac{m_1 m_0 n_1 n_0}{n^2 (n-1)}$$

En substituant ces valeurs de $E(A_1)$ et $V(A_1)$ dans l'expression de la variable Z^2 et en appliquant le test aux données observées, on obtient :

$$\chi_1^2 = \frac{(n-1)\left(a_1 n_0 - a_0 n_1\right)^2}{m_1 m_0 n_1 n_0}$$

Test de Pearson

Considérons la différence *DP* entre les deux proportions (tableau 6.1). Alors, la variable $Z^2 = \frac{\left[DP - E(DP)\right]^2}{V(DP)}$ obéit approximativement à une loi du χ^2 avec un degré de liberté. Sous l'hypothèse nulle, on a $E(DP) = 0$ et $V(DP) = \frac{m_1 m_0}{n^2}\left(\frac{1}{n_1} + \frac{1}{n_0}\right)$. En substituant ces valeurs respectives de $E(DP)$ et $V(DP)$ dans l'expression de la variable Z^2 et en appliquant le test aux données observées, on obtient : $\chi_1^2 = \frac{n\left(a_1 n_0 - a_0 n_1\right)^2}{m_1 m_0 n_1 n_0}$.

On peut aussi décrire ce test dans une forme plus usuelle : $\chi_1^2 = \sum \frac{(O-A)^2}{A}$, où la sommation se fait sur les cellules. Pour chacune d'elles, *O* représente la valeur observée et *A* la valeur attendue sous l'hypothèse nulle.

Ce test est similaire au test de Mantel-Haenszel (décrit précédemment). Les valeurs de ces deux tests sont dans le rapport $\frac{n-1}{n}$. En d'autres termes, sur de grands échantillons, les valeurs de ces deux tests sont similaires.

Test basé sur log(*RP*) ou test De Wald

Le test est construit en utilisant la transformation logarithmique du rapport de proportions, log(*RP*). Sous l'hypothèse nulle, on a $E[\log(RP)] = 0$ et par la méthode delta, $V[\log(RP)] = \left[\frac{b_1}{a_1 n_1} + \frac{b_0}{a_0 n_0}\right]$. Alors, le test appliqué aux données devient :

$$\chi_1^2 = \frac{\left[\log RP\right]^2}{V[\log RP]} = \frac{\left[\log RP\right]^2}{\frac{b_1}{a_1 n_1} + \frac{b_0}{a_0 n_0}}$$

Test basé sur log(*RC*) ou test De Wald

Le test est construit en utilisant la transformation logarithmique du rapport de cotes : log(*RC*). Sous l'hypothèse nulle, on a $E[\log(RC)] = 0$ et par la méthode delta, $V[\log RC] = \left[\frac{1}{a_1} + \frac{1}{a_0} + \frac{1}{b_1} + \frac{1}{b_0}\right]$. Alors, le test appliqué aux données devient :

$$\chi_1^2 = \frac{\left[\log RC\right]^2}{V[\log RC]} = \frac{\left[\log RC\right]^2}{\frac{1}{a_1} + \frac{1}{a_0} + \frac{1}{b_1} + \frac{1}{b_0}}$$

Test basé sur la transformation ARC

On considère la statistique $DP_T = \arcsin\sqrt{p_1} - \arcsin\sqrt{p_0}$, qui renvoie à la différence entre les deux proportions transformées par ARC. La variable $Z^2 = \frac{\left[DP_T - E(DP_T)\right]^2}{V(DP_T)}$ obéit approximativement à une loi du χ^2 avec un degré de liberté. La variance de DP_T correspond simplement à $V\left(DP_T\right) = \frac{1}{4}\left[\frac{1}{n_1} + \frac{1}{n_0}\right]$. Sous l'hypothèse nulle, on a $E(DP_T) = 0$. En substituant ces valeurs respectives de $E(DP_T)$ et $V(DP_T)$ dans l'expression de la variable Z^2, et en appliquant le test aux données observées, on obtient :

$$\chi_1^2 = \frac{\left[\arcsin\sqrt{p_1} - \arcsin\sqrt{p_0}\right]^2}{\frac{1}{4}\left[\frac{1}{n_1} + \frac{1}{n_0}\right]}$$

$$= \frac{4n_1n_0\left[\arcsin\sqrt{p_1} - \arcsin\sqrt{p_0}\right]^2}{n}$$

On désigne par χ_T^1 la valeur du khi-carré de ce test. Ce test est analogue à celui développé pour la racine carrée d'un taux (section 5.2.2, du chapitre 5).

6.2.3 TEST DU RAPPORT DE VRAISEMBLANCE POUR LA COMPARAISON DE DEUX PROPORTIONS

Test du rapport de vraisemblance

Revenons au schéma du tableau 6.1, qui décrit les distributions de deux variables binomiales indépendantes, chacune de paramètres π_j et n_j, $j = 1$ ou 0. Alors, sous l'hypothèse nulle $\Delta = 0$ (ou $\xi = 1$ ou $\psi = 1$), les proportions π_j sont égales ($\pi_1 = \pi_0 = \pi$). La fonction de vraisemblance FV_0 de ces données peut se décrire comme :

$$FV_0 = \left[C_{n_1}^{a_1}\pi^{a_1}(1-\pi)^{(n_1-a_1)}\right] \times \left[C_{n_0}^{a_0}\pi^{a_0}(1-\pi)^{(n_0-a_0)}\right]$$

L'hypothèse nulle se résume en un paramètre unique π qui sera estimé par a/n.

Par ailleurs, sous une hypothèse H_1 spécifiant des valeurs spécifiques π_j pour chacune des deux proportions, la fonction de vraisemblance FV_1 de ces données se présente comme :

$$FV_1 = \left[C_{n_1}^{a_1}\pi_1^{a_1}(1-\pi_1)^{(n_1-a_1)}\right] \times \left[C_{n_0}^{a_0}\pi_0^{a_0}(1-\pi_0)^{(n_0-a_0)}\right]$$

Les valeurs des paramètres sont fixées par les données elles-mêmes :

$$\pi_1 = \frac{a_1}{n_1} \quad \text{et} \quad \pi_0 = \frac{a_0}{n_0}$$

Le test du rapport de vraisemblance se traduit comme suit :

$$\chi_1^2 = 2\left[\begin{array}{l} a_1 \log\left(\frac{\pi_1}{\pi}\right) + a_0 \log\left(\frac{\pi_0}{\pi}\right) + (n_1 - a_1)\log\left(\frac{1-\pi_1}{1-\pi}\right) \\ +(n_0 - a_0)\log\left(\frac{1-\pi_0}{1-\pi}\right) \end{array}\right]$$

$$= 2\sum_{j=1}^{J}\left[O_j \log\left(\frac{O_j}{A_j}\right)\right]$$

Pour la cellule j, O_j correspond à la valeur observée et A_j à la valeur attendue sous l'hypothèse nulle. La sommation se fait sur les 4 cellules du tableau.6.1.

Exemple 6.1

Considérons les données du tableau 6.2, décrivant les résultats d'une étude portant sur l'association entre un facteur d'exposition X et la maladie Y. L'étude a pu être conduite sur une population fermée de type cohorte ou sur une population transversale, composée de 561 individus. Les proportions peuvent correspondre à des incidences cumulatives ou à des prévalences. Alors, on veut tester l'association entre X et Y.

Tableau 6.2

	Exposition à X		Total
	$X = 1$	$X = 0$	
$Y = 1$	25	38	63
$Y = 0$	101	397	498
Total	126	435	561

Test exact

La valeur-p est calculée dans la convention mi-p comme :

$$p = \frac{1}{C_{561}^{63}}\left[1/2 C_{126}^{25} C_{435}^{38} + \sum_{i=26}^{63} C_{126}^{i} C_{435}^{63-i}\right]$$

$$= 0{,}00054775$$

Ainsi, la valeur-p unilatérale à droite est égale à 0,00054775. La valeur-p bilatérale correspondante est de 0,00087527.

Test en approximation normale

Test de Mantel-Haenszel

$$\chi_1^2 = \frac{(n-1)\left(a_1 n_0 - a_0 n_1\right)^2}{m_1 m_0 n_1 n_0}$$
$$= \frac{(561-1)\left[25\times 435 - 38\times 126\right]^2}{63\times 498\times 126\times 435}$$
$$= 12,06604$$

Pour cette valeur du khi-carré à un degré de liberté, $p = 0,000513$.

Test de Pearson

On veut tester la l'hypothèse nulle $\Delta = 0$ sur les données du tableau 6.2. La différence *DP* observée est de 0,1111. On applique le test de Pearson.

$$\chi_1^2 = \sum \frac{(O-A)^2}{A}$$
$$= \frac{\left(25 - \frac{(63\times 126)}{561}\right)^2}{\frac{(63\times 126)}{561}} + \frac{\left(101 - \frac{(498\times 126)}{561}\right)^2}{\frac{(498\times 126)}{561}}$$
$$+ \frac{\left(38 - \frac{(63\times 435)}{561}\right)^2}{\frac{(63\times 435)}{561}} + \frac{\left(397 - \frac{(498\times 435)}{561}\right)^2}{\frac{(498\times 435)}{561}}$$
$$= 12,0876$$

Pour cette valeur 12,0876 du khi-carré, à un degré de liberté, on a $p = 0,000508$.

Test basé sur log(RP)

$$\chi_1^2 = \frac{\left(\log RP\right)^2}{\left[\frac{b_1}{a_1 n_1} + \frac{b_0}{a_0 n_0}\right]}$$
$$= \frac{\left(\log 2,2713\right)^2}{\left[\frac{101}{25\times 126} + \frac{397}{38\times 435}\right]}$$
$$= 12,0003$$

Pour cette valeur de 12,0003 du khi-carré, à un degré de liberté, $p = 0,000532$.

TEST BASÉ SUR LOG(*RC*)

$$\begin{aligned}\chi_1^2 &= \frac{\left[\log RC\right]^2}{\frac{1}{a_1}+\frac{1}{a_0}+\frac{1}{b_1}+\frac{1}{b_0}} \\ &= \frac{\left[\log 2,5860\right]^2}{\frac{1}{25}+\frac{1}{38}+\frac{1}{101}+\frac{1}{397}} \\ &= 11,4649\end{aligned}$$

Pour cette valeur du khi-carré, à un degré de liberté, $p = 0{,}000709$.

TEST BASÉ SUR LES PROPORTIONS TRANSFORMÉES PAR ARC

Le test conduit aux résultats suivants :

$$\begin{aligned}\chi_1^2 &= \frac{4n_1n_0\left[\arcsin\sqrt{p_1}-\arcsin\sqrt{p_0}\right]^2}{n} \\ &= \frac{4\times126\times435\left[\arcsin\sqrt{25/126}-\arcsin\sqrt{38/435}\right]^2}{561} \\ &= 10,21\end{aligned}$$

Pour cette valeur 10,21 du khi-carré, à un degré de liberté, on a $p = 0{,}001398$.

TEST DU RAPPORT DE VRAISEMBLANCE (RV)

$$\begin{aligned}\chi_1^2 &= 2\left[\sum_{j=1}^{4} O_j \log\left(\frac{O_j}{A_j}\right)\right] \\ &= 2\left[\begin{array}{l}25\log\left(\frac{25}{14,15}\right)+38\log\left(\frac{38}{48,85}\right) \\ +101\log\left(\frac{101}{111,85}\right)+397\log\left(\frac{397}{386,15}\right)\end{array}\right] \\ &= 2\left[14,23-9,54-10,31+11,00\right] \\ &= 10,76\end{aligned}$$

Pour cette valeur 10,76 du khi-carré, à un degré de liberté, $p = 0{,}001037$.

Nous présentons dans le tableau 6.3 les résultats de ces différents tests.

TABLEAU 6.3

Méthode	Khi-carré	Valeur-*p*		
		Unilatérale	Bilatérale	Programme
Exacte	–	0,000548	0,000875	PR6.1
Mantel-Haenszel	12,07	0,000256	0,000513	PR6.2
Pearson	12,09	0,000254	0,000508	
Basée sur log(*RP*)	12,00	0,000266	0,000532	PR6.3
Basée sur log(*RC*)	11,46	0,000355	0.000709	PR6.4
Transformation ARC	10,21	0,000699	0,001398	PR6.5
RV	10,76	0,000518	0,001037	PR6.6 OU PR6.2

Nous remarquons alors une bonne concordance entre le test exact et celui du rapport de vraisemblance pour la valeur-*p* unilatérale. Par ailleurs, tous les tests conduisent à une faible valeur-*p* bilatérale ($p < 0{,}002$) indiquant une faible compatibilité entre les résultats observés et l'hypothèse nulle.

◆

6.3 INTERVALLES DE CONFIANCE DES MESURES D'ASSOCIATION POUR DEUX PROPORTIONS

Dans cette section, nous présentons les intervalles de confiance des différentes mesures d'association déduites de la comparaison entre deux proportions. Plus spécifiquement, nous présentons les intervalles de confiance pour le rapport de cotes ψ, le rapport ξ et la différence Δ de deux proportions. Pour chacune de ces mesures, nous présentons la méthode exacte de calcul, quelques méthodes en approximation normale et la méthode du rapport de vraisemblance.

6.3.1 INTERVALLE DE CONFIANCE POUR LE RAPPORT DE COTES *RC* (ψ)

MÉTHODE EXACTE

Le rapport de cotes ψ s'avère un paramètre naturel de la variable hypergéométrique A_1. L'intervalle de confiance de ce paramètre, pour un niveau de $100(1 - \alpha)\,\%$, est donc calculé en référant à la distribution hypergéométrique de la variable A_1 (tableau 6.1).

Ainsi, la limite inférieure ψ_{inf} de l'intervalle est la valeur de ψ qui satisfait l'équation

$$\sum_{i=a_1}^{\overline{A_1}} \frac{C_{n_1}^{i} C_{n_0}^{(m_1-i)} \psi^{i}}{\sum_{j=\underline{A_1}}^{\overline{A_1}} C_{n_1}^{j} C_{n_0}^{(m_1-j)} \psi^{j}} = \alpha/2$$

De façon analogue, on définit la limite supérieure ψ_{sup} comme la valeur de ψ qui satisfait l'équation :

$$\sum_{i=\underline{A_1}}^{a_1} \frac{C_{n_1}^{i} C_{n_0}^{(m_1-i)} \psi^{i}}{\sum_{j=\underline{A_1}}^{\overline{A_1}} C_{n_1}^{j} C_{n_0}^{(m_1-j)} \psi^{j}} = \alpha/2$$

De telles équations ne peuvent être résolues que par des méthodes itératives. Nous appliquons aussi à ces calculs la convention mi-*p*.

Valeurs de A1 correspondant aux limites exactes du rapport de cotes ψ

On désigne par A_{1inf} et A_{1sup} les valeurs de la variable A_1 qui correspondent respectivement à $E(A_1 \mid \psi_{inf})$ et $E(A_1 \mid \psi_{sup})$, pour les valeurs de ψ calculées précédemment.

La valeur attendue $A_{1inf} = E\left(A_1 \mid \psi_{inf}\right)$ correspond à la valeur de A_1 qui solutionne l'équation quadratique $\dfrac{A_1(n_0 - m_1 + A_1)}{(m_1 - A_1)(n_1 - A_1)} = \psi_{inf}$.

Concrètement, cette valeur est donnée par $A_{1inf} = \dfrac{-V - \sqrt{V^2 - 4UW}}{2U}$ où $U = \psi_{inf} - 1$, $V = -[\psi_{inf}(n_1 + m_1) + (n_0 - m_1)]$ et $W = \psi_{inf} n_1 m_1$.

La valeur attendue A_{1sup} est obtenue de façon analogue à partir de la limite ψ_{sup}.

Les deux valeurs A_{1inf} et A_{1sup} permettent de déduire facilement les intervalles de confiance exacts des mesures *RP* et *DP*, comme on le verra plus loin.

Méthode en approximation normale

Méthode simple

Le calcul de l'intervalle de confiance de ψ peut se faire à l'aide de la transformation logarithmique de cette mesure. Les limites de l'intervalle de confiance de log(ψ) sont d'abord calculées par approximation normale : $\log(RC) \pm z_{\alpha/2}\sqrt{V(\log RC)}$. Puis, les limites de confiance de ψ sont déduites par transformation inverse : $RC \times \exp\left[\pm z_{\alpha/2}\sqrt{V(\log RC)}\right]$.

La variance du log(*RC*) estimée par la méthode delta correspond simplement à $\frac{1}{a_1}+\frac{1}{a_0}+\frac{1}{b_1}+\frac{1}{b_0}$. Ainsi, les limites de confiance du rapport de cotes ψ sont explicitement décrites comme :

$$\psi_{\text{inf}} = RC \times \exp\left(-z_{\alpha/2}\sqrt{\frac{1}{a_1}+\frac{1}{a_0}+\frac{1}{b_1}+\frac{1}{b_0}}\right)$$

$$\psi_{\text{sup}} = RC \times \exp\left(+z_{\alpha/2}\sqrt{\frac{1}{a_1}+\frac{1}{a_0}+\frac{1}{b_1}+\frac{1}{b_0}}\right)$$

Méthode basée sur le résultat d'un test

En utilisant le résultat du test de Mantel-Haenszel calculé sur les mêmes données, on peut déduire une estimation pour la variance V[log(*RC*)].

Ainsi, sous l'hypothèse nulle, la relation $\chi^2_{MH} = \frac{[\log(RC)]^2}{V[\log(RC)]}$ est asymptotiquement correcte. On en déduit alors une estimation de la variance V[log(*RC*)] en autant que cette variance ne soit pas trop instable :

$$V(\log RC) \approx \frac{[\log RC]^2}{\chi^2_{MH}}$$

Les limites de confiance de log (ψ) correspondent alors à :

$$\log RC \pm z_{\alpha/2}\sqrt{\frac{[\log RC]^2}{\chi^2_{MH}}} \text{ ou } \log RC\left(1 \pm \frac{z_{\alpha/2}}{\chi_{MH}}\right)$$

Par transformation inverse, on obtient les limites inférieure et supérieure de l'intervalle de confiance du rapport de cotes ψ :

$$\psi_{\inf} = RC^{\left(1-\frac{z_{\alpha/2}}{\chi_{MH}}\right)}$$

$$\psi_{\inf} = RC^{\left(1+\frac{z_{\alpha/2}}{\chi_{MH}}\right)}$$

Méthode du rapport de vraisemblance (RV)

Soulignons que le logarithme du rapport de cotes *RC* correspond au paramètre β du modèle $\text{logit}[\pi(X)] = \alpha + \beta X$, qui décrit le lien entre X et le log de la cote $\frac{\pi(X)}{1-\pi(X)}$. On a $\psi = e^{\beta}$. Si on pose $\pi(X=1) = \pi_1$ et $\pi(X=0) = \pi_0$, la fonction de vraisemblance correspondant à ce modèle est donnée par :

$$FV(\alpha,\beta) = \pi_1^{a_1}(1-\pi_1)^{b_1}\pi_0^{a_0}(1-\pi_0)^{b_0}$$

$$= \left(\frac{e^{\alpha+\beta}}{1+e^{\alpha+\beta}}\right)^{a_1}\left(\frac{1}{1+e^{\alpha+\beta}}\right)^{b_1}\left(\frac{e^{\alpha}}{1+e^{\alpha}}\right)^{a_0}\left(\frac{1}{1+e^{\alpha}}\right)^{b_0}$$

Le logarithme $L(\alpha,\beta)$ de la fonction de vraisemblance $FV(\alpha,\beta)$ se présente comme :

$$L(\alpha,\beta) = (a_1+a_0)\alpha + a_1\beta - (a_1+b_1)\log(1+e^{\alpha+\beta}) - (a_0+b_0)\log(1+e^{\alpha})$$

$L(\alpha,\beta)$ est maximal si $\alpha = \log\frac{a_0}{b_0}$ et $\beta = \log RC$.

Pour retrouver les limites de ψ, il suffit de résoudre l'équation logarithmique $2[L - L(\psi)] - \chi^2_{1,1-\alpha} = 0$, où L correspond au maximum de la fonction $L(\alpha,\beta)$ et $L(\psi)$ à la valeur de cette fonction évaluée à ψ et maximisée pour α.

La fonction $L(\psi)$ se présente comme :

$$L(\psi) = a_1\log(\psi) - (a_1+b_1)\log(1+\psi e^{\alpha}) + (a_1+a_0)\alpha - (a_0+b_0)\log(1+e^{\alpha})$$

et L comme :

$$L = a_1 \log(RC) - (a_1 + b_1)\log\left(1 + RC \times \frac{a_0}{b_0}\right) + (a_1 + a_0)\log\left(\frac{a_0}{b_0}\right) - (a_0 + b_0)\log\left(1 + \frac{a_0}{b_0}\right)$$

Les limites de confiance de ψ sont alors déterminées par itération en retraçant les deux valeurs de ψ qui satisfont l'équation logarithmique.

Dans l'itération, pour chaque valeur d'essai sur ψ, il faut déterminer la valeur de e^{α} qui maximise la fonction $L(\alpha,\beta)$. On peut montrer que

$$e^{\alpha} = \frac{-V + \sqrt{V^2 - 4UW}}{2W} \text{ où } \begin{cases} U = \psi(b_1 + b_0) \\ V = \psi(b_1 - a_0) + b_0 - a_1 \\ W = -(a_1 + a_0) \end{cases}$$

Exemple 6.2

Considérons les données du tableau 6.2. On veut calculer l'intervalle de confiance à 95 % du rapport de cotes ψ, dont la valeur est estimée à 2,59 :

$$RC = \frac{25 \times 397}{38 \times 101} = 2,58598$$

Méthode exacte

Les limites de confiance inférieure et supérieure de ψ sont données par les valeurs de ψ qui satisfont respectivement les relations suivantes :

$$\sum_{i=25}^{63} \frac{C_{126}^{i} C_{435}^{(63-i)} \psi^{i}}{\sum_{j=0}^{63} C_{126}^{j} C_{435}^{(63-j)} \psi^{j}} = 0,025 \text{ et } \sum_{i=0}^{25} \frac{C_{126}^{i} C_{435}^{(63-i)} \psi^{i}}{\sum_{j=0}^{63} C_{126}^{j} C_{435}^{(63-j)} \psi^{j}} = 0,025$$

Ces équations ne peuvent être résolues que par itération. En appliquant la convention mi-*p*, les limites inférieure et supérieure déduites sont les suivantes :

$$\psi_{\text{inf}} = 1,4746$$
$$\psi_{\text{sup}} = 4,4701$$

Les valeurs $A_{1\,\text{inf}}$ et $A_{1\,\text{sup}}$ correspondant à ces limites sont calculées de la façon suivante.

- Pour $A_{1\,\text{inf}}$, on considère $\psi_{\text{inf}} = 1,4746$ et les valeurs associées *U*, *V* et *W* :
  $U = 1,4746 - 1 = 0,4746$
  $V = -[1,4746(126 + 63) + (435 - 63)] = 650,70$
  $W = 1,4746 \times 126 \times 63 = 11\,705,40$.

La valeur $A_{1\,\text{inf}}$ recherchée est donc

$$\begin{aligned} A_{1\,\text{inf}} &= \frac{-V - \sqrt{V^2 - 4UW}}{2U} \\ &= \frac{650,7 - \sqrt{650,7^2 - 4 \times 0,4746 \times 11705,4}}{2 \times 0,4746} \\ &= 18,23 \end{aligned}$$

- Pour $A_{1\,\text{sup}}$, on considère $\psi_{\text{sup}} = 4,4701$. Le calcul conduit alors à $A_{1\,\text{sup}} = 32,10$.

Approximation normale

Méthode simple

Les limites de confiance calculées par le méthode de Wald sont de :

$$\psi_{\text{inf}} = 2,5860 \times \exp\left(-1,96\sqrt{\frac{1}{25} + \frac{1}{38} + \frac{1}{101} + \frac{1}{397}}\right) = 1,4920$$

$$\psi_{\text{sup}} = 2,5860 \times \exp\left(+1,96\sqrt{\frac{1}{25} + \frac{1}{38} + \frac{1}{101} + \frac{1}{397}}\right) = 4,4820$$

Méthode basée sur le résultat d'un test

La valeur du khi-carré de Mantel-Haenszel est de 12,07 (tableau 6.3). Sur la base de ce résultat, les limites de confiance à 95 % de ψ sont calculées comme :

$$\psi_{\text{inf}} = 2,5860^{\left(1 - \frac{1,96}{\sqrt{12,07}}\right)} = 1,5130$$

$$\psi_{\text{sup}} = 2,5860^{\left(1 + \frac{1,96}{\sqrt{12,07}}\right)} = 4,4200$$

Méthode du rapport de vraisemblance (RV)

Il faut résoudre par itération l'équation logarithmique $2[L - L(\psi)] - \chi^2_{1,1-\alpha} = 0$. Pour un niveau de confiance à 95 %, le $\chi^2_{1,(1-0,05)}$ est égal à 3,84.

La valeur de $L(\psi)$, aussi dépendante de α, correspond à :

$$L(\psi) = 25\log(\psi) - 126\log(1 + \psi e^{\alpha}) + 63\log(e^{\alpha}) - 435\log(1 + e^{\alpha})$$

La valeur numérique de L correspond au maximum de la fonction $L(\alpha,\beta)$, où $\beta = \log\psi$:

$$\begin{aligned} L &= 25\log(2,5860) - 126\log(1+2,5860\times 38/397) \\ &\quad + 63\log(38/397) - 435\log(1+38/397) \\ &= -191,70. \end{aligned}$$

L'équation à résoudre est alors : $25\log(\psi) - 126\log(1+\psi e^{\alpha}) + 63\log(e^{\alpha})$
$-435\log(1+e^{\alpha}) + 191,70 + 1/2\times 3,84 = 0$

Dans la procédure d'itération sur ψ, la valeur de la fonction $L(\psi)$ est déterminée en remplaçant ψ par sa valeur d'itération et α par son estimation du maximum de vraisemblance correspondant.

Les procédures GENMOD ou LOGISTIC de SAS permettent d'obtenir assez facilement ces limites de confiance :

$$\psi_{inf} = 1,4788$$
$$\psi_{sup} = 4,4618$$

On retrouve dans le tableau 6.4 les limites de confiance à 95 % de ψ calculées suivant les différentes méthodes présentées.

TABLEAU 6.4

Méthode	Intervalle de confiance à 95 %	Programme
Exacte	[1,4746 ; 4,4701]	PR6.7
Simple	[1,4920 ; 4,4820]	PR6.8
Résultat d'un test	[1,5129 ; 4,4203]	PR6.9
RV	[1,4788 ; 4,4618]	PR6.10 OU PR6-11

◆

6.3.2 INTERVALLE DE CONFIANCE POUR LE RAPPORT ξ DE DEUX PROPORTIONS

MÉTHODE EXACTE

Utilisant la relation (6.1) entre la variable hypergéométrique A_1 et la mesure ξ, il est facile de déterminer l'intervalle de confiance exact de cette mesure. En effet, connaissant les valeurs A_{1inf} et A_{1sup}, il suffit d'appliquer cette relation pour déterminer les limites de confiance ξ_{inf} et ξ_{sup} de ξ. On a :

$$\xi_{\text{inf}} = \frac{A_{1\,\text{inf}} n_0}{(m_1 - A_{1\,\text{inf}}) n_1}$$

$$\xi_{\text{sup}} = \frac{A_{1\,\text{sup}} n_0}{(m_1 - A_{1\,\text{sup}}) n_1}$$

Méthodes en approximation normale

Méthode simple

Comme pour le rapport de taux (section 5.3.1), l'intervalle de confiance du rapport ξ de proportions, dont *RP* est une estimation, peut aussi se calculer à l'aide de la transformation logarithmique de cette mesure. Les limites de confiance de ξ sont alors déduites par transformation inverse :

$$RP \exp\left[\pm z_{\alpha/2} \sqrt{V(\log RP)}\right].$$

L'estimation de la variance $V[\log RP]$ est obtenue par la méthode delta :

$$V[\log RP] = \frac{b_1}{n_1 a_1} + \frac{b_0}{n_0 a_0}$$

Les limites de confiance de RP sont alors simplement décrites comme :

$$\xi_{\text{inf}} = RP \times \exp\left(-z_{\alpha/2} \sqrt{\frac{b_1}{n_1 a_1} + \frac{b_0}{n_0 a_0}} \right)$$

$$\xi_{\text{sup}} = RP \times \exp\left(+z_{\alpha/2} \sqrt{\frac{b_1}{n_1 a_1} + \frac{b_0}{n_0 a_0}} \right)$$

Méthode basée sur le résultat d'un test

Le calcul de l'intervalle de confiance peut aussi se faire à partir du résultat d'un test. Nous proposons d'utiliser le résultat χ^2_{MH} du test de Mantel-Haenszel décrit à la section 6.2.2. Alors, les limites de confiance pour le niveau $100(1 - \alpha)\,\%$ pour la mesure ξ sont simplement décrites comme :

$$\xi_{\text{inf}} = RP^{\left(1 - \frac{z_{\alpha/2}}{\chi_{MH}}\right)}$$

$$\xi_{\text{sup}} = RP^{\left(1 + \frac{z_{\alpha/2}}{\chi_{MH}}\right)}$$

Méthode du rapport de vraisemblance

Soulignons que le logarithme du rapport de proportions ξ correspond au paramètre β du modèle $\log[\pi(X)] = \alpha + \beta X$, qui décrit le lien entre X et le log de la proportion $\pi(X)$: $\xi = e^{\beta}$.

Si on pose $\pi(X = 1) = \pi_1$ et $\pi(X = 0) = \pi_0$, la fonction de vraisemblance correspondant à ce modèle est donnée par :

$$\begin{aligned} FV(\alpha,\beta) &= \left[C_{n_1}^{a_1} \pi_1^{a_1} (1-\pi_1)^{(n_1-a_1)} \right] \times \left[C_{n_0}^{a_0} \pi_0^{a_0} (1-\pi_0)^{(n_0-a_0)} \right] \\ &= \left[C_{n_1}^{a_1} \left(e^{\alpha+\beta}\right)^{a_1} \left(1-e^{\alpha+\beta}\right)^{(n_1-a_1)} \right] \times \left[C_{n_0}^{a_0} \left(e^{\alpha}\right)^{a_0} \left(1-e^{\alpha}\right)^{(n_0-a_0)} \right] \end{aligned}$$

Le logarithme $L(\alpha,\beta)$ de cette fonction de vraisemblance se présente comme :

$$\begin{aligned} L(\alpha,\beta) &= a_1(\alpha+\beta) + (n_1 - a_1)\log(1-e^{\alpha+\beta}) \\ &\quad + a_0\alpha + (n_0 - a_0)\log(1-e^{\alpha}) + K \end{aligned}$$

où K est une valeur indépendante de α et β.

Par rapport à ξ, la fonction de vraisemblance se décrit comme :

$$\begin{aligned} L(\alpha,\xi) &= a_1[\alpha + \log(\xi)] + (n_1 - a_1)\log(1-\xi e^{\alpha}) \\ &\quad + a_0\alpha + (n_0 - a_0)\log(1-e^{\alpha}) + K \end{aligned}$$

Les limites de l'intervalle de confiance de niveau $100(1 - \alpha)$ % correspondent aux deux valeurs de ξ qui satisfont l'équation logarithmique $2\left[L - L(\xi)\right] - \chi^2_{1,1-\alpha} = 0$, où L correspond au maximum de la fonction $L(\alpha,\xi)$ et $L(\xi)$ est la valeur de la fonction de vraisemblance évaluée à la valeur d'itération ξ et maximisée pour α. On rappelle que la fonction de vraisemblance $L(\alpha,\beta)$ atteint ce maximum L pour $\xi = RP$ et $\alpha = \log\left(\frac{a_0}{n_0}\right)$.

Il suffit alors de résoudre pour ξ, par itération, l'équation logarithmique.

Exemple 6.3

Considérons les données du tableau 6.2. On veut calculer l'intervalle de confiance à 95 % du rapport ξ de proportions, dont la valeur est estimée à 2,27 : $RP = \frac{25 \times 435}{38 \times 126} = 2,2713.$

Méthode exacte

Les valeurs $A_{1\,\text{inf}}$ et $A_{1\,\text{sup}}$ sont respectivement de 18,2312 et 32,0985 (voir exemple 6.2, méthode exacte).

En appliquant la relation (6.1), on déduit pour ξ les limites de confiance à 95 % suivantes :

$$\xi_{\text{inf}} = \frac{18,2312 \times 435}{(63 - 18,2312) \times 126} = 1,4059$$

$$\xi_{\text{sup}} = \frac{32,0985 \times 435}{(63 - 32,0985) \times 126} = 3,5861$$

Méthodes en approximation normale

Méthode simple

Les limites de confiance à 95 % de ξ calculées par le méthode de Wald sont de :

$$\xi_{\text{inf}} = 2,2713 \times \exp\left(-1,96 \sqrt{\frac{101}{25 \times 126} + \frac{397}{38 \times 435}} \right)$$

$$= 1,4279$$

$$\xi_{\text{sup}} = 2,2713 \times \exp\left(+1,96 \sqrt{\frac{101}{25 \times 126} + \frac{397}{38 \times 435}} \right)$$

$$= 3,6129$$

Méthode basée sur le résultat d'un test

Les limites de confiance à 95 % de ξ calculées par le méthode basée sur le résultat du test de Mantel-Haenszel (voir tableau 6.3) sont les suivantes :

$$\xi_{\text{inf}} = 2,2713^{\left(1 - \frac{1,96}{\sqrt{12,0660}}\right)}$$

$$= 1,4297$$

$$\xi_{\text{sup}} = 2,2713^{\left(1 + \frac{1,96}{\sqrt{12,0660}}\right)}$$

$$= 3,6083$$

Méthode du rapport de vraisemblance

Il faut résoudre par méthode itérative l'équation logarithmique $2\left[L - L(\xi)\right] - \chi^2_{1,1-\alpha} = 0$. Les valeurs de $L(\xi)$ et L sont données respectivement par :

$$L(\xi) = 25[\alpha + \log(\xi)] + (126 - 25)\log(1 - \xi e^{\alpha})$$

$$+38\alpha + (435 - 38)\log(1 - e^{\alpha}) + K$$

et par le maximum de cette fonction, atteint lorsque $\xi = RP$ et $e^{\alpha} = 38/435$:

$$L = 25\left(\log\frac{38}{435} + \log 2,2713\right) + (126-25)\log\left(1-\frac{25}{126}\right)$$
$$+ 38\log\left(\frac{38}{435}\right) + (435-38)\log\left(1-\frac{38}{435}\right) + K$$
$$= -1916970 + K$$

Dans la procédure d'itération sur ξ, la valeur de la fonction $L(\xi)$ est déterminée en remplaçant ξ par sa valeur d'itération et α par son estimation du maximum de vraisemblance correspondant.

L'utilisation de la procédure GENMOD de SAS, dans le cadre de la distribution binomiale avec la fonction de lien log, permet d'obtenir assez facilement ces limites de confiance.

On retrouve au tableau 6.5 les limites de confiance à 95 % de ξ calculées suivant les différentes méthodes présentées.

TABLEAU 6.5

Méthode	Intervalle de confiance à 95 %	Programme
Exacte	[1,4059 ; 3,5861]	PR6.12
Simple	[1,4279 ; 3,6129]	PR6.13
Basé sur un test	[1,4297 ; 3,6083]	PR6.14
RV	[1,4092 ; 3,5902]	PR6.15 OU PR6.16

◆

6.3.3 INTERVALLE DE CONFIANCE POUR LA DIFFÉRENCE Δ DE DEUX PROPORTIONS

MÉTHODE EXACTE

En utilisant la relation (6.2) entre la variable hypergéométrique A_1 et la différence Δ entre deux proportions, il est facile de déterminer l'intervalle de confiance exact de cette mesure. En effet, si on connaît les limites A_{inf} et A_{sup}, il suffit d'appliquer la relation (6.2) pour déterminer celles de Δ : Δ_{inf} et Δ_{sup}.

On a :

$$\Delta_{inf} = \frac{A_{1inf}n - m_1 n_1}{n_1 n_0}$$

$$\Delta_{sup} = \frac{A_{1sup}n - m_1 n_1}{n_1 n_0}$$

Méthodes en approximation normale

Méthode simple

Cet intervalle est simplement obtenu par l'expression : $DP \pm z_{\alpha/2}\sqrt{V(DP)}$. Suivant les notations du tableau 6.1, la variance $V(DP)$ correspond à :

$$V(DP) = \left(\frac{p_1 q_1}{n_1} + \frac{p_0 q_0}{n_0}\right)$$
$$= \frac{a_1(n_1 - a_1)}{n_1^3} + \frac{a_0(n_0 - a_0)}{n_0^3}$$

Transformation ARC

La méthode est basée sur le résultat du test des proportions transformées par ARC : χ^2_T (section 6.2.2). On suppose que le khi-carré sur la différence DP, dont la variance a été stabilisée, est approximativement égal à χ^2_T. Cette supposition permet d'obtenir une estimation de la variance stabilisée pour DP : $V_T(DP)$. Ainsi, sous H_0,

$$\frac{DP^2}{V_T(DP)} \approx \chi^2_T \quad \Rightarrow \quad \mathrm{V_T}(DP) \approx \frac{DP^2}{\chi^2_T}$$

L'intervalle de confiance de Δ où la variance de DP a été stabilisée se présente donc comme :

$$\Delta_{\inf} = DP - z_{\alpha/2}\sqrt{V_T(DP)}$$
$$= DP\left(1 - \frac{z_{\alpha/2}}{\chi_T}\right)$$
$$\Delta_{\sup} = DP + z_{\alpha/2}\sqrt{V_T(DP)}$$
$$= DP\left(1 + \frac{z_{\alpha/2}}{\chi_T}\right)$$

Intervalle de confiance du rapport de vraisemblance

Soulignons que la différence Δ entre deux proportions correspond au paramètre β du modèle linéaire $\pi(X) = \alpha + \beta X$. Si, pour simplifier, on pose $\pi(X = 1) = \pi_1$, $\pi(X = 0) = \pi_0$ et $\beta = \Delta$, alors la fonction de vraisemblance correspondant à ce modèle linéaire est donnée par :

$$FV(\alpha,\beta) = \left[C_{n_1}^{a_1} \pi_1^{a_1} (1-\pi_1)^{n_1-a_1} \right] \times \left[C_{n_0}^{a_0} \pi_0^{a_0} (1-\pi_0)^{n_0-a_0} \right]$$
$$= \left[C_{n_1}^{a_1} (\alpha+\Delta)^{a_1} (1-\alpha-\Delta)^{n_1-a_1} \right] \times \left[C_{n_0}^{a_0} \alpha^{a_0} (1-\alpha)^{n_0-a_0} \right]$$

où Δ a été substitué à β.

Le logarithme $L(\alpha,\Delta)$ de la fonction de vraisemblance $FV(\alpha,\Delta)$ se présente comme :

$$L(\alpha,\Delta) = a_1 \log(\alpha+\Delta) + (n_1 - a_1)\log(1-\alpha-\Delta) + a_0 \log\alpha + (n_0 - a_0)\log(1-\alpha) + K$$

où K est une valeur indépendante de α et Δ.

Alors, les limites de l'intervalle de confiance de niveau $100(1-\alpha)$ % pour Δ correspondent aux deux valeurs de Δ qui solutionnent l'équation de vraisemblance suivante :

$$2\left[L - L(\Delta)\right] - \chi^2_{1,1-\alpha} = 0$$

où L est la valeur maximale de la fonction de vraisemblance $L(\alpha,\Delta)$:

$$L = a_1 \log(p_1) + (n_1 - a_1)\log(1-p_1) + a_0 \log p_0 + (n_0 - a_0)\log(1-p_0) + K$$

et $L(\Delta)$ est la valeur de la fonction de vraisemblance évaluée à la valeur Δ et maximisée pour α.

La résolution de cette équation de vraisemblance se fait par itération.

EXEMPLE 6.4

Considérons les données du tableau 6.2. On veut calculer l'intervalle de confiance à 95 % de la différence Δ entre les deux proportions. La valeur de Δ est estimée à 0,1111 :

$$DP = \frac{25}{126} - \frac{38}{435} = 0{,}1111$$

MÉTHODE EXACTE

Rappelons les valeurs $A_{1\,\text{inf}}$ et $A_{1\,\text{sup}}$ qui sont respectivement de 18,2312 et 32,0985 (voir l'exemple 6.3).

En appliquant la relation (6.2), on déduit pour Δ les limites de confiance à 95 % suivantes :

$$\Delta_{\text{inf}} = \frac{18,2312 \times 561 - 63 \times 126}{126 \times 435} = 0,04178$$

$$\Delta_{\text{sup}} = \frac{32,0985 \times 561 - 63 \times 126}{126 \times 435} = 0,18371$$

Méthodes en approximation normale

Méthode simple

Les limites de confiance à 95 % du *DP* calculées par la méthode simple sont :

$$\Delta_{\text{inf}} = 0,1111 - 1,96\sqrt{\left(\frac{25 \times 101}{126^3} + \frac{38 \times 397}{435^3}\right)} = 0,0366$$

$$\Delta_{\text{sup}} = 0,1111 + 1,96\sqrt{\left(\frac{25 \times 101}{126^3} + \frac{38 \times 397}{435^3}\right)} = 0,1856$$

Méthode basée sur la transformation ARC

La valeur du χ^2_T est de 10,21 (tableau 6.3). Sur la base de cette valeur, les limites de confiance à 95 % de Δ sont données par :

$$\Delta_{\text{inf}} = 0,1111\left(1 - \frac{1,96}{\sqrt{10,21}}\right) = 0,0439$$

$$\Delta_{\text{sup}} = 0,1111\left(1 + \frac{1,96}{\sqrt{10,21}}\right) = 0,1792$$

Méthode du rapport de vraisemblance

Il faut résoudre par méthode itérative l'équation logarithmique $2\left[L - L(\Delta)\right] - \chi^2_{1,1-\alpha} = 0$. Les valeurs de $L(\Delta)$ et L sont données respectivement par :

$$L(\Delta) = 25\log(\alpha + \Delta) + 101\log(1 - \alpha - \Delta) + 38\log\alpha + 397\log(1 - \alpha) + K$$

et par le maximum de cette fonction atteint lorsque Δ = *DP* et α = 38/435 :

$$\begin{aligned} L &= 25\log\left(\frac{25}{126}\right) + 101\log\left(\frac{101}{126}\right) + 38\log\left(\frac{38}{435}\right) \\ &\quad + 397\log\left(\frac{397}{435}\right) + K \\ &= -191,6970 + K \end{aligned}$$

Dans la procédure d'itération sur Δ, on détermine la valeur de la fonction $L(\Delta)$ en remplaçant Δ par sa valeur d'itération et α par son estimation du maximum de vraisemblance correspondant.

La procédure GENMOD de SAS, dans le cadre de la distribution binomiale avec la fonction de lien *identity*, permet d'obtenir assez facilement ces limites de confiance.

On retrouve dans le tableau 6.6 les limites de confiance à 95 % de *DP* calculées suivant les différentes méthodes présentées.

TABLEAU 6.6

Méthode	Intervalle de confiance à 95 %	Programme
Exacte	[0,0418 ; 0,1837]	PR6.17
Simple	[0,0366 ; 0,1856]	PR6.18
Transformation ARC	[0,0429 ; 0,1792]	PR6.19
RV	[0,0416 ; 0,1904]	PR6.20

◆

6.4 MESURE DU *SMR* POUR LES PROPORTIONS

Dans un groupe G de taille n, on observe x décès. On veut alors comparer la proportion p_1 ($= x/n_1$) observée sur ce groupe à la proportion π_0 d'une population P de référence. Comparer p_1 à π_0 revient à comparer le nombre x de cas observés au nombre $X_0 = n_1\pi_0$ de cas attendus, calculé sous l'hypothèse que le groupe G est un échantillon aléatoire de la population de référence. On aura reconnu le *SMR* dans le rapport $\frac{x}{X_0}$ ou $\frac{p_1}{\pi_0}$.

La variable X (nombre de cas observés) est une variable binomiale de paramètres π_1 et n_1. Le paramètre π_1 décrit la vraie proportion pour la population d'où provient le groupe sous observation. (Cette population n'est pas forcément la population de référence.) La valeur attendue et la variance de X correspondent respectivement à $E(X) = n_1\pi_1$ et $V(X) = \grave{A}n_1\pi_1(1 - \pi_1)$. Si on désigne par ϕ la vraie valeur du *SMR*, alors $\phi = \frac{\pi_1}{\pi_0}$.

L'hypothèse que le groupe G est un échantillon aléatoire de la population de référence se traduit par $\phi = 1$. Sous cette hypothèse (nulle), le paramètre π_1 correspond précisément à $\frac{X_0}{n_1}$; la valeur attendue et la variance de X sont respectivement données par $E(X) = n_1\pi_1 = n_1\pi_0 = X_0$ et $V(X) = \frac{X_0(n_1 - X_0)}{n_1}$. Sous une hypothèse quelconque ϕ du *SMR*, le para-

mètre π_1 correspond à $\frac{\phi X_0}{n_1}$; la valeur attendue et la variance de X sont respectivement données par $E(X) = \phi X_0$ et $V(X) = \frac{\phi X_0 (n_1 - \phi X_0)}{n_1}$. De là, on peut facilement établir que, sous une hypothèse ϕ pour le SMR :

- $E(SMR) = \phi$
- $V(SMR) = \frac{n_1 \phi (1 - \phi \pi_0)}{\pi_0}$. L'estimateur du maximum de vraisemblance de cette variance est donnée par $\frac{x(n_1 - x)}{n_1 \pi_0^2}$.

6.4.1 TESTS STATISTIQUES POUR LE *SMR*

Comme pour le *SMR* des taux, nous suggérons ici différents tests statistiques : un test exact, trois tests en approximation normale et le test du rapport de vraisemblance. Le test exact et les tests en approximation normale sont les mêmes que nous avons déjà définis à la section 4.4.1 du chapitre 4.

TEST EXACT

Le test exact est directement construit sur la variable binomiale X. Sous l'hypothèse nulle ($\phi = 1$), les paramètres de la variable binomiale X sont $\frac{X_0}{n_1}$ et n_1. Le test unilatéral à droite, suivant la convention mi-*p*, est alors de la forme :

$$p = 1/2 C_{n_1}^{x} \left(\frac{X_0}{n_1} \right)^{x} \left(\frac{n_1 - X_0}{n_1} \right)^{n_1 - x} + \sum_{i=x+1}^{n_1} C_{n_1}^{i} \left(\frac{X_0}{n_1} \right)^{i} \left(\frac{n_1 - X_0}{n_1} \right)^{n_1 - i}$$

où x est la valeur observée de X. On peut définir de façon analogue le test unilatéral à gauche.

TEST EN APPROXIMATION NORMALE

TEST SIMPLE

À l'aide de l'approximation normale appliquée à la variable binomiale X, il est facile de construire un test sous l'hypothèse nulle $\phi = 1$.

$$z = \frac{x - X_0}{\sqrt{V(X)}}$$
$$= \frac{x - X_0}{\sqrt{\frac{X_0(n_1 - X_0)}{n_1}}}$$

ou encore

$$\chi_1^2 = \frac{(x - X_0)^2}{X_0}\left[\frac{n_1}{n_1 - X_0}\right]$$

Test basé sur log(SMR)

Le test peut être conduit sur le logarithme du *SMR*. Il prend alors la forme suivante :

$$\chi_1^2 = \frac{(\log SMR)^2}{V[\log(SMR)]}$$

où $V[\log(SMR)] = \frac{(n_1 - X)}{n_1 X}$.

Le test se présente comme : $\chi_1^2 = \frac{n_1 x[\log SMR]^2}{(n_1 - x)}$.

Test basé sur la transformation ARC

Rappelons que le *SMR* se définit comme : $SMR = \frac{p_1}{\pi_0}$ où $p_1 = \frac{x}{n_1}$ et $\pi_0 = \frac{X_0}{n_1}$.

Le test est alors construit sur la comparaison de la valeur observée $\arcsin\sqrt{p_1}$ à sa valeur attendue sous l'hypothèse d'un *SMR* = 1, $\arcsin\sqrt{\pi_0}$. La variance de $\arcsin\sqrt{p_1}$ est égale à $\frac{1}{4n_1}$.

On construit alors la statistique $Z^2 = \dfrac{\left(\arcsin\sqrt{p_1} - \arcsin\sqrt{\pi_0}\right)^2}{\dfrac{1}{4n_1}}$ qui obéit en bonne approximation à une loi du khi-carré avec un degré de liberté.

Le test prend alors la forme :

$$\chi_1^2 = 4n_1\left(\arcsin\sqrt{\frac{x}{n_1}} - \arcsin\sqrt{\frac{X_0}{n_1}}\right)^2$$

Test du rapport de vraisemblance

Revenons aux conditions définies au début de cette section. La fonction de vraisemblance du *SMR* peut être décrite comme :

$$FV(\phi) = C_{n_1}^{x}\left(\frac{\phi X_0}{n_1}\right)^{x}\left(\frac{n_1 - \phi X_0}{n_1}\right)^{n_1 - x}$$

Sous l'hypothèse nulle d'un *SMR* égal à 1, le paramètre π_1 de la loi binomiale se réduit simplement à X_0/n_1 et la fonction de vraisemblance FV_0 peut s'écrire comme :

$$FV_0 = C_{n_1}^{x}\left(\frac{X_0}{n_1}\right)^{x}\left(\frac{n_1 - X_0}{n_1}\right)^{n_1 - x}$$

Par ailleurs, la contre-hypothèse correspond à celle entre toutes pour laquelle la fonction $FV(\phi)$ atteint son maximum. Dans ce cas, π correspond au *SMR* mesuré sur les données et $\phi X_0 = x$:

$$FV_1 = C_{n_1}^{x}\left(\frac{x}{n_1}\right)^{x}\left(\frac{n_1 - x}{n_1}\right)^{n_1 - x}$$

Le test du rapport de vraisemblance se présente alors dans une forme déjà rencontrée :

$$\begin{aligned}\chi_1^2 &= -2\log\left(\frac{FV_0}{FV_1}\right)\\ &= 2\left[O\log\left(\frac{O}{A}\right) + (n_1 - O)\log\left(\frac{n_1 - O}{n_1 - A}\right)\right]\end{aligned}$$

où O représente la valeur observée ($O = x$) et A la valeur attendue sous l'hypothèse nulle ($A = X_0$).

Exemple 6.5

Dans un groupe de 2000 bébés nés de mères ayant consommé une certaine hormone pendant la grossesse, on a observé 6 bébés avec une certaine malformation congénitale. Dans la population générale, on estime à 1 pour 1000 naissances la fréquence d'une telle malformation. Le nombre de 6 cas recensés dans le groupe des 2000 bébés correspond-il à un excès réel de cas ou est-ce une donnée compatible avec l'hypothèse d'un *SMR* égal à 1 ?

Pour solutionner ce problème, établissons d'abord que la valeur attendue X_0 de bébés ayant la malformation congénitale est de 2 : $X_0 = n_1\pi_0 = 2000 \times 0{,}001 = 2$. La valeur observée est de 6. Le *SMR* mesuré sur ces données est donc de 3 (= 6/2).

Test exact

Le test unilatéral à droite (dans la convention mi-p) se présente comme

$$p = 1/2C_{2000}^{i}\left(\frac{2}{2000}\right)^{6}\left(\frac{1998}{2000}\right)^{2000-6} + \sum_{i=7}^{2000} C_{2000}^{i}\left(\frac{2}{2000}\right)^{i}\left(\frac{1998}{2000}\right)^{2000-i}$$
$$= 0{,}0105$$

Avec une valeur-p aussi faible que 0,0105, l'hypothèse d'un *SMR* égal à 1 est peu vraisemblable. On rejette cette hypothèse nulle pour retenir celle d'une association positive entre la prise de l'hormone par la mère et la présence de la malformation congénitale chez le bébé.

Test en approximation normale

Test simple

Appliqué aux données, ce test normal simple donne :

$$\chi_1^2 = \frac{(x - X_0)^2}{X_0}\left[\frac{n_1}{n_1 - X_0}\right]$$
$$= \frac{(6-2)^2}{2}\left[\frac{2000}{2000-2}\right]$$
$$= 8{,}008$$

La valeur-p bilatérale correspondante est de 0,00466.

Test basé sur le logarithme du SMR

Appliqué aux données, ce test donne :

$$\begin{aligned}\chi_1^2 &= \frac{n_1 x\left[\log SMR\right]^2}{n_1 - x} \\ &= \frac{2000 \times 6 \times \left[\log 3\right]^2}{2000 - 6} \\ &= 7,263\end{aligned}$$

La valeur-*p* bilatérale correspondante est de 0,007.

Test basé sur la transformation ARC

Appliqué aux données, ce test en approximation normale donne :

$$\begin{aligned}\chi_1^2 &= 4 \times 2000\left(\arcsin\sqrt{\frac{6}{2000}} - \arcsin\sqrt{\frac{2}{2000}}\right)^2 \\ &= 4,2954\end{aligned}$$

La valeur-*p* bilatérale correspondante est de 0,038216.

Test du rapport de vraisemblance

La valeur du test est calculée comme :

$$2\left[6\log\left(\frac{6}{2}\right) + (2000-6)\log\left(\frac{2000-6}{2000-2}\right)\right] = 5,19136$$

La valeur-*p* bilatérale est égale à 0,022699.

Dans le tableau 6.7, nous résumons les résultats des différents tests appliqués sur les données de l'exemple.

Tableau 6.7

Méthode	Khi-carré (1 ddl)	Valeur *p*		Programme
		Unilatérale	Bilatérale	
Exact	–	0,010510	0,010510	PR6.21
Simple	8,008	0,002328	0,004657	PR6.22
Basée sur log(*SMR*)	7,26348	0,003518	0,007037	PR6.23
Transformation ARC	4,2954	0,019108	0,038216	PR6.24
RV	5,19136	0,011350	0,022699	PR6.25

Tous les tests concluent à un excès significatif du nombre de cas observés sur le nombre de cas attendus. On remarque que les tests exacts et du rapport de vraisemblance ont sensiblement les mêmes valeurs-*p* unilatérales. Le test basé sur la transformation ARC est plus conservateur et les tests simple et basé sur log(*SMR*) plus permissifs.

◆

6.4.2 INTERVALLE DE CONFIANCE DU *SMR*

L'intervalle de confiance du *SMR* (ϕ) se calcule de façon analogue à celui d'une proportion (section 4.2 du chapitre 4).

On suppose qu'on a observé x cas pour n_1 personnes dans un groupe alors que la proportion de la population générale est de π_0. L'estimation de ϕ sur les données est x/X_0. On veut déterminer l'intervalle de confiance de ϕ pour un niveau $100(1-\alpha)\,\%$.

MÉTHODE EXACTE

Dans l'approche exacte, les limites de confiance inférieure ϕ_{inf} et supérieure ϕ_{sup} de ϕ sont décrites respectivement comme suit :

ϕ_{inf} est la valeur de ϕ qui solutionne l'équation

$$\sum_{i=x}^{n_1} C_{n_1}^{i} \left(\frac{\phi X_0}{n_1}\right)^{i} \left(\frac{n_1 - \phi X_0}{n_1}\right)^{n_1 - i} = \alpha/2$$

et

ϕ_{sup} est la valeur de ϕ qui solutionne l'équation

$$\sum_{i=0}^{x} C_{n_1}^{i} \left(\frac{\phi X_0}{n_1}\right)^{i} \left(\frac{n_1 - \phi X_0}{n_1}\right)^{n_1 - i} = \alpha/2$$

Par itération sur ϕ, en appliquant la convention mi-p, on obtient assez facilement les limites inférieure et supérieure de l'intervalle.

MÉTHODE EN APPROXIMATION NORMALE

Méthode simple

On utilise directement l'expression $SMR \pm z_{\alpha/2}\sqrt{V(SMR)}$ où la variance du *SMR* est estimée par $V(SMR) = \dfrac{x(n_1 - x)}{n_1 X_0^2} = SMR \times \dfrac{(n_1 - x)}{n_1 X_0}$. Alors, en découle l'intervalle de confiance. On remarque que la variance du *SMR* est instable puisque qu'elle dépend de la valeur du *SMR* lui-même. Cette instabilité de la variance rend douteuse la méthode simple d'estimation d'intervalle.

Méthode basée sur log(SMR)

L'intervalle de confiance est d'abord calculé pour log(*SMR*). Puis par transformation inverse des limites de confiance de cet intervalle, on obtient celles de l'intervalle du *SMR*.

À cette fin, établissons d'abord que :

$$\begin{aligned} V\left[\log SMR\right] &= \left(\frac{1}{SMR}\right)^2 V\left(SMR\right) \\ &= \left(\frac{1}{SMR}\right)^2 SMR\left(\frac{n_1 - x}{n_1 X_0}\right) \\ &= \left(\frac{n_1 - x}{n_1 x}\right) \end{aligned}$$

On a alors : $\log(SMR) \pm z_{\alpha/2}\sqrt{\frac{n_1 - x}{n_1 x}}$.

Par transformation inverse, on obtient : $SMR \times \exp\left(\pm z_{\alpha/2}\sqrt{\frac{n_1 - x}{n_1 x}}\right)$.

Ainsi, les limites de confiance du *SMR* peuvent être décrites comme :

$$\phi_{\text{inf}} = SMR \times \exp\left(-z_{\alpha/2}\sqrt{\frac{n_1 - x}{n_1 x}}\right)$$

$$\phi_{\text{sup}} = SMR \times \exp\left(+z_{\alpha/2}\sqrt{\frac{n_1 - x}{n_1 x}}\right)$$

Transformation ARC

Pour stabiliser la variance du *SMR*, on utilise la transformation ARC sur la proportion $p_1 = X/n_1$. L'intervalle de confiance est calculé sur $\arcsin\sqrt{p_1}$, dont la variance est estimée à $\frac{1}{4n_1}$. Par transformation inverse, on obtient les limites de confiance de p_1 à partir de celles de $\arcsin\sqrt{p_1}$. En divisant ces limites de p_1 par la valeur π_0 de la population de référence, on obtient les limites de confiance du *SMR*.

Formellement, les limites de l'intervalle de confiance de *SMR* se décrivent comme :

$$\phi_{inf} = \frac{1}{\pi_0}\left\{\sin\left[\arcsin\sqrt{p_1} - \frac{z_{\alpha/2}}{2\sqrt{n_1}}\right]\right\}^2$$

$$\phi_{sup} = \frac{1}{\pi_0}\left\{\sin\left[\arcsin\sqrt{p_1} + \frac{z_{\alpha/2}}{2\sqrt{n_1}}\right]\right\}^2$$

MÉTHODE QUADRATIQUE

L'intervalle de confiance peut aussi être calculé sans condition sur la variance.

On considère la variable binomiale X de paramètres $\pi = \frac{\phi X_0}{n_1}$ et n_1, telle qu'elle a été décrite au début de la section 6.4. On a : $E(X) = \phi X_0$ et $V(X) = \frac{\phi X_0 (n_1 - \phi X_0)}{n_1}$. On considère alors la statistique $Z^2(\phi) = \frac{n_1 (X - \phi X_0)^2}{\phi X_0 (n_1 - \phi X_0)}$ qui, pour une valeur fixe de ϕ, obéit en bonne approximation à une loi du khi-carré à un degré de liberté. Si $\chi^2_{1,1-\alpha}$ désigne la valeur du khi-carré correspondant au seuil de signification α, alors toute valeur de ϕ telle que $Z^2(\phi) \leq \chi^2_{1,1-\alpha}$ est une valeur compatible avec les données au seuil de signification α. Alors pour $X = x$, il s'agit de solutionner pour ϕ l'équation quadratique de la forme :

$\frac{n_1 (x - \phi X_0)^2}{\phi X_0 (n_1 - \phi X_0)} = \chi^2_{1,1-\alpha}$ où x, n_1, X_0 et $\chi^2_{1,1-\alpha}$ sont des valeurs connues. Cela revient alors à solutionner pour ϕ l'équation quadratique suivante :

$$f(\phi) = \left[X_0^2 (n_1 + \chi^2_{1,1-\alpha})\right] \phi^2 - n_1 X_0 (2x + \chi^2_{1,1-\alpha})\, \phi + n_1 x^2 = 0$$

On a, $\chi^2_{1,(1-0,05)} = 3,84$.

Méthode du rapport de vraisemblance

On se rappelle que la fonction de vraisemblance pour le *SMR* (= ϕ) est de la forme

$$FV(\phi) = C_{n_1}^{x}\left(\frac{\phi X_0}{n_1}\right)^{x}\left(\frac{n_1 - \phi X_0}{n_1}\right)^{n_1 - x}$$

Il suffit alors de résoudre pour ϕ l'équation logarithmique : $2\left[L - L(\phi)\right] - \chi^2_{1,1-\alpha} = 0.$

Ainsi, puisque $L(\phi) = x\log\left(\phi X_0\right) + (n_1 - x)\log(n_1 - \phi X_0) + K$ et que $L = x\log\left(x\right) + (n_1 - x)\log(n_1 - x) + K,$ alors l'équation logarithmique devient $\left[x\log x \ + (n_1 - x)\ \log(n_1 - x)\right] -$

$$\left[x\log\left(\phi X_0\right) + (n_1 - x)\log(n_1 - \phi X_0)\right] - 1/2\chi^2_{1,1-\alpha} = 0$$

Les deux valeurs de ϕ qui satisfont cette dernière équation correspondent aux limites de l'intervalle de confiance à $100(1 - \alpha)\,\%$ pour le *SMR*. La résolution de l'équation peut se faire assez facilement par itération. On peut aussi utiliser la procédure GENMOD de SAS.

De façon équivalente, on peut considérer la variable binomiale *X* dont la fonction de vraisemblance se présente comme :

$FV(\pi) = C_{n_1}^{x}\pi^{x}\left(1 - \pi\right)^{n_1 - x}$. L'estimateur du maximum de vraisemblance de π étant x/n_1, ses limites de confiance, π_{inf} et π_{sup}, peuvent facilement être calculées par la méthode du maximum de vraisemblance.

Celles du SMR en découlent directement :

$$\phi_{inf} = \frac{n_1 \times \pi_{inf}}{X_0} = \frac{X_{inf}}{X_0}$$

$$\phi_{sup} = \frac{n_1 \times \pi_{sup}}{X_0} = \frac{X_{sup}}{X_0}$$

Exemple 6.6

Considérons les données de l'exemple 6.5. On a $x = 6$, $n_1 = 2000$ et $\pi_0 = 0{,}001$; *SMR* = 3. On veut calculer les limites de confiance à 95 % du *SMR* mesuré sur ces données.

Méthode exacte

Le calcul des limites de confiance du *SMR* par la méthode exacte passe par la résolution des équations numériques qui les définissent :

ϕ_{inf} est la valeur de ϕ qui solutionne l'équation

$$\sum_{i=6}^{2000} C_{2000}^{i} \left(\frac{2\phi}{2000}\right)^{i} \left(\frac{2000-2\phi}{2000}\right)^{2000-i} = 0,025$$

ϕ_{sup} est la valeur de ϕ qui solutionne l'équation

$$\sum_{i=0}^{6} C_{2000}^{i} \left(\frac{2\phi}{2000}\right)^{i} \left(\frac{2000-2\phi}{2000}\right)^{2000-i} = 0,025$$

Les limites de confiance obtenues sont :

$$\phi_{inf} = 1,2166$$
$$\phi_{sup} = 6,2286$$

Méthodes en approximation normale

Méthode simple

Les limites de confiance obtenues par la méthode simple sont :

$$\phi_{inf} = 3 - 1,96\sqrt{\frac{6(2000-6)}{2000\times 2^2}}$$
$$= 0,6031$$
$$\phi_{sup} = 3 + 1,96\sqrt{\frac{6(2000-6)}{2000\times 2^2}}$$
$$= 5,3969$$

Méthode basée sur log(SMR)

Les limites de confiance obtenues par la méthode basée sur log(*SMR*) sont :

$$\phi_{inf} = 3\times \exp\left(-1,96\sqrt{\frac{2000-6}{2000\times 6}}\right)$$
$$= 1,3494$$
$$\phi_{sup} = 3\times \exp\left(+1,96\sqrt{\frac{2000-6}{2000\times 6}}\right)$$
$$= 6,6696$$

Méthode de la transformation ARC

Les limites de confiance obtenues par la méthode de la transformation ARC sont :

$$\phi_{inf} = \frac{1}{0,001}\left\{\sin\left[\arcsin\sqrt{0,003} - \frac{1,96}{2\sqrt{2000}}\right]\right\}^2$$
$$= 1,0811$$

$$\phi_{sup} = \frac{1}{0,001}\left\{\sin\left[\arcsin\sqrt{0,003} + \frac{1,96}{2\sqrt{2000}}\right]\right\}^2$$
$$= 5,8734$$

Méthode quadratique

L'équation quadratique en ϕ qu'il faut résoudre pour déterminer les limites de confiance du *SMR* est

$$f(\phi) = \left[2^2(2000 + 3,84)\right]\phi^2 - 2000 \times 2(2 \times 6 + 3,84)\,\phi + 2000 \times 6^2$$
$$= 8015,36\phi^2 - 63360\phi + 72000 = 0$$

Les racines de cette équation correspondent aux limites recherchées :

$$\phi_{inf} = 1,3756$$
$$\phi_{sup} = 6,5299$$

Méthode du rapport de vraisemblance

L'équation logarithmique qu'il faut résoudre pour déterminer les limites de confiance du *SMR* est :

$$[6\log 6 + (2000 - 6)\log(2000 - 6)] -$$
$$[6\log(2\phi) + (2000 - 6)\log(2000 - 2\phi)] - 1/2 \times 3,84 = 0$$

Les racines de cette équation correspondent aux limites recherchées :

$$\phi_{inf} = 1,1935$$
$$\phi_{sup} = 6,0697$$

On retrouve dans le tableau 6.8 les limites de confiance du *SMR* calculées suivant les différentes méthodes présentées.

Tableau 6.8

Méthode	Intervalle de confiance	Programme
Exacte	[1,2166 ; 6,2286]	PR6.26
Normale simple	[0,6032 ; 5,3969]	PR6.27
Basée sur log(*SMR*)	[1,3494 ; 6,6696]	PR6.28
Transformation ARC	[1,0811 ; 5,8733]	PR6.29
Méthode quadratique	[1,3756 ; 6,5299]	PR6.30
RV	[1,1935 ; 6,0697]	PR6-31

La méthode à éviter est celle de l'approximation normale simple. Son résultat ici ne concorde ni avec les tests d'hypothèse ni avec les autres méthodes de calcul d'intervalle de confiance. Les trois autres méthodes en approximation normale montrent aussi des différences entre elles et avec les méthodes exactes et du rapport de vraisemblance. Ces deux dernières méthodes sont les plus concordantes.

◆

CHAPITRE 7

LES MESURES FRACTIONNAIRES ET LEURS INTERVALLES DE CONFIANCE

Sous le vocable de « mesures fractionnaires », nous regroupons les mesures qui s'expriment comme des fractions des mesures de base, principalement les fractions de taux et de proportions. Dans cette famille de mesures, on retrouve la fraction attribuable ou étiologique, la fraction prévenue ou évitable. Nous présentons d'abord la fraction attribuable ou étiologique, puis la fraction prévenue ou évitable. En évitant toute discussion sur la signification épidémiologique de ces mesures, nous décrivons les méthodes de calcul de leurs intervalles de confiance.

Le calcul de l'intervalle de confiance des mesures fractionnaires n'est pas simple en général. Ici, nous proposons une approche basée sur la condition des marges fixes qui permet d'alléger les calculs dans le contexte des études de cohortes. Dans cette approche, la mesure fractionnaire est définie comme la

fonction d'une variable binomiale ou hypergéométrique suivant le type de données. Si l'étude de cohorte est conduite dans une population totale ou sur un échantillon aléatoire de celle-là, alors la mesure est directement valide. Si l'étude de cohorte est conduite sur deux groupes d'exposés et de non-exposés dont les fractions d'échantillonnage sont connues, il suffit d'appliquer à la mesure une correction tenant compte de ces fractions d'échantillonnage. Cette approche *des marges fixes* n'a pas été définie pour les études cas-témoins, où il est difficile d'assurer une certaine validité à la mesure. Dans ce dernier cas, nous renvoyons le lecteur aux méthodes traditionnelles de calcul, méthodes qui ne sont pas présentées ici.

7.1 FRACTION ATTRIBUABLE

Soit X un facteur de risque pour la maladie Y. On dénote par $X = 1$ et par $X = 0$ respectivement l'exposition et la non-exposition à ce facteur. On considère la mesure de fréquence R (un taux ou une proportion) évaluée chez les exposés (R_1), chez les non-exposés (R_0) ou dans la population totale (R_t). On désigne par RR le rapport des mesures R_1 et R_0 (tableau 7.1).

TABLEAU 7.1

	$X = 1$	$X = 0$	Total
$Y = 1$	a_1	a_0	m_1
$Y = 0$*	b_1	b_0	m_0
Total	n_1	n_0	n

* La ligne $Y = 0$ n'est pas définie pour les données de personnes-temps.

On rappelle que le risque total R_t est la somme pondérée des risques R_1 et R_0 comme : $R_1 = P_1R_1 + (1 - P_1)R_0$, où P_1 désigne la proportion d'exposés dans la population.

La fraction attribuable peut être calculée parmi les cas exposés. Désignée par FA_1, elle est dite fraction attribuable chez les exposés . Elle peut aussi être calculée parmi tous les cas recensés dans la population. Désignée par FA_t, elle est dite fraction attribuable totale ou de population.

En utilisant les notations décrites ci-dessus, ces mesures se définissent formellement comme suit :

$$FA_1 = \frac{R_1 - R_0}{R_1} = \frac{RR - 1}{RR} \qquad (7.1)$$

$$\begin{aligned} FA_t &= \frac{R_t - R_0}{R_t} \\ &= p_c \times \frac{RR-1}{RR} \\ &= p_c \times FA_1 \end{aligned} \tag{7.2}$$

où $p_c \approx \frac{a_1}{m_1}$ correspond à la proportion de cas exposés parmi la totalité des cas.

Ces mesures n'ont de sens que si $RR \geq 1$, ce qu'il nous faut dès lors supposer.

7.1.1 FRACTION ATTRIBUABLE CHEZ LES EXPOSÉS : FA_1

On peut considérer FA_1 comme une transformation monotone croissante du *RR*. En conséquence, le calcul de l'intervalle de confiance de cette fraction attribuable découle directement de celui du *RR*. Si RR_{inf} et RR_{sup} désignent les limites de confiance du *RR*, alors celles de FA_1 sont données par :

$$FA_{1\,\text{inf}} = \frac{RR_{\text{inf}} - 1}{RR_{\text{inf}}}$$

$$FA_{1\,\text{sup}} = \frac{RR_{\text{sup}} - 1}{RR_{\text{sup}}}$$

Dans le contexte des études de cohortes sur une spopulation totale ou sur un échantillon aléatoire de celle-là, on peut aussi présenter la fraction attribuable chez les exposés en fonction du nombre A_1 de cas exposés. En effet, si on suppose que m_1 est une valeur fixe, alors

$$FA_1 = \frac{n}{n_0} \times \left[\frac{A_1 - E_0(A_1)}{A_1} \right] \tag{7.3}$$

où $E_0(A_1) = \frac{n_1 m_1}{n}$ correspond à la valeur attendue de A_1 sous l'hypothèse nulle.

Les limites de confiance de FP_1 peuvent être calculées comme :

$$FA_{1\,\text{inf}} = \frac{n}{n_0} \times \left[\frac{A_{1\,\text{inf}} - E_0(A_1)}{A_{1\,\text{inf}}} \right]$$

$$FA_{1\,\text{sup}} = \frac{n}{n_0} \times \left[\frac{A_{1\,\text{sup}} - E_0(A_1)}{A_{1\,\text{sup}}} \right]$$

où $A_{1\,\text{inf}} = E(A_1 \mid RR_{\text{inf}})$ et $A_{1\,\text{sup}} = E(A_1 \mid RR_{\text{sup}})$.

Si l'étude n'est basée ni sur la population totale ni sur un échantillon aléatoire de celle-là, alors il est possible de définir une formule corrigée si les fractions d'échantillonnage f_1 pour les exposés et f_0 pour les non-exposés sont connues : $f_1 = \frac{n_1}{N_1}$ et $f_0 = \frac{n_0}{N_0}$, où N_1 et N_0 représentent les effectifs totaux respectivement des sujets exposés et non exposés dans la population. Dans ce cas, on peut montrer que

$$FA_1 = \frac{nf_0}{n_0 F_t} \times \left[\frac{A_1 - E_0(A_1) F_t / F_1}{A_1} \right] \tag{7.4}$$

où $F_1 = \dfrac{m_1}{\frac{a_1}{f_1} + \frac{a_0}{f_0}}$ et $F_t = \dfrac{n}{\frac{n_1}{f_1} + \frac{n_0}{f_0}}$ sont considérées comme des valeurs fixes.

Remarquons que F_1 et F_t sont des moyennes harmoniques de f_1 et f_0.

Ces dernières formulations peuvent être utiles pour le calcul de l'intervalle de confiance exact. Les valeurs RR_{inf} et RR_{sup} sont alors assimilées aux limites inférieure et supérieure du RR, successivement associées au paramètre de la distribution de la variable A_1, variable binomiale ou hypergéométrique suivant le type de données.

Exemple 7.1

Fraction attribuable chez les exposés pour le rapport de taux

Considérons les données du tableau 5.2 du chapitre 5. Nous les reproduisons ici, au tableau 7.2.

TABLEAU 7.2

	$X = 1$	$X = 0$	Total
Y (décès)	37	51	88
Personnes-années	712	1535	2247

Le rapport de taux est donné par : $RT = \frac{37}{712} \Big/ \frac{51}{1535} = 1,5641$ et la fraction attribuable FA_1 est estimée par $FA_1 = \frac{1,5641 - 1}{1,5641} = 0,3610$ (voir la relation 7.1).

Ses limites de confiance à 95 % peuvent alors facilement être déduites à partir de celles du *RT*. Au tableau 7.3, nous reprenons les intervalles de confiance du *RT* déjà décrits au tableau 5.4 suivant différentes méthodes de calcul et présentons les intervalles de confiance de FA_1 correspondants.

TABLEAU 7.3

Méthode de calcul	Intervalle de confiance à 95 %			Programme
	Du *RT*	De A_1	De FA_1	
Exact	[1,0178 ; 2,3864]	[28,22 ; 46,23]	[0,0175 ; 0,5810]	PR7.1
Normale simple	[1,0243 ; 2,3883]	[28,34 ; 46,25]	[0,0237 ; 0,5813]	PR7.2
Résultat d'un test	[1,0279 ; 2,3799]	[28,41 ; 46,17]	[0,0271 ; 0,5798]	PR7.3
RV	[1,0178 ; 2,3807]	[28,22 ; 46,18]	[0,0175 ; 0,5800]	PR7.4

Par exemple, la limite inférieure de l'intervalle de confiance à 95 % du FA_1 par la méthode exacte (tableau 7.3) est simplement obtenue à partir de la limite inférieure de l'intervalle de confiance du *RT* par la méthode exacte :

$$FA_{1\,\text{inf}} = \frac{1,0178 - 1}{1,0178} = 0,0175$$

Les limites de confiance peuvent aussi être calculées à partir de la relation (7.3).

Pour obtenir les valeurs de $A_{1\text{inf}}$ et $A_{1\text{sup}}$ correspondant aux valeurs RT_{sup} et RT_{inf}, on rappelle la relation $E(A_1 \mid RT) = \pi \times m_1$ où $\pi = \frac{RT \times n_1}{RT \times n_1 + n_0}$. Ainsi, $A_{1\,\text{inf}} = \frac{RT_{\text{inf}} n_1}{RT_{\text{inf}} n_1 + n_0} \times m_1$. De façon analogue, on obtient la valeur de $A_{1\,\text{inf}}$ à partir de RT_{sup}. Pour l'approche exacte, on obtient les valeurs suivantes : $A_{1\,\text{inf}} = 28,2215$ et $A_{1\,\text{sup}} = 46,2328$. De plus, $E_0(A_1) = \frac{712 \times 88}{2247} = 27,8843.$

Les limites de confiance sont alors données par :

$$FA_{1\,\text{inf}} = \frac{2247}{1535} \times \left[\frac{28,2215 - 27,8843}{28,2215} \right] = 0,0175$$

$$FA_{1\,\text{sup}} = \frac{2247}{1535} \times \left[\frac{46,2328 - 27,8843}{46,2328} \right] = 0,5810$$

◆

Exemple 7.2

Fraction attribuable chez les exposés pour le rapport de proportions

Considérons les données du tableau 6.2, du chapitre 6, que nous rappelons ici, au tableau 7.4.

Tableau 7.4

	Exposition à X		Total
	$X = 1$	$X = 0$	
$Y = 1$	25	38	63
$Y = 0$	101	397	498
Total	126	435	561

Le rapport de proportions *RP* est donné par $RP = \frac{25}{126} \Big/ \frac{38}{435} = 2,2713$. La fraction attribuable FA_1 est estimée par $FA_1 = \frac{2,2713 - 1}{2,2713} = 0,5597$ (voir la relation 7.1). Ses limites de confiance à 95 % peuvent alors facilement être déduites à partir de celles du *RP* obtenues par les différentes méthodes de calcul. Au tableau 7.5 nous reprenons les intervalles de confiance du *RP* déjà décrits au tableau 6.5 du chapitre 6, suivant différentes méthodes de calcul, et présentons les intervalles de confiance du *RP* correspondants.

Tableau 7.5

Méthode de calcul	Intervalle de confiance à 95 %			Programme
	Du *RT*	De A_1	De FA_1	
Exact	[1,4059 ; 3,5861]	[18,23 ; 32,10]	[0,2887 ; 0,7211]	PR7.5
Normale simple	[1,4279 ; 3,6129]	[18,43 ; 32,22]	[0,2997 ; 0,7232]	PR7.6
Résultat d'un test	[1,4297 ; 3,6083]	[18,45 ; 32,20]	[0,3006 ; 0,7229]	PR7.7
RV	[1,4092 ; 3,5902]	[18,26 ; 32,12]	[0,2904 ; 0,7215]	PR7.8

◆

7.1.2 FRACTION ATTRIBUABLE TOTALE OU DE POPULATION : FA_t

Considérons une étude de cohorte conduite dans la population totale, ou sur un échantillon de celle-là. Les données peuvent être disposées suivant le schéma du tableau 7.1. On suppose alors que m_1 est une valeur fixe. Dans ce cas, on peut montrer que

$$FA_t = \frac{A_1 - E_0(A_1)}{m_1 - E_0(A_1)} \tag{7.5}$$

Remarquons que FA_t est une fonction monotone croissante de A_1. La variable A_1 est binomiale pour des données de personnes-temps et hypergéométrique pour des données de personnes. Les limites de l'intervalle de confiance de la fraction attribuable de la population peuvent être alors déduites comme :

$$FA_{t\,\text{inf}} = \frac{A_{1\text{inf}} - E_0(A_1)}{m_1 - E_0(A_1)}$$

$$FA_{t\,\text{sup}} = \frac{A_{1\text{sup}} - E_0(A_1)}{m_1 - E_0(A_1)}$$

où $A_{1\text{inf}} = E(A_1 \mid RR_{\text{inf}})$ et $A_{1\text{sup}} = E(A_1 \mid RR_{\text{sup}})$.

Il est parfois plus commode d'exprimer ces limites de confiance en fonction de celles du *RR*. À cette fin, on peut utiliser la relation équivalente :

$$FA_t = \frac{RR - 1}{RR + \dfrac{n_0}{n_1}} \tag{7.6}$$

L'intervalle de confiance de la fraction attribuable totale se déduit alors de celui du *RR* comme :

$$FA_{t\,\text{inf}} = \frac{RR_{\text{inf}} - 1}{RR_{\text{inf}} + \dfrac{n_0}{n_1}}$$

$$FA_{t\,\text{sup}} = \frac{RR_{\text{sup}} - 1}{RR_{\text{sup}} + \dfrac{n_0}{n_1}}$$

Si l'étude n'est basée ni sur la population totale ni sur un échantillon aléatoire de celle-là, alors il est possible de définir une formule corrigée si les fractions d'échantillonnage f_1 pour les exposés et f_0 pour les non-exposés sont connues. Dans ce cas, on peut montrer que

$$FA_t = \frac{A_1F_1 - E_0(A_1)F_t}{m_1f_1 - E_0(A_1)F_t} \tag{7.7}$$

où les facteurs F_1 et F_t sont définis comme précédemment.

Ces formulations peuvent aussi être utiles pour le calcul de l'intervalle de confiance exact.

Exemple 7.3

Fraction attribuable totale pour le rapport de taux

Considérons les données du tableau 7.2, pour lesquelles le rapport de taux est donné par : $RT = \frac{37}{712} \Big/ \frac{51}{1535} = 1,5641$. On suppose que ces données représentent bien les distributions du facteur et de la maladie dans la population.

La fraction attribuable FA_t, définie par $FA_t = p_c \times \frac{RT - 1}{RT}$, est alors estimée par $FA_t = \frac{37}{88} \times \frac{1,5641 - 1}{1,5641} = 0,1516$.

Les limites de confiance à 95 % de FA_t peuvent facilement être déduites (tableau 7.6) à partir de celles du RT ou de celles de A_1, décrites au tableau 7.3. Rappelons le lien entre la valeur attendue $E(A_1)$ et RT :

$$E(A_1) = m_1 \times \frac{RT \times n_1}{RT \times n_1 + n_0}$$

Tableau 7.6

Méthode de calcul	Intervalle de confiance à 95 %			
	Du RT	De A_1	De FA_1	Programme
Exact	[1,0178 ; 2,3864]	[28,22 ; 46,23]	[0,0056 ; 0,3052]	PR7.9
Normale simple	[1,0243 ; 2,3883]	[28,34 ; 46,25]	[0,0076 ; 0,3055]	PR7.10
Résultat d'un test	[1,0279 ; 2,3799]	[28,41 ; 46,17]	[0,0088 ; 0,3042]	PR7.11
RV	[1,0178 ; 2,3807]	[28,22 ; 46,18]	[0,0056 ; 0,3043]	PR7.12

Par exemple, pour la méthode exacte (tableau 7.6), la limite supérieure de l'intervalle de confiance à 95 % de FA_t s'obtient simplement de la façon suivante :

$$FA_{t\,\text{sup}} = \frac{2,3864 - 1}{2,3864 + \dfrac{1535}{712}} = 0,3052$$

On remarque que cette limite peut être donnée par $\dfrac{46,23 - 27,88}{88 - 27,88} = 0,3052$ sachant que $A_{1\,\text{sup}} = E(A_1 \mid RT = 2,3864) = 46,23$.

◆

EXEMPLE 7.4

FRACTION ATTRIBUABLE TOTALE POUR LE RAPPORT DE PROPORTIONS

Considérons les données de personnes décrites au tableau 7.4, pour lesquelles le rapport de proportions est de 2,2713.

On suppose que les données représentent bien les distributions du facteur et de la maladie dans la population. La fraction attribuable FA_t, définie par $FA_t = p_c \times \dfrac{RP - 1}{RP}$, est alors estimée à 0,2221. Les limites de confiance à 95 % de FA_t peuvent être déduites (tableau 7.7) à partir de celles du *RP* ou de celles de A_1 suivant la méthode de calcul utilisée.

TABLEAU 7.7

Méthode de calcul	Intervalle de confiance à 95 %			
	Du *RT*	De A_1	De FA_1	Programme
Exact	[1,4059 ; 3,5861]	[18,23 ; 32,10]	[0,0836 ; 0,3674]	PR7.13
Normale simple	[1,4279 ; 3,6129]	[18,43 ; 32,22]	[0,0877 ; 0,3698]	PR7.14
Résultat d'un test	[1,4297 ; 3,6083]	[18,45 ; 32,20]	[0,0880 ; 0,3694]	PR7.15
RV	[1,4092 ; 3,5902]	[18,26 ; 32,12]	[0,0842 ; 0,3678]	PR7.16

◆

7.2 FRACTION PRÉVENUE OU ÉVITABLE

Soit un facteur X auquel l'exposition et la non-exposition correspondent respectivement à $X = 1$ et $X = 0$. Ce facteur est un facteur protecteur de la maladie. On considère la mesure de fréquence R (un taux ou une proportion) évaluée chez les exposés (R_1), chez les non-exposés (R_0) ou dans la population totale (R_t). Dans ce contexte, $R_1 < R_0$. On désigne par *Ef* le rapport de R_0 à R_1. Les données sont disposées comme au tableau 7.1.

Alors, on peut calculer la fraction évitable (ou prévenue) parmi les cas exposés potentiels. Désignée par FP_1, elle est dite fraction évitable chez les exposés. Elle peut aussi être calculée parmi tous les cas potentiels de la population. Désignée par FP_t, elle est dite fraction évitable de population. En utilisant les notations ci-dessus, ces mesures se définissent formellement comme :

$$\begin{aligned} FP_1 &= \frac{R_0 - R_1}{R_0} \\ &= \frac{Ef - 1}{Ef} \end{aligned} \tag{7.8}$$

$$\begin{aligned} FP_t &= \frac{R_0 - R_t}{R_0} \\ &= p_1 \times \frac{Ef - 1}{Ef} \\ &= p_1 \times FP_1 \end{aligned} \tag{7.9}$$

où $p_1 = \frac{n_1}{n}$ estime la proportion d'exposés dans la population.

La fraction complémentaire de FP_1 se présente simplement comme : $(1 - FP_1) = \frac{1}{Ef} = RR$.

Ces mesures n'ont de sens que si $Ef \geq 1$, ce qu'il nous faut dès lors supposer.

7.2.1 FRACTION PRÉVENUE CHEZ LES EXPOSÉS : FP_1

Le calcul de l'intervalle de confiance de la fraction prévenue FP_1 découle de celui du *RR*.

Ainsi, on considère la fraction complémentaire ($1 - FP_1$), qui coïncide avec le *RR*, et pour laquelle les limites de confiance sont facilement calculables. Par transformation complémentaire (relation 2.8), on déduit celles de la mesure FP_1.

$$FP_{1\inf} = 1-\left(1-FP_1\right)_{\sup}$$
$$= 1-RR_{\sup}$$
$$FP_{1\sup} = 1-\left(1-FP_1\right)_{\inf}$$
$$= 1-RR_{\inf}$$

Puisque $1-RR=\dfrac{Ef-1}{Ef}$, on pourrait aussi écrire :

$$FP_{1\inf} = \frac{Ef_{\inf}-1}{Ef_{\inf}}$$
$$FP_{1\sup} = \frac{Ef_{\sup}-1}{Ef_{\sup}}$$

Dans le contexte des études de cohortes sur une population totale ou sur un échantillon aléatoire de celle-ci, on peut aussi présenter la fraction prévenue chez les exposés en fonction du nombre A_0 de cas non exposés :

$$FP_1 = \frac{n}{n_1}\times\left[\frac{A_0-E_0(A_0)}{A_0}\right] \tag{7.10}$$

où $E_0(A_0)=\dfrac{n_0m_1}{n}$.

Ainsi, les limites de confiance de FP_1 peuvent aussi être calculées comme :

$$FP_{1\inf} = \frac{n}{n_1}\times\left[\frac{A_{0\inf}-E_0(A_0)}{A_{0\inf}}\right]$$
$$FP_{1\sup} = \frac{n}{n_1}\times\left[\frac{A_{0\sup}-E_0(A_0)}{A_{0\sup}}\right]$$

où $A_{0\inf}=E(A_0\mid Ef_{\inf})=m_1-E(A_1\mid RR_{\sup})=m_1-A_{1\sup}$
et $A_{0\sup}=E(A_0\mid Ef_{\sup})=m_1-E(A_1\mid RR_{\inf})=m_1-A_{1\inf}$.

Si l'étude n'est basée ni sur la population totale ni sur un échantillon aléatoire de celle-là, alors il est possible de définir une formule corrigée si les fractions d'échantillonnage f_1 pour les exposés et f_0 pour les non-exposés sont connues. Dans ce cas,

$$FP_1 = \frac{nf_1}{n_1 F_t} \times \left[\frac{A_0 - E_0(A_0) F_t / F_1}{A_0} \right] \tag{7.11}$$

où les facteurs F_1 et F_t sont définis comme précédemment.

Ces formulations permettent le calcul exact des intervalles de confiance.

EXEMPLE 7.5

FRACTION PRÉVENUE CHEZ LES EXPOSÉS POUR LE RAPPORT DE TAUX

Considérons les données du tableau 7.8, qui décrit l'effet protecteur d'un facteur *X* contre la maladie *Y*.

TABLEAU 7.8

	$X = 1$	$X = 0$	Total
$Y = 1$	10	50	60
Total	2000	1000	3000

L'effet protecteur *Ef* du facteur *X* est mesuré par : $Ef = \frac{50}{1000} \Big/ \frac{10}{2000} = 10$.

Cette mesure indique de combien de fois le taux de la maladie est diminué chez les vaccinés par rapport à celui chez les non-vaccinés.

La fraction prévenue FP_1, définie par $FP_1 = \frac{Ef - 1}{Ef}$, est alors estimée à

$FP_1 = \frac{10 - 1}{10} = 0{,}9$. On observe aussi que $(1 - FP_1) = RR = 0{,}10$.

Les limites de confiance à 95 % de FP_1 peuvent alors facilement être déduites de celles du *RR* obtenues par les différentes méthodes de calcul.

Par exemple, si on suppose que A_1 et A_0 sont des variables de Poisson, alors en fixant le nombre de cas à $m_1 = 60$, la variable A_1 peut être considérée comme une variable binomiale. Dans ces conditions, les limites de confiance exactes pour le rapport ϕ (ou *RT*) de deux taux, calculées suivant la méthode décrite à

la section 5.3.1 du chapitre 5, correspondent à 0,048175 et 0,19155. Ainsi, par transformation complémentaire, on déduit celles de FP_1 :

$$FP_{1\text{inf}} = 1 - 0{,}1916$$
$$= 0{,}8084$$
$$FP_{1\text{sup}} = 1 - 0{,}0482$$
$$= 0{,}9518$$

Le tableau 7.9 décrit les intervalles de confiance du *RT* calculés suivant différentes méthodes et ceux de FP_1 correspondants.

TABLEAU 7.9

Méthode de calcul	Intervalle de confiance à 95 %			
	Du *RT*	De A_0	De FP_1	Programme
Exact	[0,0482 ; 0,1916]	[43,38 ; 54,73]	[0,8084 ; 0,9518]	PR7.17
Normale simple	[0,0507 ; 0,1972]	[43,03 ; 54,48]	[0,8028 ; 0,9493]	PR7.18
Résultat d'un test	[0,0577 ; 0,1732]	[44,56 ; 53,79]	[0,8268 ; 0,9423]	PR7.19
RV	[0,0478 ; 0,1886]	[43,57 ; 54,76]	[0,8114 ; 0,9522]	PR7.20

◆

EXEMPLE 7.6

FRACTION PRÉVENUE CHEZ LES EXPOSÉS POUR LE RAPPORT DE PROPORTIONS

Considérons les données suivantes, qui décrivent les résultats d'une étude sur l'efficacité d'un vaccin contre la maladie *Y* (tableau 7.10).

TABLEAU 7.10

	Vaccinés	Non-vaccinés	Total
$Y = 1$	40	80	120
$Y = 0$	160	20	180
Total	200	100	300

L'effet protecteur du vaccin est mesuré par : $Ef = \frac{80}{100} \Big/ \frac{40}{200} = 4$. Cette mesure indique de combien de fois le risque de la maladie est diminué chez les vaccinés par rapport au risque chez les non-vaccinés. La fraction prévenue correspondante est donc $FP_1 = \frac{4-1}{4} = 0{,}75$. Mais cette fraction est aussi donnée par $FP_1 = (1 - 0{,}25) = 0{,}75$, sachant que $RP = 0{,}25$.

Le tableau 7.11 décrit les intervalles de confiance du *RP*, calculés suivant différentes méthodes. Les limites de confiance à 95 % de FP_1 peuvent facilement être déduites de celles du *RP* obtenues par les différentes méthodes de calcul déjà décrites.

TABLEAU 7.11

Méthode de calcul	Intervalle de confiance à 95 %			Programme
	Du *RT*	De A_0	De FP_1	
Exact	[0,1993 ; 0,3223]	[72,96 ; 85,80]	[0,67766 ; 0,80074]	PR7.21
Normale simple	[0,1863 ; 0,3354]	[71,82 ; 87,43]	[0,66456 ; 0,81368]	PR7.22
Résultat d'un test	[0,1904 ; 0,3282]	[72,45 ; 86,91]	[0,67180 ; 0,80957]	PR7.23
RV	[0,1832 ; 0,3308]	[72,22 ; 87,82]	[0,66920 ; 0,81679]	PR7.24

◆

7.2.2 FRACTION PRÉVENUE TOTALE OU DE POPULATION : FP_t

Le calcul de l'intervalle de confiance de FP_t peut se traiter directement à partir de la relation 7.9. Les limites de confiance sont alors données par :

$$FP_{t\,\mathrm{inf}} = \frac{n_1}{n} \times \frac{Ef_{\mathrm{inf}} - 1}{Ef_{\mathrm{inf}}}$$

$$= \frac{n_1}{n} \times \left(1 - RR_{\mathrm{sup}}\right)$$

$$FP_{t\,\mathrm{sup}} = \frac{n_1}{n} \times \frac{Ef_{\mathrm{sup}} - 1}{Ef_{\mathrm{sup}}}$$

$$= \frac{n_1}{n} \times \left(1 - RR_{\mathrm{inf}}\right)$$

Remarquons que la fraction prévenue totale peut aussi s'exprimer comme une fonction de la variable A_0 :

$$FP_t = \frac{A_0 - E_0(A_0)}{A_0} \tag{7.11}$$

À partir de cette formule, l'intervalle de confiance se calcule comme :

$$FP_{t\,\mathrm{inf}} = \frac{A_{0\,\mathrm{inf}} - E_0(A_0)}{A_{0\,\mathrm{inf}}}$$

$$FP_{t\,\mathrm{sup}} = \frac{A_{0\,\mathrm{sup}} - E_0(A_0)}{A_{0\,\mathrm{sup}}}$$

Cette dernière formulation permet aussi des calculs exacts.

EXEMPLE 7.7

FRACTION PRÉVENUE TOTALE POUR LE RAPPORT DE TAUX

Considérons les données du tableau 7.8, qui décrit l'effet protecteur d'un facteur *X* contre la maladie *Y*.

L'effet protecteur *Ef* du vaccin est mesuré par : $Ef = \dfrac{50}{1000} \Big/ \dfrac{10}{2000} = 10$.

La fraction prévenue FP_t est estimée par $FP_t = \dfrac{2000}{3000} \times \dfrac{10-1}{10} = 0,60$.

On remarque aussi qu'à un $RT = \dfrac{10 \times 1000}{50 \times 2000} = 0,10$ correspond une fraction prévenue totale de $FP_t = \dfrac{2000}{3000} \times (1 - 0,10) = 0,6$.

Les limites de confiance à 95 % de FP_t peuvent alors facilement être déduites de celles du *RT* obtenues par les différentes méthodes de calcul.

Le tableau 7.12 présente les intervalles de confiance à 95 % du *RT* suivant différentes méthodes de calcul, et ceux de FP_t qui leur correspondent.

TABLEAU 7.12

Méthode de calcul	Intervalle de confiance à 95 %			
	Du *RT*	De A_0	De FP_t	Programme
Exact	(0,0481 ; 0,1916)	(43,38 ; 54,73)	(0,5390 ; 0,6346)	PR7.25
Normale simple	(0,0507 ; 0,1972)	(43,03 ; 54,48)	(0,5352 ; 0,6329)	PR7.26
Résultat d'un test	(0,0577 ; 0,1732)	(44,56 ; 53,79)	(0,5512 ; 0,6282)	PR7.27
RV	(0,0478 ; 0,1886)	(43,57 ; 54,76)	(0,5409 ; 0,6348)	PR7.28

Pour obtenir les valeurs de $A_{0\,\text{inf}}$ et $A_{0\,\text{sup}}$ correspondant aux valeurs RT_{sup} et RT_{inf}, on rappelle la relation $E(A_1 \mid RT) = \pi \times m_1$ où $\pi = \dfrac{RT \times n_1}{RT \times n_1 + n_0}$. Ainsi : $A_{0\,\text{inf}} = m_1 - A_{1\,\text{sup}} = m_1 - \dfrac{RT_{\text{sup}} n_1}{RT_{\text{sup}} n_1 + n_0} \times m_1$. De façon analogue, on obtient la valeur de $A_{0\,\text{sup}}$ à partir de RT_{inf}. Pour l'approche exacte, on obtient les valeurs suivantes : $A_{0\,\text{inf}} = 43,3777$ et $A_{0\,\text{sup}} = 54,7345$. On en dérive les intervalles de confiance en utilisant les relations 7.9 ou 7.10.

◆

Exemple 7.8

Fraction prévenue totale pour le rapport de proportions

Considérons les données du tableau 7.10 décrivant l'effet d'un vaccin dans la protection de la maladie *M*. L'effet protecteur du vaccin est mesuré par *Ef*, l'inverse du *RP* : $Ef = \frac{80 \times 200}{40 \times 100} = 4$. Cette mesure indique de combien de fois le risque de la maladie chez les non-vaccinés est plus fort que celui des vaccinés.

La fraction prévenue FP_t est estimée par $FP_t = \frac{200}{300} \times \frac{4-1}{4} = 0,50$.

On remarque aussi qu'à un $RP = \frac{40 \times 100}{80 \times 200} = 0,25$ correspond une fraction prévenue totale de $FP_t = \frac{200}{300} \times (1 - 0,25) = 0,50$.

Les limites de confiance à 95 % de FP_t peuvent alors facilement être déduites de celles du *RP* obtenues par les différentes méthodes de calcul.

Le tableau 7.13 présente les intervalles de confiance à 95 % du *RP* suivant différentes méthodes de calcul, et ceux de FP_t qui leur correspondent.

Tableau 7.13

Méthode de calcul	Intervalle de confiance à 95 %			
	Du *RT*	De A_0	De FP_t	Programme
Exact	[0,1993 ; 0,3223]	[72,96 ; 85,80]	[0,45178 ; 0,53383]	Pr7.29
Normale simple	[0,1863 ; 0,3354]	[71,82 ; 87,43]	[0,44304 ; 0,54245]	Pr7.30
Résultat d'un test	[0,1904 ; 0,3282]	[72,45 ; 86,91]	[0,44787 ; 0,53971]	Pr7.31
RV	90,1832 ; 0,3308]	[72,22 ; 87,82]	[0,44614 ; 0,54453]	Pr7.32

◆

PARTIE

4

ANALYSE STRATIFIÉE : PLUSIEURS TABLEAUX 2 × 2

CHAPITRE 8

LES TECHNIQUES DE BASE EN ANALYSE STRATIFIÉE

On s'intéresse ici à une étude sur l'association entre un facteur d'exposition *X* et une maladie *Y*, en présence d'un tiers facteur *F*. On veut alors faire ressortir l'effet de *X* sur *Y* tout en contrôlant l'influence que peut avoir le facteur *F* sur cet effet de *X*. On distingue principalement deux types d'influence du facteur *F* : la confondance et la modification (ou modifiance), concepts que nous définissons maintenant.

La force d'association entre *X* et *Y* peut varier suivant les catégories de *F*. En épidémiologie, dans le contexte de la causalité, cette situation s'interprète souvent comme la présence de modification. Le facteur *F*, alors dit modifiant, influence la force d'association entre *X* et *Y*.

Il peut arriver aussi que le facteur *F*, non considéré dans les analyses, perturbe la mesure d'association établie entre *X* et *Y*. Cette perturbation peut se produire si *F* est un facteur de risque de la maladie *Y* et un facteur associé à *X* de façon concomitante ou prédictive. L'association de *F* à *Y,* via celle de *F* à *X,* se mélange à celle que l'on veut mesurer entre *X* et *Y*. Pour cette dernière, le facteur *F* est dit alors facteur de confusion ou confondant.

Ainsi, soit pour faire ressortir les effets de modification de *F* sur l'association entre *X* et *Y*, soit pour contrôler son effet de confondance sur la mesure globale, il peut être intéressant, sinon nécessaire, d'utiliser une analyse stratifiée suivant les catégories de *F*.

8.1 ANALYSE STRATIFIÉE

Pour simplifier, on suppose que *X* et *Y* sont des variables dichotomiques et *F*, une variable catégorielle comprenant *k* catégories. Pour chacune des catégories (ou strates) *i* de *F*, les variables A_{1i} et A_{0i} désignent les nombres de cas respectivement chez les n_{1i} exposés et les n_{0i} non-exposés au facteur *X*. Ces variables sont des variables de Poisson ou binomiales suivant le type de données. Pour la strate *i* de *F*, la description des données se fait suivant le schéma du tableau 8.1. Les valeurs a_{1i} et a_{0i} sont les valeurs observées pour les variables A_{1i} et A_{0i} respectivement. Si les données sont en personnes-temps, les nombres b_{1i} et b_{0i} ne sont pas définis.

Les risques (taux ou proportions) R_{1i} chez les exposés à *X* et R_{0i} chez les non-exposés sont définis comme précédemment :

$$R_{1i} = \frac{a_{1i}}{n_{1i}} \text{ et } R_{0i} = \frac{a_{0i}}{n_{0i}}$$

TABLEAU 8.1

F Strate *i*	*X* = 1	*X* = 0	Total
Y = 1	a_{1i}	a_{0i}	m_{1i}
Y = 0*	b_{1i}	b_{0i}	m_{1i}
Total	n_{1i}	n_{0i}	n_i

* Si les données sont en personnes-temps, cette ligne n'est pas définie.

Désignons par *m* l'une quelconque des mesures d'association spécifiées plus haut ou une transformation de celles-là. La mesure spécifique à la strate *i* est désignée par m_i, alors que m_B et m_a représentent respectivement la mesure brute et la mesure ajustée (ou pondérée). La mesure brute est celle obtenue sur le tableau global (tableau 8.2). La mesure ajustée m_a

(que l'on désignera aussi simplement par m) est obtenue par une somme pondérée des différentes mesures spécifiques : $m_a = \sum_i \lambda_i m_i$, où les λ_i sont des valeurs telles que $0 < \lambda_i < 1$ et $\Sigma\lambda_i = 1$. Ces λ_i définissent un système de poids.

TABLEAU 8.2

Toutes les strates		$X = 1$	$X = 0$	Total
	$Y = 1$	Σa_{1i}	Σa_{0i}	Σm_{1i}
	$Y = 0$*	Σb_{1i}	Σb_{0i}	Σm_{1i}
	Total	Σn_{1i}	Σn_{0i}	Σn_i

* Si les données sont en personnes-temps, cette ligne n'est pas définie.

8.1.1 MESURES SPÉCIFIQUES ET MODIFICATION

En analyse stratifiée, les mesures spécifiques m_i varient en général d'une strate à l'autre. Ces variations peuvent être compatibles avec l'hypothèse d'homogénéité des mesures spécifiques. Cette hypothèse peut formellement s'énoncer comme suit :

Hypothèse d'homogénéité : $\mu_1 = \mu_2 = \ldots = \mu_k = \mu$

où les μ_i représentent les mesures spécifiques réelles d'association entre X et Y.

Les mesures spécifiques m_i sont alors toutes des estimations d'une même mesure théorique μ. Sous cette hypothèse, les variations observées entre les mesures m_i s'expliquent par les simples fluctuations aléatoires. Dans un tel cas, les mesures spécifiques pourraient être résumées en une mesure pondérée globale m. Les poids utilisés pour définir la mesure pondérée sont le plus souvent proportionnels à l'inverse des variances des mesures spécifiques. Un tel choix permet de définir une mesure résumée à variance minimale.

Si les variations observées sont trop importantes pour être compatibles avec l'hypothèse d'homogénéité, elles pourraient alors refléter la présence d'une réelle hétérogénéité entre les mesures spécifiques.

Hypothèse d'hétérogénéité : il existe un certain couple de mesures (μ_i, μ_j) telles que $\mu_i \neq \mu_j$.

Souvent cette hétérogénéité peut s'interpréter comme la modification de F sur l'association entre X et Y ou la synergie entre les deux facteurs X et F sur le risque de Y. (Dans la section, 8.1.2 qui suit, nous rappelons ce concept de synergie ou d'interaction entre deux facteurs.)

Pour faire ressortir l'hétérogénéité des mesures spécifiques, il suffit de les présenter par strates. Parfois, pour des raisons pratiques, il peut être désirable de résumer ces mesures spécifiques en une mesure pondérée *m*. Dans ce cas, les systèmes de poids sont choisis en fonction de l'intérêt qu'a la mesure pondérée.

8.1.2 MESURE D'INTERACTION OU DE SYNERGIE ENTRE DEUX FACTEURS

Au concept de modification (ou de modifiance) se joint celui de synergie entre deux facteurs. Ces deux notions traduisent chacune une perspective particulière dans l'interprétation du concept statistique plus vaste d'interaction. Quand la modification s'intéresse à l'influence que peut avoir un tiers facteur (le facteur modifiant) sur l'intensité de l'association entre un facteur principal et une maladie, la synergie entre deux facteurs s'intéresse à l'effet supplémentaire que peut avoir la seule combinaison de ces deux facteurs sur le risque de la maladie. Mais, au plan statistique, l'analyse d'une synergie ou d'un effet modifiant revient à considérer l'interaction entre deux facteurs.

En épidémiologie, on distingue principalement deux types de synergie. Le premier marque l'influence que peut avoir la combinaison de deux facteurs sur l'association mesurée par la différence *DR* des risques et le second, sur l'association mesurée par le rapport *RR* des risques. Le premier type est dit de modèle additif et le second, de modèle multiplicatif.

Pour bien fixer les concepts, considérons les données d'une étude portant sur l'association entre les deux facteurs d'exposition, X_1 et X_2, et le risque d'une maladie *Y*. Les données sont disposées suivant un schéma comme celui du tableau 8.3.

TABLEAU 8.3

	X_1X_2				Total
	11	**01**	**10**	**00**	
$Y = 1$	a_{11}	a_{01}	a_{10}	a_{00}	m_1
$Y = 0$	b_{11}	b_{01}	b_{10}	b_{00}	m_0
Total	n_{11}	n_{01}	n_{10}	n_{00}	n

Les différentes modalités d'exposition aux facteurs X_1 et X_2 sont désignées par ij : (i ou j) = (1 ou 0) suivant l'exposition ou la non-exposition à X_1 ou à X_2.

Remarquons que ces données pourraient être aussi disposées suivant un schéma stratifié, comme au tableau 8.1, où X est mis pour X_1 et F (dichotomique) pour X_2.

On désigne par R_{ij} le risque de la maladie Y pour la modalité (ij). Si on prend la modalité 00 comme référence, le risque R_{00} peut être considéré comme le risque de base. On comparera les risques des autres modalités à ce risque de base pour composer les différentes mesures d'association DR_{ij} ou RR_{ij}. Ainsi, DR_{10} correspond à la différence entre les risques R_{10} et R_{00}.

Synergie (interaction) dans le modèle additif

La synergie dans le modèle additif (ou synergie additive) mesure l'effet présumé d'un tiers facteur (l'interaction) sur le risque de la maladie Y, effet qui s'ajoute aux effets indépendants des deux facteurs combinés ; ces effets sont mesurés par la différence des risques. Les effets des facteurs X_1 et X_2 pris de façon indépendante, sont respectivement mesurés par DR_{10} et DR_{01}. Par ailleurs, pour une exposition simultanée aux deux facteurs, le risque R_{11} est déterminé par les effets indépendants des deux facteurs plus, peut-être, celui d'un tiers facteur créé par la combinaison des deux premiers. Si tel est le cas, alors la mesure d'effet DR_{11} devrait différer de la somme des mesures DR_{10} et DR_{01}. Il y a synergie additive positive si DR_{11} est plus grande, ou synergie additive négative si DR_{11} est plus petite que la somme $(DR_{10} + DR_{01})$. On peut alors mesurer la synergie additive I_+ comme :

$$I_+ = DR_{11} - (DR_{10} + DR_{01})$$

Dans le tableau 8.4, nous reprenons la description des effets liés aux différents facteurs, X_1, X_2 et I_+ sur le risque de base R_{00}.

Tableau 8.4

	$X_1 = 0$	$X_1 = 1$
$X_2 = 0$	R_{00}	$R_{00} + e_1$
$X_2 = 1$	$R_{00} + e_2$	$R_{00} + e_1 + e_2 + \mathrm{I}_+$

Les termes e_1 et e_2 correspondent respectivement aux effets DR_{10} et DR_{01} des facteurs X_1 et X_2.

Sous la forme I_+, la mesure d'interaction dans le modèle additif n'est pas applicable aux données d'études cas-témoins. Pour adapter la mesure à ce contexte, on propose de la transformer en la rapportant à la mesure de fréquence de base R_{00}. Elle devient ainsi :

$$\frac{I_+}{R_{00}} = RR_{11} - RR_{10} - RR_{01} + 1$$

Synergie (interaction) dans le modèle multiplicatif

La synergie dans le modèle multiplicatif (ou synergie multiplicative) mesure l'effet présumé d'un tiers facteur (l'interaction) sur le risque de la maladie Y, effet qui s'ajoute aux effets indépendants des deux facteurs combinés ; ces effets sont mesurés par le rapport des risques. S'il y a interaction multiplicative entre les facteurs X_1 et X_2, alors la mesure RR_{11} devrait différer du produit des mesures RR_{10} et RR_{01}. Il y a synergie multiplicative positive si RR_{11} est plus grand, ou synergie multiplicative négative si RR_{11} est plus petit que le produit ($RR_{10} \times RR_{01}$). On peut alors mesurer la synergie multiplicative $I_\times$ comme :

$$I_\times = \frac{RR_{11}}{RR_{10} \times RR_{01}}$$

Dans le tableau 8.5, nous reprenons la description des effets liés aux différents facteurs, X_1, X_2 et $I_\times$ sur le risque de base R_{00}.

Tableau 8.5

	$X_1 = 0$	$X_1 = 1$
$X_2 = 0$	R_{00}	$R_{00} \times e_1$
$X_2 = 1$	$R_{00} \times e_2$	$R_{00} \times e_1 \times e_2 \times I_\times$

Les termes e_1 et e_2 correspondent respectivement aux effets RR_{10} et RR_{01} des facteurs X_1 et X_2.

8.1.3 Mesure brute et confondance

L'effet confondant du facteur F sur l'association entre X et Y met en cause la validité de la mesure brute de cette association. Au plan pratique, on peut distinguer deux types de confondance : la confondance *absolue* et la confondance *relative*.

La confondance *absolue* est celle qui se manifeste lorsque la mesure brute m_B se situe à l'extérieur de l'intervalle déterminé par la plus petite et la plus grande des mesures spécifiques m_i (figure 8.1) :

$$m_B < \min \{m_i\} \text{ ou } m_B > \max \{m_i\}$$

On comprend bien alors qu'aucune somme pondérée des mesures spécifiques ne peut conduire à la valeur de m_B. Un tel type de confondance peut se présenter lorsque la modification est faible et que les associations de F à X et de F à Y sont conjointement fortes. Dans une telle situation, la mesure brute n'est pas valide et devrait être remplacée par une somme pondérée m des mesures spécifiques. Cette mesure est dite aussi ajustée pour F.

La confondance *relative* est une situation complémentaire à la précédente. Bien que les associations conjointes de F à X et de F à Y soient non nulles, elles ne sont pas assez fortes pour créer une confondance *absolue*. La mesure brute se retrouve quelque part entre la plus petite et la plus grande des mesures spécifiques (figure 8.2) :

$$\min\{m_i\} \leq m_B \leq \max\{m_i\}$$

Dans une telle situation, le jugement pratique sur la confondance est relatif au type d'ajustement que l'on décide de pratiquer. En effet, on peut montrer qu'il existe au moins un système de poids $\{\lambda_i\}$ tel que $\Sigma\lambda_i m_i = m_B$. Mais, pour des raisons pratiques, on peut très bien décider d'utiliser une mesure ajustée conventionnelle dont la valeur sera le plus souvent différente de m_B.

On peut donc énoncer la règle pratique générale suivante :

un facteur F a un effet de confondance sur la mesure brute d'association m_B si cette mesure est différente d'une mesure m ajustée pour ce facteur.

FIGURE 8.1 : CONFONDANCE ABSOLUE

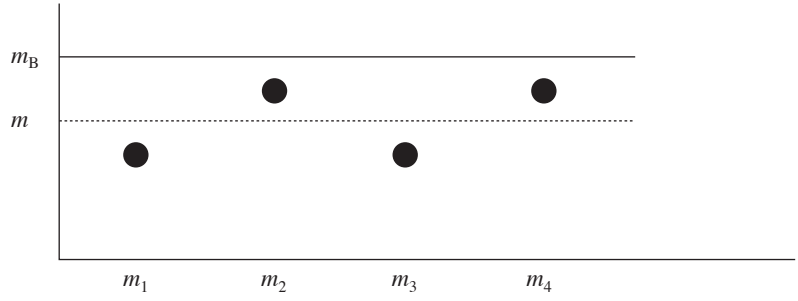
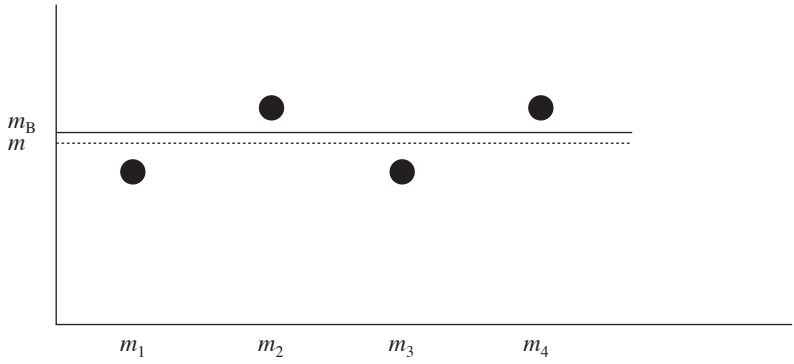


Figure 8.2 : Confondance relative



8.1.4 Association globale et homogénéité

S'il y a association entre X et Y sur chacune des k strates du facteur F, on peut s'intéresser aux trois questions suivantes :

i) Comment peut-on tester l'homogénéité des m_i ?

ii) Comment peut-on tester l'association globale ?

iii) Comment peut-on mesurer l'association globale ?

Pour répondre à ces trois questions, nous proposons deux approches : l'approche en approximation normale et celle par la méthode du rapport de vraisemblance. Nous présentons ici sommairement ces deux approches. Par la suite, dans les chapitres 9 et 10, nous les appliquerons aux différentes mesures d'association.

8.2 Analyse stratifiée : approche en approximation normale

8.2.1 Poids proportionnels à l'inverse des variances

Pour simplifier la description de l'approche, on suppose que, sous l'hypothèse nulle, les mesures m_i fluctuent autour de 0 ; sinon, on leur applique une transformation appropriée pour qu'il en soit ainsi.

Rappelons que la mesure pondérée m est définie comme $m = \frac{\sum w_i m_i}{\sum w_i}$, où les w_i définissent le système de poids $\{\lambda_i\}$: $\lambda_i = \frac{w_i}{\sum_i w_i}$. On considère ici en particulier les poids proportionnels à l'inverse de la variance V_i des mesures m_i : $w_i = \frac{1}{V_i}$. Alors, dans la pondération de m, la mesure m_i a un poids d'autant plus important qu'elle est stable. Dans ces conditions, on peut montrer que $V(m) = \frac{1}{\sum_i w_i}$.

Si les strates sont indépendantes, la statistique $\sum_{i=1}^{k} w_i m_i^2$ obéit approximativement à un χ^2 à k degrés de liberté. Cette statistique, dite le khi-carré total et notée $\chi_k^2(total)$, marque la variation totale liée aux variations de chaque mesure m_i autour de sa valeur nulle (figure 8.3). Par exemple, pour la mesure spécifique m_2 de la figure, la déviation totale **T** est expliquée par la déviation **A** (= m – 0) de la mesure pondérée m par rapport à la valeur nulle et par la déviation **B** (= $m_2 - m$) de la mesure spécifique par rapport à la moyenne m.

Ainsi, ce khi-carré total peut être partitionné suivant deux sources de variation.

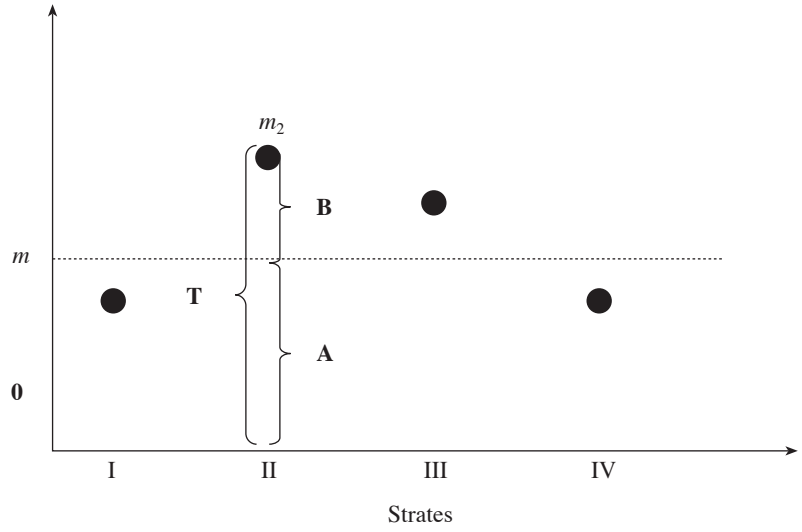
1. La première source est liée à la tendance générale qu'ont les mesures à se détacher de la valeur nulle. En d'autres termes, cette première source réfère à la tendance qu'a la moyenne générale m de s'éloigner de la valeur nulle et traduit ainsi une association globale entre X et Y. Le test sur l'association se définit simplement comme :
$$\chi_1^2(\text{assoc}) = \frac{m^2}{V(m)} = \frac{\left(\sum w_i m_i\right)^2}{\sum w_i}$$

2. La deuxième source est celle liée à la tendance qu'ont les mesures à s'éloigner de leur moyenne commune m. Cette variation traduit donc une hétérogénéité entre les mesures. Le test sur l'homogénéité des mesures se définit comme :
$$\chi_{k-1}^2(\text{homog}) = \sum w_i \left(m_i - m\right)^2$$

On peut donc facilement montrer que
$\chi^2_k(\text{total}) = \chi^2_1(\text{assoc}) + \chi^2_{k-1}(\text{homog})$

En pratique, il est plus simple de calculer $\chi^2_k(\text{total})$ et $\chi^2_1(\text{assoc})$. Le $\chi^2_{k-1}(\text{homog})$ peut alors être déduit par simple soustraction.

FIGURE 8.3



Dans le cadre de cette méthode, on peut facilement décrire l'intervalle de confiance de niveau $100(1 - \alpha)\%$ pour la mesure pondérée m : $m \pm z_{a/2}\sqrt{V(m)}$.

8.2.2 APPROCHES PLUS SPÉCIFIQUES AUX MESURES *DR*

Considérons la différence *DR* (de taux *DT* ou de proportions *DP*). Suivant la notation du tableau 8.1, pour une strate i, sous les conditions de marges fixes, on désigne par d_i la déviation de la valeur observée a_{1i} par rapport à sa valeur attendue $E_0(A_{1i})$ calculée sous l'hypothèse nulle : $d_i = [a_{1i} - E_0(A_{1i})]$. Si w_i représente l'inverse de la variance de A_{1i} estimée sous l'hypothèse nulle, alors la statistique $\chi^2_i = w_i d_i^2$ permet de tester l'hypothèse d'une association nulle sur cette strate. Par ailleurs, la statistique $\sum_i w_i d_i^2$ mesure la variation totale liée aux fluctuations des mesures DR_i à travers les strates i, sous cette même hypothèse. En d'autres termes, on a

$$\sum_i w_i d_i^2 = \sum_i w_i^* DR_i^2 \text{ , où } w_i^* = \frac{1}{V_0\left[DR_i\right]}.$$

Suivant la technique décrite au paragraphe précédent, le khi-carré total peut être partitionné en

- un khi-carré d'association :

$$\chi_1^2(\text{assoc}) = \frac{\left(\sum_i w_i d_i\right)^2}{\sum_i w_i} = \frac{\left(\sum_i w_i^* DR_i\right)^2}{\sum_i w_i^*}$$

- et un khi-carré d'homogénéité :

$$\chi_{k-1}^2(\text{homog}) = \sum_i w_i \left(d_i - \bar{d}\right)^2$$
$$= \sum_i w_i^* \left(DR_i - \overline{DR}\right)^2$$

Le premier permet de porter un jugement sur la signification statistique de la différence moyenne $\overline{DR}$.

En résumé, on peut dire que l'application de la technique de la pondération par l'inverse des variances à la mesure de déviation $[a_{1i} - E_0(A_{1i})]$ définit des approches pour les différences de taux ou de proportions, suivant le cas.

8.2.3 TEST DE MANTEL-HAENSZEL D'ASSOCIATION EN ANALYSE STRATIFIÉE

Le test de Mantel-Haenszel en analyse stratifiée se rapporte plus directement aux rapports de taux ou aux rapports de cotes.

Suivant les notations du tableau 8.1, pour une strate *i*, le khi-carré de Mantel-Haenszel se présente comme :

$$\chi_{MH}^2 = \frac{\left[\sum_i a_{1i} - \sum_i E_0(A_{1i})\right]^2}{\sum_i V_0(A_{1i})} = \frac{\left(\sum_i d_i\right)^2}{\sum_i V_0(d_i)}$$ avec un degré de liberté. Il permet de tester l'association globale mesurée en terme de rapport de taux ou de rapport de cotes, suivant le cas. Par ailleurs, en référence à ce qui a été dit précédemment, il pourrait être trompeur d'utiliser le khi-carré de Mantel-Haenszel comme composante d'une partition du khi-carré total $\sum_i w_i d_i^2$ défini à partir des déviations d_i. On peut facilement construire des contre-exemples numériques où, en situation de parfaite homogénéité

pour les rapports de taux ou de cotes, le khi-carré total soit manifestement plus grand que le khi-carré d'association correspondant à ces mesures, ce qui laisserait supposer une situation d'hétérogénéité alors qu'il n'y en a pas.

8.2.4 TEST DE BRESLOW-DAY SUR L'HOMOGÉNÉITÉ DES MESURES

Le test de Breslow-Day permet de juger l'homogénéité des rapports de cotes en analyse stratifiée. Sous l'hypothèse d'homogénéité (hypothèse nulle), le rapport de cotes est constant à travers les strates : $RC = \psi$. De faibles écarts mesurés sur l'ensemble des strates entre les nombres observés et les nombres attendus de cas exposés soutiennent cette hypothèse de l'homogénéité ; des écarts importants invitent à son rejet.

Considérons les variables hypergéométriques A_{1i}, de paramètres $\{n_i, m_{1i}, n_{1i}, \psi\}$.

On désigne par $E(A_{1i}|\psi)$ et $V(A_{1i}|\psi)$ respectivement la valeur attendue et la variance de A_{1i} sous l'hypothèse d'un rapport de cotes constant et égal à ψ. Le test se présente alors comme :

$$\chi^2_{k-1} = \sum_{i=1}^{K} \frac{\left(a_{1i} - E(A_{1i} \mid \psi)\right)^2}{V(A_{1i} \mid \psi)}$$

Pour la conduite du test, la valeur de ψ peut être toute valeur de la forme $\psi = \sum_i \lambda_i \psi_i$ où les λ_i forment un système de poids arbitraire. L'estimateur du maximum de vraisemblance ψ et celui de Mantel-Haenszel sont les plus souvent utilisés.

Le test de Breslow-Day peut facilement être étendu au problème de l'homogénéité de différences ou de rapports de taux ou de proportions en analyse stratifiée. La variable A_{1i} est alors une variable binomiale ou hypergéométrique, suivant le contexte.

8.2.5 INTERVALLE DE CONFIANCE D'UNE MESURE PONDÉRÉE

Considérons un système de poids $\{\lambda_i\}$ à partir duquel sont résumées les mesures spécifiques m_i en une mesure globale : $m = \sum_{i=1}^{k} \lambda_i m_i$. Alors, nous voulons calculer l'intervalle de confiance de m au niveau $100(1 - \alpha)$ % de

confiance. Comme on l'a vu précédemment, cet intervalle est donné par : $m \pm z_{\alpha/2}\sqrt{V(m)}$. La variance de m peut être estimée simplement par $V(m) = \sum_i \lambda_i^2 V_i$, où V_i désigne la variance de la mesure spécifique m_i.

Supposons maintenant que m représente le logarithme d'un *RR* pondéré (*RT*, *RP* ou *RC* pondéré) : $m = \log(RR)$ où $RR = \sum_i \lambda_i RR_i$. Alors, à partir de l'intervalle de confiance sur $\log(RR)$ on peut facilement déduire celui du *RR* par transformation inverse :

$$\log(RR) \pm z_{\alpha/2}\sqrt{V[\log(RR)]} \Rightarrow RR \times e^{\pm z_{\alpha/2}\sqrt{V[\log(RR)]}}$$

Le problème est résolu si l'on sait calculer la variance $V[\log(RR)]$. Par la méthode delta (section 2.10.3 du chapitre 2), cette variance se décrit simplement comme $V[\log(RR)] \approx \sum_i \left\{ \frac{\lambda_i^2 RR_i^2 V_i}{RR^2} \right\}$ et se simplifie à $V[\log(RR)] \approx \sum_i \lambda_i^2 V_i$ sous l'hypothèse d'uniformité des mesures RR_i.

8.3 ANALYSE STRATIFIÉE : APPROCHE PAR LA MÉTHODE DU RAPPORT DE VRAISEMBLANCE

Nous décrivons d'abord le risque $\pi(X,F)$ de la maladie Y comme une fonction linéaire du facteur d'exposition X (dichotomique) et du facteur de stratification F. À cette fin, nous proposons trois modèles :

Modèle 0 : $\pi(\cdot,F) = \alpha_0 + \beta_{01}F$

Modèle 1 : $\pi(X,F) = \alpha_1 + \beta_{11}F + \beta_{12}X$

Modèle 2 : $\pi(X,F) = \alpha_2 + \beta_{21}F + \beta_{22}X + \beta_{23}XF$

Le premier modèle ignore l'effet de X. Le deuxième modèle intègre les effets de X et de F et le troisième ajoute à ces effets celui de l'interaction XF entre les deux facteurs. Les coefficients des modèles sont estimés par la méthode du maximum de vraisemblance.

Pour une codification de $X = 1$ ou 0 désignant respectivement le statut d'exposé et de non-exposé à X, le coefficient β_{12} de X du modèle 1 correspond à la différence m entre les deux risques, ajustée pour la variable F : $m = \pi(1,F) - \pi(0,F)$.

Le coefficient β_{23} du facteur d'interaction *XF* du modèle 2 marque la modifiance de F sur l'effet de *X*.

8.3.1 TESTS STATISTIQUES SUR L'ASSOCIATION ET L'HOMOGÉNÉITÉ

Pour juger de la signification du coefficient β_{12} sous l'hypothèse nulle, on peut utiliser les tests de Wald ou du rapport de vraisemblance. Ce dernier test s'obtient en comparant la vraisemblance du modèle 0 à celle du modèle 1 :

$$\chi_1^2 = -2\log\left[\frac{FV(\alpha_0,\beta_{01})}{FV(\alpha_1,\beta_{11},\beta_{12})}\right]$$

Le jugement sur l'homogénéité des mesures m_i passe simplement par un jugement sur le coefficient β_{23} de l'interaction entre la variable *X* et *F* du modèle 2. Basé sur la comparaison des vraisemblances des modèles 1 et 2, le test du rapport de vraisemblance sur le coefficient β_{23} se présente comme :

$$\chi_1^2 = -2\log\left[\frac{FV(\alpha_1,\beta_{11},\beta_{12})}{FV(\alpha_2,\beta_{21},\beta_{22},\beta_{23})}\right]$$

Si *F* a plusieurs catégories, alors on utilise autant de variables indicatrices qu'il y a de catégories moins un pour entrer la variable *F* dans les modèles.

8.3.2 INTERVALLES DE CONFIANCE D'UNE MESURE AJUSTÉE OU D'UNE INTERACTION

Dans la suite de ces tests, on peut aussi définir les intervalles de confiance par les mêmes méthodes, de Wald ou du rapport de vraisemblance.

Intervalle de confiance d'une mesure ajustée

Pour l'intervalle de confiance de *m* ajustée pour *F*, dans le cadre du modèle 1, il suffit de résoudre l'équation

$$2\left[L - L(\beta^*)\right] = \chi^2_{1,1-\alpha}$$

où *L* correspond à la valeur maximale de la fonction du $\log FV(\alpha_1,\beta_{11},\beta_{12})$ et $L(\beta^*)$ à la valeur maximale de la fonction du $\log FV(\alpha_1,\beta_{11},\beta^*)$ pour une valeur fixe β^* de β_{12}.

Intervalle de confiance d'une interaction

Pour l'intervalle de confiance de l'interaction *XF*, dans le cadre du modèle 2, il suffit de résoudre l'équation

$$2\left[L - L(\beta^{*}) \right] = \chi^{2}_{1,1-\alpha}$$

où L correspond à la valeur maximale de la fonction du log $FV(\alpha_1, \beta_{21}, \beta_{22}, \beta_{23})$ et $L(\beta^*)$ à la valeur maximale de la fonction du log $FV(\alpha_2, \beta_{21}, \beta_{22}, \beta^*)$ pour une valeur fixe β^* de β_{23}.

Il peut être fastidieux, voire pratiquement impossible de décrire de façon plus explicite les approches de vraisemblance en analyse stratifiée. Nous proposons donc d'utiliser directement la procédure GENMOD qui permet de telles analyses. Il suffit de bien préciser la fonction de lien, c'est-à-dire la transformation appropriée de la mesure de base, et la loi de distribution sous-jacente, le plus souvent une loi binomiale ou une loi de Poisson.

CHAPITRE

9

MESURES D'ASSOCIATION BASÉES SUR LES TAUX EN ANALYSE STRATIFIÉE

Les mesures d'association décrites dans ce chapitre sont les différences de taux *DT*, les rapports de taux *RT*, les interactions additive et multiplicative et le *SMR*. Pour chacune de ces mesures, nous présentons d'abord les tests statistiques, puis les intervalles de confiance, en approximation normale et par la méthode du rapport de vraisemblance. Pour le *SMR*, nous y ajouterons la méthode exacte de calcul.

9.1 DIFFÉRENCE DE DEUX TAUX EN ANALYSE STRATIFIÉE

On considère les données (en personnes-années) d'une étude de cohortes portant sur l'association entre le facteur X et la maladie Y. Le facteur F est un facteur à contrôler. Ce facteur F a k catégories. Dans le tableau 9.1 sont décrites les données de l'étude pour la strate ou la catégorie i du facteur F.

TABLEAU 9.1

Facteur F		$X = 1$	$X = 0$	Total
Strate i	$Y = 1$	a_{1i}	a_{0i}	m_{1i}
	Personnes-années	n_{1i}	n_{0i}	n_i

Sur cette strate, les taux t_{1i} et t_{0i} sont respectivement définis comme :

$$t_{1i} = \frac{a_{1i}}{n_{1i}} \text{ et } t_{0i} = \frac{a_{0i}}{n_{0i}}.$$

La différence DT_i correspondante des taux est alors : $DT_i = t_{1i} - t_{0i}$.

Toute mesure pondérée DT, résumant les mesures spécifiques DT_i en une mesure totale, se présente comme $DT = \sum_i \lambda_i DT_i$, où les λ_i constituent un système de poids.

9.1.1 TESTS STATISTIQUES EN APPROXIMATION NORMALE

Pour répondre aux problèmes de l'homogénéité des mesures spécifiques DT_i et de la signification statistique de la mesure pondérée DT, nous référons à la partition du $\chi^2_k(\text{total})$ en $\chi^2_1(\text{assoc})$ et $\chi^2_{k-1}(\text{homog})$. En choisissant les poids λ_i proportionnels à l'inverse des variances V_i des mesures DT_i, nous obtenons simplement les statistiques :

$$\begin{cases} \chi^2_{\text{total}} = \sum_i w_i DT_i^2 \\ \chi^2_{\text{assoc}} = \dfrac{\left(\sum_i w_i DT_i\right)^2}{\sum_i w_i} \\ \chi^2_{\text{homog}} = \sum_i w_i \left(DT_i - DT\right)^2 \end{cases}$$

où $w_i = \dfrac{1}{V_i}$. La variance V_i de la mesure DT_i est donnée par : $V_i = \dfrac{a_{1i}}{n_{1i}^2} + \dfrac{a_{0i}}{n_{0i}^2}$.

Ainsi, le khi-carré d'association porte sur la mesure pondérée DT qui est confrontée à l'hypothèse nulle d'une différence égale à 0. Par ailleurs, le khi-carré d'homogénéité porte sur l'hypothèse de l'homogénéité des mesures spécifiques DT_i entre elles.

9.1.2 TEST DE BRESLOW-DAY SUR L'HOMOGÉNÉITÉ DES MESURES SPÉCIFIQUES DT_i

Sous l'hypothèse de leur homogénéité, les DT_i fluctuent aléatoirement autour d'une même mesure paramétrique Δ. C'est donc dire que les déviations $\left[a_{1i} - E(A_{1i} \mid \Delta)\right]$ fluctuent aléatoirement autour de 0 avec une variance de $V(A_{1i} \mid \Delta)$.

Le test de Breslow-Day se présente alors comme :

$$\chi^2_{k-1} = \sum_{i=1}^{K} \frac{\left[a_{1i} - E(A_{1i} \mid \Delta)\right]^2}{V(A_{1i} \mid \Delta)}$$

La différence Δ est estimée par la différence DT pondérée par l'inverse des variances V_i. Sous l'hypothèse d'une différence homogène de valeur Δ, la variable A_{1i} obéit à une loi binomiale de paramètres $\pi_i = \dfrac{\Delta n_{1i} n_{0i} + m_{1i} n_{1i}}{m_{1i} n_i}$ et m_{1i}. La valeur attendue $E(A_{1i} \mid \Delta)$ et la variance $V(A_{1i} \mid \Delta)$ de A_{1i} sont alors respectivement données par :

$$E(A_{1i} \mid \Delta) = m_{1i}\pi_i$$
$$V(A_{1i} \mid \Delta) = m_{1i}\pi_i(1 - \pi_i)$$

9.1.3 INTERVALLE DE CONFIANCE EN APPROXIMATION NORMALE DE *DT* PONDÉRÉE

Supposons que l'on s'intéresse à la différence pondérée des taux $DT = \sum_i \lambda_i DT_i$, où $\{\lambda_i\}$ représente le système de poids. Alors, l'intervalle de confiance s'exprime comme $DT \pm z_{\alpha/2}\sqrt{V}$. Une bonne estimation de la variance V est donnée par $V = \sum_i \lambda_i^2 V_i$, où $V_i = \dfrac{a_{1i}}{n_{1i}^2} + \dfrac{a_{0i}}{n_{0i}^2}$ est la variance de la mesure spécifique DT_i. Si le système de poids est tel que $\lambda_i \propto 1/V_i$, c'est-à-dire proportionnel à la stabilité de la mesure DT_i, alors la variance V correspond simplement à $\left[\sum_i \dfrac{1}{V_i}\right]^{-1}$. On rappelle que le système de poids $\{\lambda_i\}$ est théoriquement arbitraire.

Ci-dessous, nous proposons différents systèmes de poids, parmi les plus usités. Pour que le lecteur s'y retrouve plus facilement, nous rappelons l'expression de ces différents poids en utilisant la notation du tableau 9.1.

En posant $\lambda_i = \dfrac{w_i}{\sum_i w_i}$, on a :

- le poids proportionnel, à l'inverse des variances,

  $w_i = \dfrac{1}{V_i} = \left[\dfrac{a_{1i}}{n_{1i}^2} + \dfrac{a_{0i}}{n_{0i}^2}\right]^{-1}$, qui conduit à la mesure pondérée DT_V ;

- le poids de Mantel-Haenszel, $w_i = \dfrac{n_{1i} \times n_{0i}}{n_i}$, qui conduit à la mesure pondérée DT_{MH} ;
- le poids proportionnel, à la distribution du facteur F chez les exposés : $w_i = n_{1i}$, qui conduit à la mesure pondérée DT_a ;
- le poids proportionnel, à la distribution du facteur F chez les non-exposés : $w_i = n_{0i}$, qui conduit à la mesure pondérée DT_S.

9.1.4 TESTS STATISTIQUES ET INTERVALLES DE CONFIANCE PAR LA MÉTHODE DU RAPPORT DE VRAISEMBLANCE

Comme on l'a mentionné précédemment, le test du rapport de vraisemblance se prête très bien au jugement tant sur l'homogénéité des mesures que sur la mesure globale d'association. Il suffit de considérer la modélisation linéaire du taux comme fonction de X et de F, $\tau(X, F)$, sans et avec terme d'interaction :

Modèle 1 : $\tau(X,F) = \alpha_1 + \beta_{11}X + \beta_{12}F$

Modèle 2 : $\tau(X,F) = \alpha_2 + \beta_{21}X + \beta_{22}F + \beta_{23}XF$

Dans le premier modèle, le coefficient β_{11} représente la différence des taux ajustée pour le facteur F.

Dans le second modèle, le coefficient β_{23} représente le terme d'interaction additive. Il peut correspondre alors à une mesure de l'hétérogénéité entre les DT_i.

Comme on le verra dans une autre section, la procédure GENMOD de SAS se prête très bien à ce type de modélisation.

Exemple 9.1

Considérons les données suivantes d'une étude portant sur l'association entre un facteur d'exposition X et une maladie Y (tableau 9.2). Les nombres de cas incidents et de personnes-années sont répartis suivant les catégories d'un facteur F dichotomique à contrôler et suivant celles du facteur d'exposition X.

Tableau 9.2

	$F = 0$			$F = 1$		
	$X = 1$	$X = 0$	Total	$X = 1$	$X = 0$	Total
$Y = 1$	6	4	10	24	6	30
Pers-années	1000	2000	3000	2000	1000	3000

Approche en approximation normale

Dans le tableau 9.3 sont présentées les valeurs numériques nécessaires au calcul des tests statistiques, de la mesure DT pondérée et de son intervalle de confiance au niveau désiré.

Tableau 9.3

Strate	DT_i	V_i	w_i	w_iDT_i	$w_iDT_i^2$
$F = 0$	0,004	0,000007	142 857,14	571,43	2,28571
$F = 1$	0,006	0,000012	83 333,33	500	3,0
Total	–	–	226 190,47	1071,43	5,28571

(**PR9.1**)

La mesure pondérée DT est estimée comme $DT = \dfrac{1071,43}{226190,47} = 0,0047.$

La partition du khi-carré nous conduit aux statistiques décrites ci-dessous :

$$\begin{cases} \chi^2_{\text{total}} = 5,2857 \\ \chi^2_{\text{assoc}} = \dfrac{1\,071,43^2}{226\,190,47} = 5,0752\,, & p = 0,0243 \\ \chi^2_{\text{homog}} = 5,2857 - 5,0752 = 0,2105\,, & p = 0,6464 \end{cases}$$

On peut donc conclure à une faible différence entre les deux mesures spécifiques puisque le khi-carré d'homogénéité est de 0,21 avec un degré de liberté, pour une valeur-p de 0,65 environ. Par contre, la mesure pondérée DT est significativement différente de 0 ($p = 0,0243$). Il pourrait être justifié ici de ne présenter que la mesure pondérée : les mesures spécifiques sont homogènes et la mesure globale brute est biaisée ($DT_B = 0,00667$).

Test sur l'homogénéité des mesures DT_i par l'approche de Breslow-Day

En évaluant Δ à 0,0047, on peut facilement estimer pour chaque strate la valeur de π_i et celles de $E(A_{1i} \mid \Delta)$ et de $V(A_{1i} \mid \Delta)$. Le test suivant en découle (tableau 9.4).

Tableau 9.4

Strate	M_{1i}	π_i	$E(A_{1i} \mid \Delta)$	$V(A_{1i} \mid \Delta)$	$\frac{[a_{1i} - E(A_{1i} \mid \Delta)]^2}{V(A_{1i} \mid \Delta)}$
1	10	0,65	6,5	2,278	0,1059
2	30	0,77	23,2	5,282	0,1343
Total	42	–	–	–	0,2402

Le résultat du test Breslow-Day sur l'homogénéité est de 0,24, similaire au résultat du test précédent issu de la partition du khi-carré ($\chi^2_1(\text{homog}) = 0{,}21$].

Intervalle de confiance du *DT* pondéré en approximation normale

Pour chacune des pondérations décrites à la section 9.1.3, nous présentons la mesure pondérée, sa variance et son intervalle de confiance à 95 % (tableau 9.5).

Tableau 9.5

Estimateur	Estimation	Variance	IC à 95 %
DT_V	0,0047	0,000 0044	[0,00062 ; 0,00886]
DT_{MH}	0,0050	0,000 0047	[0,00073 ; 0,00927]
DT_a	0,0053	0,000 0061	[0,00049 ; 0,01018]
DT_S	0,0047	0,000 0044	[0,00053 ; 0,00880]

(PR9.2)

Les résultats varient en fonction des systèmes de poids. Y a-t-il un système de poids préférable aux autres ? Le choix d'un système de poids repose sur différentes considérations. Au plan strictement statistique, le système de poids basé sur l'inverse des variances pourrait être le meilleur choix, puisqu'il conduit à la variance la plus faible. La pondération de Mantel-Haenszel a l'avantage d'être applicable dans les situations où les données sont éparses. Enfin, un système de poids défini à partir d'une distribution observée peut conduire à des mesures qui traduisent plus correctement la réalité.

Test et intervalle de confiance pour le *DT* par le rapport de vraisemblance

On considère les deux modèles spécifiés à la section 9.1.4 que l'on applique aux données du tableau 9.2. On peut les déterminer à l'aide de GENMOD de SAS. (PR9.3)

Dans les résultats, on remarque d'abord une bonne homogénéité entre les mesures DT_i : $\beta_{23} = 0{,}002$ (modèle 2). Le test du rapport de vraisemblance conduit sur l'hypothèse nulle $\beta_{23} = 0$ donne un χ^2 de 0,20 (1 degré de liberté) pour une valeur-*p* de 0,6518. Ce résultat est similaire aux deux précédents : χ^2_1 (homog) = 0,21 et 0,24.

Par ailleurs, le test du rapport de vraisemblance sur le coefficient β_{11} du modèle 1 conduit à un χ^2 de 5,38 pour une valeur-*p* de 0,0203. Ce résultat du test est similaire à celui obtenu dans la partition du khi-carré : χ_1^2 (assoc) = 5,0752. La différence *DT*, correspondant au coefficient β_{11}, est estimée à 4,8 pour 1000 (ou 0,0048) avec un intervalle de confiance à 95 % de [0,0007 ; 0,0095] (ou [0,7 ; 9,5] par 1000 de population). Ces valeurs concordent avec celles obtenues par la méthode des poids proportionnels, à l'inverse des variances et des poids de Mantel-Haenszel.

◆

9.1.5 INTERVALLE DE CONFIANCE D'UNE INTERACTION ADDITIVE

Considérons les données du tableau 9.1 pour lequel *F* est un facteur dichotomique.

On obtient alors le tableau 9.6.

TABLEAU 9.6

	X_1X_2				Total
	11	01	10	00	
$Y = 1$	a_3	a_2	a_1	a_0	m_1
Total	n_3	n_2	n_1	n_0	N

APPROXIMATION NORMALE

On rappelle que l'interaction additive I_+ se mesure par : $I_+ = DT_{11} - DT_{10} - DT_{01}$.

On peut montrer que la variance $V(I_+)$ de I_+ est donnée par :

$$V(I_+) = \sum_{i=0}^{3} \frac{a_i}{n_i^2}.$$

Connaissant la variance de I_+, et utilisant l'approximation normale, il est facile d'en déduire l'intervalle de confiance pour un niveau $100(1-\alpha)\,\%$:

$$(I_+)_{\text{inf}} = I_+ - z_{\alpha/2}\sqrt{\sum_{i=0}^{3}\frac{a_i}{n_i^2}}$$

$$(I_+)_{\text{sup}} = I_+ + z_{\alpha/2}\sqrt{\sum_{i=0}^{3}\frac{a_i}{n_i^2}}$$

Méthode du rapport de vraisemblance

La méthode est celle décrite à la section 8.3.2 du chapitre 8.

Il suffit de résoudre l'équation $2\left[L - L(\beta^*)\right] = \chi^2_{1,1-\alpha}$

où L correspond à la valeur maximale de la fonction du $\log FV(\alpha_2, \beta_{21}, \beta_{22}, \beta_{23})$ et $L(\beta^*)$ à la valeur maximale de la fonction du $\log FV(\alpha_2, \beta_{21}, \beta_{22}, \beta^*)$ pour une valeur fixe β^* de β_{23} (voir modèle 2 de la section 9.1.4).

Exemple 9.2

Reportons-nous aux données du tableau 9.2. Sur ces données, on obtient les mesures suivantes :

$$DT_{11} = \frac{24}{2000} - \frac{4}{2000} = \frac{10}{1000},\quad DT_{10} = \frac{6}{1000} - \frac{4}{2000} = \frac{4}{1000} \text{ et}$$

$$DT_{01} = \frac{6}{1000} - \frac{4}{2000} = \frac{4}{1000}.$$

De ces mesures, on peut facilement obtenir celle de l'interaction additive I_+ :

$$\begin{aligned} I_+ &= DT_{11} - DT_{10} - DT_{01} \\ &= \frac{10}{1000} - \frac{4}{1000} - \frac{4}{1000} \\ &= \frac{2}{1000} \end{aligned}$$

Méthode en approximation normale

La variance de I_+ est estimée par :

$$V(I_+) = \frac{4}{2000^2} + \frac{6}{1000^2} + \frac{6}{1000^2} + \frac{24}{2000^2} = 0{,}000019$$

Les limites de confiance à 95 % sont alors données par :

$$(I_+)_{\text{inf}} = 0,002 - 1,96\sqrt{0,000019}$$
$$= -0,0065$$
$$(I_+)_{\text{sup}} = 0,002 + 1,96\sqrt{0,000019}$$
$$= 0,0105$$

(**PR9.4**)

◆

MÉTHODE DU RAPPORT DE VRAISEMBLANCE

En utilisant GENMOD de SAS, on obtient facilement les limites de confiance du rapport de vraisemblance pour l'interaction I_+. Pour les données du tableau 9.2, on obtient :

$$(I_+)_{\text{inf}} = -0,0073$$
$$(I_+)_{\text{sup}} = 0,0102$$

(**PR9.5**)

◆

9.2 RAPPORT DE DEUX TAUX EN ANALYSE STRATIFIÉE

En référence aux notations de la section 9.1, on définit les rapports de taux spécifiques comme $RT_i = \dfrac{t_{1i}}{t_{0i}}$ et le rapport de taux pondéré comme $RT = \sum_i \lambda_i RT_i$.

9.2.1 TESTS STATISTIQUES EN APPROXIMATION NORMALE SUR LES RAPPORTS DE TAUX

MÉTHODE DE LA PARTITION DU KHI-CARRÉ

Pour juger de l'homogénéité des mesures RT_i et de la signification statistique d'une mesure d'association RT pondérée, nous recourons aussi à la partition du χ^2(total) en χ^2(association) et χ^2(homogénéité). Cependant, cette technique de la partition se pratique, non pas sur la base des RT_i, mais sur la transformation logarithmique de ces mesures.

La transformation de RT par le logarithme présente certains avantages : 1) alors que la distribution d'une statistique comme le RT n'est pas symétrique, celle de sa transformation logarithmique l'est, ce qui confère à la distribution de cette dernière des propriétés de convergence rapide

vers la loi normale ; 2) la transformation logarithmique d'un rapport conduit à des calculs simplifiés sur une échelle additive ; 3) la variance de log (*RT*) se calcule plus facilement que celle du *RT*.

Une fois les calculs faits, il est toujours possible de revenir par transformation inverse à des expressions pour le rapport *RT*.

La variance V_i de $\log(RT_i)$, estimée par la méthode delta, se décrit comme :

$$V\left[\log RT_i\right] = V_i = \frac{1}{a_{1i}} + \frac{1}{a_{0i}}$$

Si on pose $w_i = \frac{1}{V_i}$, on a les relations suivantes :

$$\begin{cases} \chi^2_{\text{total}} = \sum_i w_i \left(\log RT_i\right)^2 \\ \chi^2_{\text{assoc}} = \dfrac{\left[\sum_i w_i \log RT_i\right]^2}{\sum_i w_i} \\ \chi^2_{\text{homog}} = \sum_i w_i \left(\log RT_i - \overline{\log RT}\right)^2 \end{cases}$$

Cette approche définit une mesure pondérée

$\overline{\log RT} = \dfrac{\sum_i w_i \log(RT_i)}{\sum_i w_i}$. Cette mesure pondérée a comme variance :

$$V\left(\overline{\log RT}\right) = \frac{1}{\sum_i w_i}$$

Par transformation inverse, on obtient la mesure RT_V. Cette mesure n'est pas à proprement parler une mesure pondérée des RT_i, mais plutôt une moyenne géométrique de ces mesures :

$$\begin{aligned} e^{\overline{\log RT}} &= RT_V \\ &= \prod_i RT_i^{\lambda_i} \end{aligned}$$

où les λ_i sont proportionnels aux inverses des variances V_i.

Test de Mantel-Haenszel pour l'association

Le test se présente simplement comme $\chi^2_{MH} = \dfrac{\left[\sum_i [a_{1i} - E_0(A_{1i})]\right]^2}{\sum_i V_0(A_{1i})}$.

Sous l'hypothèse H_0, on a $E_0(A_{1i}) = \dfrac{n_{1i} \times m_{1i}}{n_i}$ et $V_0(A_{1i}) = \dfrac{m_{1i} n_{1i} n_{0i}}{n_i^2}$.

9.2.2 Test de Breslow-Day sur l'homogénéité des mesures RT_i

Sous l'hypothèse d'homogénéité, les mesures RT_i fluctuent aléatoirement autour d'une même mesure paramétrique φ. C'est donc dire que les déviations $\left[a_{1i} - E(A_{1i} \mid \varphi)\right]$ fluctuent aléatoirement autour de 0 avec une variance de $V(A_{1i} \mid \varphi)$.

Le test de Breslow-Day se présente alors comme :

$$\chi^2_{k-1} = \sum_{i=1}^{K} \frac{\left(a_{1i} - E(A_{1i} \mid \varphi)\right)^2}{V(A_{1i} \mid \varphi)}$$

Le rapport φ peut être l'estimateur du maximum de vraisemblance ou de Mantel-Haenszel. Sous l'hypothèse des rapports de taux homogènes φ_i, la variable A_{1i} obéit à une loi binomiale de paramètres $\pi_i = \dfrac{\varphi n_{1i}}{\varphi n_{1i} + n_{0i}}$ et m_{1i}. La valeur attendue $E(A_{1i} \mid \varphi)$ et la variance $V(A_{1i} \mid \varphi)$ de A_{1i} sont alors respectivement données par $E(A_{1i} \mid \varphi) = m_{1i}\pi_i$ et $V(A_{1i} \mid \varphi) = m_{1i}\pi_i(1 - \pi_i)$.

9.2.3 Intervalle de confiance en approximation normale de *RT* pondéré

Dans le cadre d'une partition du khi-carré

Rappelons que la mesure RT_V est une moyenne géométrique des différentes mesures spécifiques RT_i (voir la section 9.2.1). De même que RT_V est la transformation inverse de la somme pondérée des $\log(RT_i)$, de même ses limites de confiance s'obtiennent aussi par transformation inverse des limites de confiance de cette même somme pondérée :

Ainsi :

$$\varphi_{\text{inf}} = RT_V \times e^{-z_{\alpha/2}\sqrt{\frac{1}{\sum_i w_i}}}$$

$$\varphi_{\text{sup}} = RT_V \times e^{+z_{\alpha/2}\sqrt{\frac{1}{\sum_i w_i}}}$$

où $w_i = \dfrac{1}{V_i}$.

Dans le cadre d'une pondération arithmétique

On s'intéresse à l'intervalle de confiance d'un rapport du taux pondéré *RT* tel que $RT = \sum_i \lambda_i RT_i$, où les λ_i constituent un système de poids. Par ailleurs, suivant la méthode delta, la variance de log(*RT*) peut être correctement estimée par : $V\left[\log\left(\sum_i \lambda_i RT_i\right)\right] = \sum_i \lambda_i^2 V_i$. On déduit alors les limites de confiance de *RT* par :

$$RT_{\text{inf}} = RT \times e^{-z_{\alpha/2}\sqrt{\sum_i \lambda_i^2 V_i}}$$

$$RT_{\text{sup}} = RT \times e^{+z_{\alpha/2}\sqrt{\sum_i \lambda_i^2 V_i}}$$

où $V_i = \dfrac{1}{a_{1i}} + \dfrac{1}{a_{0i}}$.

Rappelons que le système de poids $\{\lambda_i\}$ est théoriquement arbitraire.

Comme pour la mesure *DT*, nous proposons différents systèmes de poids pour la mesure *RT* ; ces poids sont parmi les plus usités.

Pour que le lecteur s'y retrouve plus facilement, nous rappelons l'expression de ces différents poids en utilisant la notation du tableau 9.1.

En posant $\lambda_i = \dfrac{w_i}{\sum_i w_i}$, on a :

- le poids de Mantel-Haenszel, $w_i = \dfrac{a_{0i} \times n_{1i}}{n_i}$, qui conduit à la mesure pondérée RT_{MH};

- le poids de type *SMR*, qui conduit à la mesure pondérée RT_a ;
- le poids standardisé, proportionnel à la distribution du facteur F chez les cas non exposés, qui conduit à la mesure pondérée RT_s.

9.2.4 TEST STATISTIQUE ET INTERVALLE DE CONFIANCE PAR LA MÉTHODE DU RAPPORT DE VRAISEMBLANCE

Dans le cadre de la méthode du rapport de vraisemblance, le rapport de taux est traité dans le cadre de la loi de Poisson à l'aide de la transformation logarithmique des taux. Si $\tau(X, F)$ désigne le taux comme une fonction des variables indépendantes X et F, alors les modèles considérés peuvent avoir la forme suivante :

$$\textbf{Modèle 1 :}\ \log\ \tau(X,F) = \alpha_1 + \beta_{11}X + \beta_{12}F$$

$$\textbf{Modèle 2 :}\ \log\ \tau(X,F) = \alpha_2 + \beta_{21}X + \beta_{22}F + \beta_{23}XF$$

Dans le second modèle, le coefficient β_{23} représente le logarithme de l'interaction multiplicative entre X et F ; il marque l'hétérogénéité des RT_i à travers les strates de F. Si X et F sont dichotomiques, alors $e^{\beta_{23}} = \dfrac{RT_1}{RT_0}$.

On peut écrire aussi $e^{\beta_{23}} = \dfrac{RT_{11}}{RT_{10} \times RT_{01}}$, où RT_{ij} représente le rapport de taux issu de la comparaison entre la catégorie ($X = i$, $F = j$) et la catégorie de référence ($X = 0$, $F = 0$).

Dans le premier modèle, le coefficient β_{11} correspond au logarithme du rapport des taux, ajusté pour le facteur F : $e^{\beta_{11}} = RT_{RV}$. Ainsi, pour porter un jugement sur la signification statistique du rapport de taux ajusté pour le facteur F, il suffit d'appliquer le test du rapport de vraisemblance au coefficient β_{11} du modèle 1.

On obtient les intervalles de confiance en solutionnant l'équation de vraisemblance : $\left[L - L(\beta)\right] - 0,5\chi^2_{1,1-\alpha} = 0$.

EXEMPLE 9.3

TESTS STATISTIQUES SUR LE *RT* EN APPROXIMATION NORMALE

Une étude a été conduite sur la relation entre l'exposition au facteur X et la maladie Y. Les résultats stratifiés pour l'âge sont donnés au tableau 9.7.

TABLEAU 9.7

	Âge en années			
	20-49		50-79	
	$X = 1$	$X = 0$	$X = 1$	$X = 0$
Cas	78	39	783	52
Personnes-années	12 546	39 853	22 314	11 221

Le tableau 9.8 présente les valeurs numériques nécessaires aux calculs des tests.

TABLEAU 9.8

Âge (années)	RT_i	$\log(RT_i)$	V_i	w_i	$w_i \log(RT_i)$	$w_i[\log RT_i]^2$
20-49	6,35	1,85	0,0385	26,00	48,0725	88,88
50-79	7,57	2,02	0,0205	48,76	98,7161	199,85
Total	–	–	–	74,76	146,7886	288,73

D'abord, on établit que $\log RT = \dfrac{146,79}{74,76} = 1,9635$ et qu'en conséquence le RT est donné par : $RT = e^{1,9635} = 7,12$.

La partition du khi-carré nous conduit aux statistiques décrites ci-dessous.

$$\begin{cases} \chi^2_{total} = 288,73 \\ \chi^2_{assoc} = \dfrac{146,79^2}{74,76} = 288,21 \quad p \approx 0 \\ \chi^2_{\text{hom}og} = 0,52 \qquad\qquad p = 0,4698 \end{cases}$$

On peut donc conclure à une faible différence entre les deux mesures spécifiques puisque le khi-carré d'homogénéité est de 0,52, qui pour 1 degré de liberté donne une valeur-p de 0,47 environ. Le test de Breslow-Day conduit sensiblement au même résultat (khi-carré = 0,523). Par contre, la mesure globale log(RT) est significativement différente de 0, puisque le khi-carré d'association est de 288,21. Sans discuter du contexte de l'étude, il pourrait être justifié ici de ne présenter que la mesure globale RT_V correspondante.

La valeur du test de Mantel-Haenszel est de 371,31.

TESTS STATISTIQUES SUR LE *RT* PAR LA MÉTHODE DU RAPPORT DE VRAISEMBLANCE

On applique le test aux données du tableau 9.7 à l'aide de la procédure GENMOD, suivant le schéma des modèles présentés à la section 9.2.4. Le facteur X correspond à l'exposition et F au facteur âge. Les résultats nous conduisent aux constatations suivantes.

On remarque d'abord une faible hétérogénéité : $\beta_{23} \approx 0{,}1755$; le test du rapport de vraisemblance conduit à un χ^2 de 0,5210 pour une valeur-*p* de 0,4704. Par ailleurs, le coefficient β_{11} du modèle 1 est égal à 1,9652 et le test du rapport de vraisemblance sur ce coefficient conduit à un χ^2 de 457,28 pour un $p \approx 0$, ce qui suggère un rapport de taux *RT* fortement significatif. Ce rapport est estimé à $e^{1,9652} = 7{,}1363$. Ces résultats sont similaires à ceux obtenus précédemment par la méthode des poids proportionnels à l'inverse des variances, sauf la valeur du χ^2 d'association (288,21 par la première méthode et 457,28 par la méthode du rapport de vraisemblance).

Enfin, remarquons que $e^{\beta_{23}} = e^{0,1755} = \dfrac{7,5720}{6,3531} = 1{,}1919$.

Nous résumons au tableau 9.9 les résultats de ces différents tests.

Tableau 9.9

Méthode	Test d'association sur le *RT* pondéré			Test d'homogénéité sur les RT_i			Programme
	χ^2	ddl*	Valeur-*p*	χ^2	ddl*	Valeur-*p*	
Partition du khi-carré	288,21	1	≈ 0	0,52	1	0,47	PR9.6
MH	371,31	1	≈ 0	–	–	–	PR9.7
Breslow-Day	–	–	–	0,52	1	0,47	PR9.8
RV	457,28	1	≈ 0	0,52	1	0,47	PR9.9

* Nombre de degrés de liberté. ◆

Exemple 9.4

Intervalle de confiance du *RT* pondéré suivant les différentes méthodes de calcul

Pour les données du tableau 9.7, nous présentons pour chacun des systèmes de poids décrits à la section 9.2.3 et pour la méthode du *RV* la mesure globale *RT*, son logarithme et la variance du logarithme, et l'intervalle de confiance à 95 % du *RT*.

Tableau 9.10

RT	Valeur du *RT*	Log *RT*	*V*[log *RT*]	IC à 95 %	Programme
RT_V	7,12366	1,9634	0,013376	[5,68 ; 8,94]	PR9.10
RT_{MH}	7,31298	1,9872	0,014455	[5,78 ; 9,26]	
RT_a	7,44267	2,0058	0,016819	[5,77 ; 9,60]	PR9.11
RT_S	7,04963	1,9492	0,013761	[5,60 ; 8,87]	
RT_{RV}	7,13634	1,9652	0,013110	[5,73 ; 8,99]	PR9.12

Nous remarquons que les *RT* obtenus diffèrent peu d'un système de poids à l'autre. Cela est dû principalement à la grande homogénéité entre les deux mesures spécifiques. La mesure *RT* ajustée, quel que soit le système de poids, doit obligatoirement se situer entre ces deux valeurs spécifiques, $RT_1 = 6{,}3531$ et $RT_2 = 7{,}5720$.

◆

9.2.5 INTERVALLE DE CONFIANCE D'UNE INTERACTION MULTIPLICATIVE

Considérons les données du tableau 9.6, que nous reprenons dans le tableau 9.11.

TABLEAU 9.11

	X_1X_2				Total
	11	01	10	00	
$Y = 1$	a_3	a_2	a_1	a_0	m_1
Total	n_3	n_2	n_1	n_0	n

APPROXIMATION NORMALE

Rappelons que l'interaction multiplicative $I_\times$ se mesure par :

$$I_\times = \frac{RT_{11}}{RT_{10} \times RT_{01}}$$

On peut montrer que la variance $V[\log(I_\times)]$ est donnée par :

$$V\left[\log(I_\times)\right] = \sum_{i=0}^{3} \frac{1}{a_i}$$

Connaissant la variance de $\log(I_\times)$, et en utilisant l'approximation normale, il est facile d'en déduire l'intervalle de confiance pour un niveau $100(1 - \alpha)\,\%$:

$$(I_\times)_{\text{inf}} = \left(I_\times\right) \times \exp\left(-z_{\alpha/2} \sqrt{\sum_{i=0}^{3} \frac{1}{a_i}} \right)$$

$$(I_\times)_{\text{sup}} = \left(I_\times\right) \times \exp\left(+z_{\alpha/2} \sqrt{\sum_{i=0}^{3} \frac{1}{a_i}} \right)$$

Méthode du rapport de vraisemblance

Il suffit de résoudre l'équation $2\left[L - L(\beta^*)\right] = \chi^2_{1,1-\alpha}$

où L correspond à la valeur maximale de la fonction du $\log FV(\alpha_2, \beta_{21}, \beta_{22}, \beta_{23})$ et $L(\beta^*)$ à la valeur maximale de la fonction du $\log FV(\alpha_2, \beta_{21}, \beta_{22}, \beta^*)$ pour une valeur fixe β^* de β_{23} (voir modèle 2 de la section 9.2.4).

Exemple 9.5

Reportons-nous aux données du tableau 9.7. Supposons que le groupe d'âge 20-49 constitue la référence pour la variable âge. Alors sur ces données, on obtient les mesures suivantes :

$$RT_{11} = \frac{783 \times 39853}{39 \times 22314} = 35,86\,,\ RT_{10} = \frac{78 \times 39853}{39 \times 12546} = 6,35 \text{ et}$$
$$RT_{01} = \frac{52 \times 39853}{39 \times 11221} = 4,74.$$

De ces mesures, on peut facilement obtenir celle de l'interaction multiplicative $I_\times$:

$$\begin{aligned} I_\times &= \frac{RT_{11}}{RT_{10} \times RT_{01}} \\ &= \frac{35,86}{6,35 \times 4,74} \\ &= 1,19 \end{aligned}$$

Méthode en approximation normale

La variance de $I_\times$ est estimée par : $V\left(\log I_\times\right) = \frac{1}{39} + \frac{1}{78} + \frac{1}{52} + \frac{1}{783} = 0,05897.$

Les limites de confiance à 95 % sont alors données par :

$$\begin{aligned} (I_\times)_{\text{inf}} &= 1,19 \times \exp\left(-1,96\sqrt{0,05897}\right) \\ &= 0,74 \\ (I_\times)_{\text{sup}} &= 1,19 \times \exp\left(+1,96\sqrt{0,05897}\right) \\ &= 1,92 \end{aligned}$$

(**Pr9.13**)

MÉTHODE DU RAPPORT DE VRAISEMBLANCE

En utilisant GENMOD de SAS, on obtient les limites de confiance du rapport de vraisemblance pour l'interaction $I_{\times}$. Pour les données du tableau 9.7, ces limites de confiance sont sensiblement les mêmes que celles obtenues par approximation normale :

$$(I_{\times})_{\text{inf}} = 0,74$$
$$(I_{\times})_{\text{sup}} = 1,92$$

(PR9.14)

◆

9.3 MESURE DU *SMR* POUR LES TAUX EN ANALYSE STRATIFIÉE

Dans la section 5.4 du chapitre 5, nous avons considéré le *SMR* dans sa plus simple expression. Les propriétés de la loi de Poisson vont permettre ici d'étendre assez facilement aux analyses stratifiées les principaux outils statistiques définis pour les analyses simples.

On considère une variable de stratification F comportant k catégories. Pour chaque strate i de la variable F, on désigne par n_i et τ_{0i} respectivement les personnes-temps observées sur l'échantillon et le taux estimé à partir de la population standard. La valeur attendue de cas X_{0i} correspond à $n_i\tau_{0i}$. Si, sur cette strate, le nombre de cas observés est de x_i, alors on a :

$$SMR_i = \frac{x_i}{X_{0i}}$$

On peut aussi considérer la définition théorique suivante. Pour la strate i, le nombre de cas observés x_i est la réalisation d'une variable de Poisson X_i de paramètre $n_i\tau_i$ où τ_i est un taux inconnu. La valeur x_i est la meilleure estimation disponible pour $n_i\tau_i$. Le SMR_i correspond alors simplement à l'estimation du rapport φ_i des taux τ_i et τ_{0i} :

$$SMR_i = \frac{n_i\tau_i}{n_{0i}\tau_{0i}} = \frac{\tau_i}{\tau_{0i}} = \varphi_i$$

La mesure globale du *SMR* peut être définie comme :

$$SMR = \frac{\sum_{i=1}^{k} X_{0i} SMR_i}{\sum_{i=1}^{k} X_{0i}} = \frac{\sum_{i=1}^{k} x_i}{\sum_{i=1}^{k} X_{0i}} = \frac{x}{X_0}$$

où x désigne le nombre total de cas observés et X_0 le nombre total de cas attendus sous l'hypothèse nulle : $X_0 = \sum_i X_{0i}$. On remarque que le *SMR* est la somme des SMR_i spécifiques pondérés par les valeurs attendues X_{0i}. Cet estimateur du *SMR* est celui du maximum de vraisemblance (voir ci-dessous).

On rappelle que si $\{X_i\}$ désigne k variables de Poisson indépendantes de paramètres $\{n_i\tau_i\}$, alors $X = \Sigma_i X_i$ est aussi une variable de Poisson de paramètre $E(X) = V(X) = \Sigma_i n_i\tau_i$.

La fonction de vraisemblance construite pour les observations sur les variables X_i prend la forme suivante :

$FV\left(\{\varphi_i\}\right) = \prod_{i=1}^{k} FV(\varphi_i) = \prod_{i=1}^{k} \frac{e^{-\varphi_i X_{0i}} \left(\varphi_i X_{0i}\right)^{x_i}}{x_i!}$. Par la transformation logarithmique, elle devient :

$$L\left(\{\varphi_i\}\right) = \sum_{i=1}^{k} L(\varphi_i) = \sum_{i=1}^{k} \left[-\varphi_i X_{0i} + x_i \log(\varphi_i X_{0i}) + K_i\right]$$

où les K_i sont des constantes qui ne comportent pas d'information pertinente à l'estimation des φ_i.

Pour une certaine somme pondérée φ des φ_i spécifiques, on pose $FV(\varphi) = \prod_{i=1}^{k} FV(\varphi_i = \varphi)$. On a donc $L(\varphi) = \sum_{i=1}^{k} L(\varphi_i = \varphi)$. De là, on peut montrer que $\frac{x}{X_0}$ est l'estimateur du maximum de vraisemblance de φ.

Par les propriétés de la loi de Poisson, les tests statistiques et les intervalles de confiance sur le *SMR* sont analogues à ceux que nous avons déjà décrits dans le chapitre 5, aux sections 5.4.1 et 5.4.2, pour une variable de Poisson. Après un bref rappel de ces tests et intervalles de confiance, nous décrirons certains exemples numériques.

9.3.1 TESTS STATISTIQUES SUR LE *SMR*

TEST EXACT

Si on suppose $x > X_0$, le test exact unilatéral à droite sous H_0 est de la forme $p = P(X \geq x \mid \mu)$ où $\mu = X_0$. Le test exact bilatéral est de la forme $p = P(X \leq u \mid \mu) + P(X \geq x \mid \mu)$ où u est la plus grande valeur de X inférieure à X_0 telle que $P(X = u \mid \mu) \leq P(X = x \mid \mu)$.

De façon analogue, on peut définir le test unilatéral à gauche et le test bilatéral correspondant.

En corrigeant pour la valeur mi-*p*, on obtient la valeur-*p* recherchée.

Tests en approximation normale

Simple

On se rappelle simplement que $\frac{(x - X_0)^2}{X_0} = \chi_1^2$.

Test basé sur log(SMR)

Ce test est simplement défini par : $\chi_1^2 = \frac{(\log SMR)^2}{V(\log SMR)} = x(\log SMR)^2$

Test du rapport de vraisemblance

Dans ce qui suit, nous considérons trois modèles (ou hypothèses) : le modèle de base $\varphi = 1$, le modèle spécifié de φ, et le modèle qui sature les données. En chacun de ces modèles, nous évaluons la fonction de vraisemblance $L(\varphi)$. La comparaison de ces valeurs permet de définir les tests du rapport de vraisemblance pour la signification du *SMR* global et pour l'homogénéité des SMR_i spécifiques. Ces fonctions sont :

MODÈLE DE BASE $\varphi = 1$

Ce modèle correspond à l'hypothèse nulle d'un $SMR = 1$. On rappelle que sous l'hypothèse nulle $\phi = 1$, on a $E(X) = V(X) = X_0$.

Sous cette hypothèse, la fonction de vraisemblance prend la valeur L_0.

$$L_0 = -X_0 + \sum_{i=1}^{k} x_i \log(X_{0i})$$

MODÈLE DE φ (DU MAXIMUM DE VRAISEMBLANCE)

Ce modèle correspond à l'hypothèse du maximum de vraisemblance. Sous cette hypothèse, $\varphi = \frac{x}{X_0}$ qui, on le rappelle, est l'estimateur du maximum de vraisemblance. Sous cette hypothèse, la fonction de vraisemblance prend la valeur L_1.

$$L_1 = -\varphi X_0 + \sum_{i=1}^{k} x_i \log(\varphi X_{0i})$$

MODÈLE QUI SATURE LES DONNÉES

Le modèle qui sature les données est celui obtenu en remplaçant chaque φ_i de la fonction L par son estimation $\frac{x_i}{X_{0i}}$, qui est celle du maximum de vraisemblance.

$$L_S = -x + \sum_{i=1}^{k} x_i \log(x_i)$$

Le test du rapport de vraisemblance peut être considéré comme émanant d'une partition de la déviance totale à expliquer sur la base du modèle $\varphi = 1$.

La déviance totale est mesurée par la différence $[L_S - L_0]$. Elle peut faire l'objet d'une partion suivant deux composantes : la déviance expliquée $[L_1 - L_0]$ par le modèle $\varphi = SMR$ (estimateur du maximum de vraisemblance) et la déviance résiduelle $[L_S - L_1]$ marquée par la différence entre le modèle $\varphi = SMR$ et le modèle saturé ($\varphi_i = SMR_i$).

Nous présentons cette partition, et les tests qui en résultent, à l'aide du tableau de la déviance (tableau 9.12).

TABLEAU 9.12

Composante	Modèle	Degrés de liberté	Déviance	Test du khi-carré
Totale	$\varphi = 1$	k*	$[L_S - L_0]$	$2\left[-x + X_0 + \sum_{i=1}^{k} x_i \log\left(\frac{x_i}{X_{0i}}\right)\right]$
Expliquée	$\varphi = SMR$	1	$[L_1 - L_0]$	$2\left[-\varphi X_0 + X_0 + x \log \varphi\right]$
Résiduelle	–	$k - 1$	$[L_S - L_1]$	$2\sum_{i=1}^{k} x_i \log\left(\frac{x_i}{\varphi X_{0i}}\right)$

* Nombre de strates impliquées dans la définition du *SMR*.

Le test sur le modèle $\varphi = SMR$ est le test d'association qui compare la valeur du *SMR* observé à la valeur 1 (sous l'hypothèse nulle $\varphi = 1$) :

$$\chi^2_1 = 2\left[-x + X_0 + \sum_{i=1}^{k} x_i \log\left(\frac{x_i}{X_{0i}}\right)\right]$$

Le test sur la déviance résiduelle permet de porter un jugement sur l'homogénéité des SMR_i spécifiques, comme on le précise plus loin :

$$\chi^2_{k-1} = 2\sum_{i=1}^{k} x_i \log\left(\frac{x_i}{\varphi X_{0i}}\right)$$

Exemple 9.6

Dans une région donnée, sur une période d'un an, on observe 114 cas incidents de la maladie *Y* chez les personnes âgées de 40 ans et plus. Le tableau 9.13 décrit la répartition des cas (x_i) de maladie *Y* observée dans la région *R*, les effectifs n_i (en personnes-années) de cette même région, et les taux d'incidence (τ_i) pour la population générale dans la même année d'observation, suivant différents groupes d'âges. On y ajoute aussi les valeurs attendues et les *SMR*.

Tableau 9.13

Age	x_i	n_i ($\times 10^3$)	τ_i ($\times 10^{-3}$)	X_{0i}	SMR_i
40-49	15	120	0,1	12	1,25
50-59	20	75	0,2	15	1,33
60-69	54	50	0,5	25	2,16
70 et +	25	10	1,0	10	2,50
Total	114	255	0,51	62	1,84

À partir de ces données, peut-on dire qu'il y a un excès de cas incidents dans la région *R* ?

En d'autres termes, le nombre observé ($X = 114$) de cas est-il significativement supérieur à celui attendu ($X_0 = 62$) sous l'hypothèse que, tenant compte de l'âge, cette région soit affectée des mêmes taux d'incidence de la maladie *Y* que ceux de la population générale ?

Nous considérons alors le *SMR* global :

$$SMR = \frac{114}{62} = 1,84$$

Le nombre de cas total *X* obéit à une loi de Poisson de paramètre $X_0 = 62$.

Test exact

Le test exact bilatéral est de la forme :

$$p = P(X \leq 20 \mid \mu = 62) + P(X \geq 114 \mid \mu = 62).$$

La valeur 20 correspond à la plus grande valeur de *X* dans la partie gauche de la distribution telle que $P(X = x \mid \mu = 62) \leq P(X = 114 \mid \mu = 62)$.

Dans la convention mi-*p*, la valeur-*p* obtenue est de 0,000000021507.

Tests en approximation normale

Simple

En appliquant ce test aux données du tableau 9.13, on a :

$\chi_1^2 = \dfrac{(114-62)^2}{62} = 43,61$, ce qui donne une valeur-$p \approx 0$.

Test basé sur log(*SMR*)

En appliquant ce test aux données du tableau 9.13, on a

$$\chi_1^2 = 114\left[\log\frac{114}{62}\right]^2 = 42,29 \text{ ce qui donne aussi une valeur-}p \approx 0.$$

Test du rapport de vraisemblance

En appliquant ce test aux données du tableau 9.13, on obtient :

$$\chi_1^2 = 2\left[-114+62+114\log\frac{114}{62}\right] = 34,87 \text{ pour une valeur-}p \approx 0.$$

Nous résumons dans le tableau 9.14 les résultats des différents tests conduits sur les données du tableau 9.13.

Tableau 9.14

Test	Khi-carré	*p* (bilatéral)	Programme
Exact	–	$2,15 \times 10^{-9}$	PR9.15
Normal simple	43,61	$4,00 \times 10^{-11}$	PR9.16
Log(*SMR*)	42,29	$7,87 \times 10^{-11}$	PR9.17
RV	34,87	$3,53 \times 10^{-9}$	PR9.18

Les résultats des tests exacts et du RV sont très similaires. Ceux qui utilisent plus directement la loi normale (normal simple et basé sur log(*SMR*)) se distinguent des deux autres.

Soulignons cependant que les quatre tests conduisent en pratique à une même valeur-*p*, soit approximativement 0.

◆

9.3.2 Intervalle de confiance pour le *SMR*

Intervalle de confiance exact

Les limites de l'intervalle de confiance de φ sont définies comme suit : la limite inférieure φ_{inf} de l'intervalle est la valeur de φ qui satisfait l'équation

$$\sum_{i=x}^{\infty} \frac{e^{-\varphi X_0}(\varphi X_0)^i}{i!} = \alpha/2$$

de façon analogue, on définit la limite supérieure φ_{sup} comme la valeur de φ qui satisfait l'équation :

$$\sum_{i=0}^{x} \frac{e^{-\varphi X_0}(\varphi X_0)^i}{i!} = \alpha/2$$

On applique la convention mi-*p*.

Intervalle de confiance par approximation normale

Basée sur la transformation RAC

Formellement, les limites de confiance du *SMR* se décrivent comme :

$$\phi_{\text{inf}} = \frac{\left(\sqrt{x} - z_{\alpha/2} \times 0,50\right)^2}{X_0}$$

$$\phi_{\text{sup}} = \frac{\left(\sqrt{x} + z_{\alpha/2} \times 0,50\right)^2}{X_0}$$

(Nous suggérons d'éviter l'intervalle de confiance calculé à partir d'une simple application de la loi normale à la variable *X* de Poisson.)

Basée sur log(SMR)

On peut montrer que $V[\log(SMR)] = \frac{1}{x}$. De là, on déduit que :

$$\varphi_{\text{inf}} = SMR \times e^{-z_{\alpha/2}\sqrt{1/x}}$$

$$\varphi_{\text{sup}} = SMR \times e^{+z_{\alpha/2}\sqrt{1/x}}$$

Intervalle de confiance par la méthode du rapport de vraisemblance

La résolution pour φ de l'équation de vraisemblance

$$2\left\{(x\log x - x) - \left[x\log\left(\varphi X_0\right) - \varphi X_0\right]\right\} - \chi^2_{1,1-\alpha} = 0$$

conduit aux limites de confiance à 100(1 – α) % pour le *SMR*.

Exemple 9.7

Revenons aux données du tableau 9.13. Nous rappelons que le *SMR* est égal à 1,838 7.

Intervalle de confiance exact à 95 % du *SMR*

Les limites de confiance de φ (le *SMR*) sont calculées comme :

φ_{inf} est la solution pour φ de l'équation $\sum_{i=114}^{\infty} \frac{e^{-62\varphi}(62\varphi)^i}{i!} = \alpha/2$

et

φ_{sup} est la solution pour φ de l'équation $\sum_{i=0}^{114} \frac{e^{-62\varphi}(62\varphi)^i}{i!} = \alpha/2$.

On obtient les limites suivantes :

$\varphi_{inf} = 1{,}52$
$\varphi_{sup} = 2{,}20$

Intervalle de confiance à 95 % en approximation normale

Par transformation RAC

Les limites de l'intervalle de confiance de φ (le *SMR*) se décrivent comme :

$$\varphi_{inf} = \frac{\left(\sqrt{114} - 1{,}96 \times 0{,}50\right)^2}{62}$$
$$= 1{,}52$$
$$\varphi_{sup} = \frac{\left(\sqrt{114} + 1{,}96 \times 0{,}50\right)^2}{62}$$
$$= 2{,}19$$

Basé sur log(SMR)

Les limites de l'intervalle de confiance de φ (le *SMR*) se décrivent comme :

$$\varphi_{inf} = 1{,}8387 \times e^{-1{,}96\sqrt{1/114}}$$
$$= 1{,}5303$$
$$\varphi_{sup} = 1{,}8387 \times e^{+1{,}96\sqrt{1/114}}$$
$$= 2{,}2092$$

Intervalle de confiance à 95 % par la méthode du rapport de vraisemblance

En solutionnant pour φ l'équation de vraisemblance

$$2\left\{(114\log 114 - 114) - \left[114\log(62\varphi) - 62\varphi\right]\right\} - 3{,}84 = 0$$

on obtient les limites suivantes :

$\varphi_{inf} = 1{,}52$
$\varphi_{sup} = 2{,}20$

Nous reprenons dans le tableau 9.15 les résultats des différentes méthodes utilisées pour l'exemple.

TABLEAU 9.15

Méthode	*SMR*	IC 95 %	Programme
Exacte	1,8387	[1,5237 – 2,2004]	PR9.19
Transformation RAC	1,8387	[1,5167 – 2,1917]	PR9.20
Log(*SMR*)	1,8387	[1,5304 – 2,2092]	PR9.21
RV	1,8387	[1,5215 – 2,1972]	

On peut dire que les quatre méthodes conduisent essentiellement aux mêmes limites. La méthode exacte et celle du *RV* conduisent à des résultats très similaires.

◆

9.3.3 TESTS SUR L'HOMOGÉNÉITÉ DES *SMR* SPÉCIFIQUES EN ANALYSE STRATIFIÉE

Sans aborder spécifiquement la comparabilité entre deux ou plusieurs *SMR*, nous présentons dans un contexte simplifié deux tests statistiques qui permettent de porter un jugement sur l'homogénéité entre deux ou plusieurs *SMR*.

Ainsi, on suppose que F est un facteur pour lequel on pratique la stratification. Pour chaque strate i de F, on a l'estimation d'un SMR_i : $SMR_i = \dfrac{x_i}{X_{0i}}$. La situation se présente alors comme :

TABLEAU 9.16

	F				Total
	1	2	….	*k*	
Observé (a_i)	x_1	x_2	….	x_k	x
Attendu (A_{0i})	X_{01}	X_{02}	….	X_{0k}	X_0
SMR_i	SMR_1	SMR_2	….	SMR_k	*SMR*

L'hypothèse à tester est la suivante :

$$H_0 : SMR_1 = SMR_2 = \ldots. = SMR_k = SMR = \varphi, \text{ où } \varphi = \frac{x}{X_0}.$$

Test en approximation normale

Le test d'homogénéité peut être défini simplement à partir des variables de Poisson X_i sous l'hypothèse d'un *SMR* égal à φ. Les variables de Poisson X_i ont alors les valeurs attendues $E(X_i) = \varphi X_{0i}$. Le test se présente alors simplement comme :

$$\chi^2_{k-1} = \sum_{i=1}^{k} \frac{\left(x_i - \varphi X_{0i}\right)^2}{\varphi X_{0i}} = \frac{X_0}{x}\left[\sum_{i=1}^{k} \frac{x_i^2}{X_{0i}} - \frac{x^2}{X_0}\right].$$

C'est un test analogue à celui de Breslow-Day déjà présenté pour l'homogénéité des mesures, au chapitre 8, à la section 8.2.4.

Test du rapport de vraisemblance

Le test du rapport de vraisemblance pour l'homogénéité des SMR_i est décrit au tableau 9.12 comme la composante résiduelle de la partition de la déviance totale :

$$\chi^2_{k-1} = 2\sum_{i=1}^{k} x_i \log\left(\frac{x_i}{\varphi X_{0i}}\right)$$

Exemple 9.8

Test en approximation normale

Considérons les données du tableau 9.13. Alors, le test en approximation normale se décrit comme :

$$\chi^2_{k-1} = \frac{X_0}{x}\left(\sum_{i=1}^{k} \frac{x_i^2}{X_{0i}} - \frac{x^2}{X_0}\right) = \frac{62}{114}\left(\left[\frac{15^2}{12} + \frac{20^2}{15} + \frac{54^2}{25} + \frac{25^2}{10}\right] - \frac{114^2}{62}\right) = 8{,}12731.$$

La valeur-*p* correspondante est de 0,043453. Ces résultats sont peu compatibles avec l'hypothèse d'homogénéité des *SMR*. (**PR9.22**)

TEST DU RAPPORT DE VRAISEMBLANCE

Considérons les données du tableau 9.13. Alors, le test du rapport de vraisemblance s'applique comme :

$$\begin{aligned}\chi^2_{k-1} &= 2\sum_{i=1}^{k} x_i \log\left(\frac{x_i}{\varphi X_{0i}}\right)\\ &= 2\left[\begin{array}{l} 15\log\dfrac{15}{1,839\times 12} + 20\log\dfrac{20}{1,839\times 15} + \\ 54\log\dfrac{54}{1,839\times 25} + 25\log\dfrac{25}{1,839\times 10}\end{array}\right]\\ &= 2\times 4,1606 = 8,3212\end{aligned}$$

La valeur-*p* correspondante est de 0,0398. (**PR9.23**)

Remarquons que ces deux tests sont ici très concordants.

◆

CHAPITRE

10

MESURES D'ASSOCIATION BASÉES SUR LES PROPORTIONS EN ANALYSE STRATIFIÉE

Les mesures d'association qui sont décrites dans ce chapitre sont les différences de proportions *DP*, les rapports de proportions *RP*, les rapports de cotes *RC* et le *SMR*. Pour chacune de ces mesures, nous présentons d'abord les tests statistiques, puis les intervalles de confiance, en approximation normale, puis par la méthode du rapport de vraisemblance. Les différentes procédures appliquées aux proportions sont analogues à celles déjà présentées pour les taux au chapitre 9.

10.1 DIFFÉRENCE ENTRE DEUX PROPORTIONS EN ANALYSE STRATIFIÉE

On considère les données d'une étude de cohortes (ou de prévalence) portant sur l'association entre le facteur X et la maladie Y.

Le facteur F est un facteur à contrôler. Ce facteur F pour lequel on veut pratiquer une analyse stratifiée a k catégories. Dans le tableau 10.1 sont décrites les données de l'étude pour la strate i du facteur F.

TABLEAU 10.1

Facteur F		$X = 1$	$X = 0$	Total
Strate i	$Y = 1$	a_{1i}	a_{0i}	m_{1i}
	$Y = 0$	b_{1i}	b_{0i}	m_{0i}
	Total (personnes)	n_{1i}	n_{0i}	n_i

Sur cette strate, les proportions p_{1i} et p_{0i} sont respectivement définies comme : $p_{1i} = \dfrac{a_{1i}}{n_{1i}}$ et $p_{0i} = \dfrac{a_{0i}}{n_{0i}}$.

La différence DP_i correspondante entre les proportions est alors : $DP_i = p_{1i} - p_{0i}$.

Toute mesure pondérée DP, résumant les mesures spécifiques DP_i, se présente comme $DP = \sum_i \lambda_i DP_i$, où les λ_i constituent un système de poids.

10.1.1 TESTS STATISTIQUES EN APPROXIMATION NORMALE

Pour répondre au problème de l'homogénéité des mesures spécifiques DP_i et de la signification statistique de la mesure pondérée DP, nous faisons appel à la partition du $\chi^2_k(\text{total})$ en $\chi^2_1(\text{assoc})$ et $\chi^2_{k-1}(\text{homog})$. En choisissant les poids λ_i proportionnels à l'inverse des variances V_i des mesures DP_i, nous obtenons simplement les statistiques :

$$\begin{cases} \chi^2_{\text{total}} = \sum_i w_i DP_i^2 \\ \chi^2_{\text{assoc}} = \dfrac{\left(\sum_i w_i DP_i\right)^2}{\sum_i w_i} \\ \chi^2_{\text{homog}} = \sum_i w_i \left(DP_i - DP\right)^2 \end{cases}$$

où $w_i = 1 \Big/ V_i$. La variance V_i de la mesure DP_i est donnée par : $V_i = \dfrac{a_{1i} b_{1i}}{n_{1i}^3} + \dfrac{a_{0i} b_{0i}}{n_{0i}^3}$.

Ainsi, le khi-carré d'association porte sur la mesure pondérée *DP* qui est confrontée à l'hypothèse nulle d'une différence égale à 0. Par ailleurs, le khi-carré d'homogénéité porte sur l'hypothèse de l'homogénéité des mesures spécifiques DP_i entre elles.

10.1.2 TEST DE BRESLOW-DAY SUR L'HOMOGÉNÉITÉ DES MESURES SPÉCIFIQUES DP_i

Sous l'hypothèse d'homogénéité, les mesures DP_i fluctuent aléatoirement autour d'une même mesure paramétrique Δ. En conséquence, les déviations $[a_{1i} - E(A_{1i} \mid \Delta)]$, elles aussi, fluctuent aléatoirement autour de 0 avec une variance de $V(A_{1i} \mid \Delta)$.

Le test de Breslow-Day se présente alors comme suit :

$$\chi^2_{k-1} = \sum_{i=1}^{k} \frac{\left[a_{1i} - E(A_{1i} \mid \Delta)\right]^2}{V(A_{1i} \mid \Delta)}$$

Pour la strate i, la variable A_{1i} obéit à une loi hypergéométrique de paramètres n_i, n_{1i} et m_{1i}. Sous l'hypothèse Δ, la valeur attendue $E(A_i \mid \Delta)$ et la variance $V(A_i \mid \Delta)$ sont respectivement estimées par :

$$E(A_{1i} \mid \Delta) = \frac{\Delta n_{1i} n_{0i} + m_{1i} n_{1i}}{n_i}$$

$$V(A_{1i} \mid \Delta) = \left[\frac{1}{A_i} + \frac{1}{B_i} + \frac{1}{C_i} + \frac{1}{D_i}\right]^{-1}$$

où $A_i = E(A_{1i} \mid \Delta)$, $B_i = m_{1i} - A_i$, $C_i = n_{1i} - A_i$ et $D_i = n_{0i} - B_i$.

La différence Δ peut être estimée par la différence *DP* pondérée par l'inverse des variances V_i ou par l'estimation du maximum de vraisemblance ou encore par l'estimateur de Mantel-Haenszel.

10.1.3 INTERVALLE DE CONFIANCE EN APPROXIMATION NORMALE

Si on connaît la variance V de la mesure $DP = \sum_i \lambda_i DP_i$, on peut facilement calculer son intervalle de confiance. Une bonne estimation de V est donnée par $V = \sum_i \lambda_i^2 V_i$, où $V_i = \frac{a_{1i} b_{1i}}{n_{1i}^3} + \frac{a_{0i} b_{0i}}{n_{0i}^3}$.

L'intervalle de confiance se présente alors comme : $DP \pm z_{\alpha/2}\sqrt{V}$.

Si le système de poids $\{\lambda_i\}$ est tel que $\lambda_i \propto 1/V_i$, c'est-à-dire proportionnel à la stabilité de la mesure DP_i, alors la variance V correspond simplement à $\dfrac{1}{\sum_i w_i}$.

Ci-après, nous proposons différents systèmes de poids $\{\lambda_i\}$ parmi les plus usités.

Pour que le lecteur s'y retrouve plus facilement, nous rappelons l'expression de ces différents poids en utilisant la notation du tableau 10.1.

Ainsi, en posant $\lambda_i = \dfrac{w_i}{\sum_i w_i}$, on a :

- le poids proportionnel à l'inverse des variances,

  $w_i = \dfrac{1}{V_i} = \dfrac{1}{\dfrac{a_{1i}b_{1i}}{n_{1i}^3} + \dfrac{a_{0i}b_{0i}}{n_{0i}^3}}$, qui conduit à la mesure pondérée DP_V ;

- le poids de Mantel-Haenszel, $w_i = \dfrac{n_{1i} \times n_{0i}}{n_i}$, qui conduit à la mesure pondérée DP_{MH} ;
- le poids proportionnel à la distribution du facteur F chez les exposés, $w_i = n_{1i}$, qui conduit à la mesure pondérée DP_a ;
- le poids proportionnel à la distribution du facteur F chez les non-exposés, $w_i = n_{0i}$, qui conduit à la mesure pondérée DP_S.

10.1.4 TESTS STATISTIQUES ET INTERVALLES DE CONFIANCE PAR LA MÉTHODE DU RAPPORT DE VRAISEMBLANCE

Comme on l'a déjà mentionné, le test du rapport de vraisemblance se prête aussi bien au jugement sur l'homogénéité des mesures qu'à celui sur la mesure globale d'association. Il suffit de considérer la modélisation linéaire de la proportion comme fonction de X et de F, $\pi(X,F)$, sans et avec terme d'interaction :

Modèle 1 : $\pi(X,F) = \alpha_1 + \beta_{11}X + \beta_{12}F$

Modèle 2 : $\pi(X,F) = \alpha_2 + \beta_{21}X + \beta_{22}F + \beta_{23}XF$

Dans le premier modèle, le coefficient β_{11} représente la différence des proportions, DP_{RV}, ajustée pour le facteur F par la méthode du rapport de vraisemblance.

Dans le second modèle, le coefficient β_{23} représente le terme d'interaction additive. Il peut correspondre alors à une mesure de l'hétérogénéité entre les DP_i.

Comme pour les taux, la procédure GENMOD de SAS se prête très bien à la modélisation des proportions.

Exemple 10.1

Considérons les données suivantes (tableau 10.2) d'une étude fictive. Le facteur *X* est mis en relation avec la maladie *Y* et la stratification est faite sur le facteur *F*.

Tableau 10.2

	F = 0			*F* = 1		
	X = 1	*X* = 0	Total	*X* = 1	*X* = 0	Total
Y = 1	6	6	12	30	6	36
Y = 0	94	194	288	170	94	264
Total	100	200	300	200	100	300

Le tableau 10.3 ci-dessous présente les valeurs numériques nécessaires aux calculs des tests.

Tableau 10.3

Strate	DP_i	V_i	w_i	w_iDP_i	$w_iDP_i^2$
F = 1	0,03	0,0007095	1409,44	42,28	1,2685
F = 2	0,09	0,0012015	832,29	74,91	6,7416
Total	–	–	2241,73	117,19	8,0101

On estime la mesure globale comme $DP = \dfrac{117,19}{2241,73} = 0,05228.$

Méthode de la partition du khi-carré

La partition du khi-carré nous conduit aux statistiques suivantes :

$$\begin{cases} \chi^2_{\text{total}} = 8,0101 \\ \chi^2_{\text{assoc}} = \dfrac{117,19^2}{2241,73} = 6,1263 \qquad p = 0,0133 \\ \chi^2_{\text{homog}} = 8,0101 - 6,1263 = 1,8838 \quad p = 0,1699 \end{cases}$$

On peut donc conclure à une faible différence entre les deux mesures spécifiques [$\chi^2_1(\text{homog}) = 1{,}88$, avec un degré de liberté, $p = 0{,}1699$]. Par contre, la mesure globale *DP* est significativement différente de 0, puisque le $\chi^2_1(\text{assoc})$ est de 6,1263 ($p = 0{,}0133$). Sans discuter du contexte de l'étude, il pourrait être justifié ici de ne présenter que la mesure pondérée. (**PR10.1**)

Test de Breslow-Day pour l'homongénéité des mesures spécifiques

Nous résumons dans le tableau 10.4 les résultats du test d'homogénéité suivant trois estimateurs : celui issu de la méthode de la partition (DP_V), celui du rapport de vraisemblance (DP_{RV}) et celui de Mantel-Haenszel (DP_{MH}).

Tableau 10.4

Estimateur Δ	Strate	a_i	A_i	$V(A_i)$	$(a_i - A_i)^2/V(A_i)$	Programme
DP_V Δ = 0,0523	1	6	7,4851	2,69533	0,81826	PR10.2
	2	30	27,4851	5,86332	1,07870	
	Total	–	–	–	1,89696* (= χ^2)	
DP_{RV} Δ = 0,0538	1	6	7,5867	2,67142	0,94247	PR10.3
	2	30	27,5867	5,81978	1,00070	
	Total	–	–	–	1,94317** (= χ^2)	
DP_{MH} Δ = 0,06		6	8	2,55773	1,56389	PR10.4
		30	28	5,63705	0,70959	
	Total	–	–	–	2,27348*** (= χ^2)	

* $p = 0{,}16842$ ** $p = 0{,}16333$ ***$p = 0{,}13160$

Nous remarquons ici une bonne concordance entre les tests, particulièrement ceux qui sont basés sur les estimateurs DP_V et DP_{RV}. Par ailleurs, aucun des trois tests ne conduit au rejet de l'hypothèse de l'homogénéité.

Intervalle de confiance normal de *DP* pondéré

Pour les données du tableau 10.2, nous présentons la mesure pondérée, sa variance et son intervalle de confiance à 95 %, pour trois types de pondération, en approximation normale (tableau 10.5).

Tableau 10.5

Estimateur	Différence pondérée	Variance	IC à 95 %	Programme
DP_V	0,0523	0,00044608	[0,01088 ; 0,09367]	PR10.5
DP_{MH}	0,06	0,00047775	[0,01716 ; 0,10284]	PR10.6
DP_A	0,07	0,00061283	[0,02148 ; 0,11852]	PR10.7

TEST ET INTERVALLE DE CONFIANCE POUR LE *DP* PAR LE RAPPORT DE VRAISEMBLANCE

On considère les deux modèles décrits à la section 10.1.4, que l'on applique aux données du tableau 10.2. Les résultats peuvent être déterminés à l'aide de GENMOD de SAS. (**PR10.8**)

Dans ces résultats, on remarque d'abord une bonne homogénéité entre les mesures DP_i. Le test du rapport de vraisemblance, conduit sur l'hypothèse nulle $\beta_{23} = 0$, donne un $\chi^2_1(\text{homog}) = 1{,}7276$ ($p = 0{,}1887$) (modèle 2). Ce résultat est similaire aux résultats des tests précédents sur l'homogénéité.

Par ailleurs, le test du rapport de vraisemblance sur le coefficient β_{11} du modèle 1 conduit à un χ^2_1 de 5,4264 ($p = 0{,}0198$). Ce résultat du test concorde avec celui obtenu dans la partition du khi-carré : χ^2_1 (assoc) de 6,1263 ($p = 0{,}0133$). La différence *DP*, correspondant au coefficient β_{11}, est estimée à 0,0538 avec un intervalle de confiance à 95 % de [0,0080 ; 0,1047]. Ces valeurs sont concordantes avec celles obtenues par la méthode des poids proportionnels à l'inverse des variances.

◆

10.1.5 INTERVALLE DE CONFIANCE D'UNE INTERACTION ADDITIVE

Considérons les données du tableau 10.1 pour lequel *F* est un facteur dichotomique.

On obtient alors le tableau 10.6.

TABLEAU 10.6

	X_1X_2				Total
	11	01	10	00	
$Y = 1$	a_3	a_2	a_1	a_0	m_1
$Y = 0$	b_3	b_2	b_1	b_0	m_0
Total	n_3	n_2	n_1	n_0	n

APPROXIMATION NORMALE

L'interaction additive I_+ se mesure par : $I_+ = DP_{11} - DP_{10} - DP_{01}$.

On peut montrer que la variance $V(I_+)$ de I_+ est donnée par :

$$V(I_+) = \sum_{i=0}^{3} \frac{a_i b_i}{n_i^3}$$

Si on connaît la variance de I_+, et qu'on utilise l'approximation normale, il est facile d'en déduire l'intervalle de confiance pour un niveau $100(1-\alpha)\,\%$:

$$(I_+)_{\text{inf}} = I_+ - z_{\alpha/2}\sqrt{\sum_{i=0}^{3}\frac{a_i b_i}{n_i^3}}$$

$$(I_+)_{\text{sup}} = I_+ + z_{\alpha/2}\sqrt{\sum_{i=0}^{3}\frac{a_i b_i}{n_i^3}}$$

Méthode du rapport de vraisemblance

Cette méthode est décrite à la section 8.3.2 du chapitre 8.

Il suffit de résoudre l'équation $2\left[L - L(\beta^*)\right] = \chi^2_{1,1-\alpha}$

où L correspond à la valeur maximale de la fonction du $\log FV(\alpha_2, \beta_{21}, \beta_{22}, \beta_{23})$ et $L(\beta^*)$ à la valeur maximale de la fonction du $\log FV(\alpha_2, \beta_{21}, \beta_{22}, \beta^*)$ pour une valeur fixe β^* de β_{23} (voir le modèle 2 de la section 10.1.4).

Exemple 10.2

Revenons aux données du tableau 10.2. Sur ces données, on obtient les mesures suivantes :

$$DP_{11} = \frac{30}{200} - \frac{6}{200} = \frac{12}{100},\ DP_{10} = \frac{6}{100} - \frac{6}{200} = \frac{3}{100} \text{ et } DP_{01} = \frac{6}{100} - \frac{6}{200} = \frac{3}{100}.$$

De ces mesures, on peut facilement obtenir celle de l'interaction additive I_+ :

$$\begin{aligned} I_+ &= DP_{11} - DP_{10} - DP_{01} \\ &= \frac{12}{100} - \frac{3}{100} - \frac{3}{100} \\ &= \frac{6}{100} \end{aligned}$$

Méthode en approximation normale

La variance de I_+ est estimée par :

$$V(I_+) = \frac{30 \times 170}{200^3} + \frac{6 \times 94}{100^3} + \frac{6 \times 94}{100^3} + \frac{6 \times 194}{200^3} = 0{,}001911.$$

Les limites de confiance à 95 % sont alors données par :

$$(I_+)_{\text{inf}} = 0,06 - 1,96\sqrt{0,001911}$$
$$= -0,0257$$
$$(I_+)_{\text{sup}} = 0,06 + 1,96\sqrt{0,001911}$$
$$= 0,1457$$

(**PR10.9**)

Méthode du rapport de vraisemblance

En utilisant GENMOD de SAS, on obtient facilement les limites de confiance du rapport de vraisemblance pour l'interaction I_+. Pour les données du tableau 9.2, on obtient :

$$(I_+)_{\text{inf}} = -0,0317$$
$$(I_+)_{\text{sup}} = 0,1425$$

(**PR10.10**)

◆

10.2 RAPPORT *RP* DE DEUX PROPORTIONS EN ANALYSE STRATIFIÉE

En référence aux notations de la section 10.1, on définit les rapports de proportions spécifiques comme $RP_i = \dfrac{p_{1i}}{p_{0i}}$ et le rapport de taux pondéré comme $RP = \sum_i \lambda_i RP_i$.

10.2.1 TESTS STATISTIQUES EN APPROXIMATION NORMALE SUR LES RAPPORTS DE PROPORTIONS

Méthode de la partition du khi-carré

Soient les rapports de proportions $\{RP_i\}$ en analyse stratifiée. À partir d'un système de poids $\{\lambda_i\}$, on résume ces mesures en une mesure pondérée *RP* telle que $RP = \sum_i \lambda_i RP_i$. Alors, pour répondre aux problèmes de l'homogénéité des mesures RP_i et de la signification statistique de la mesure pondérée *RP*, nous faisons appel à la partition du χ^2(total) en χ^2(association) et χ^2(homogénéité). Comme pour le *RT* (section 9.2.1 du chapitre 9), la partition sera pratiquée sur la transformation logarithmique de *RP*.

Une fois les calculs faits, on obtient par transformation inverse les expressions désirées pour le rapport de proportions.

Si on pose $w_i = \dfrac{1}{V\left[\log RP_i\right]} = \dfrac{1}{V_i}$, on a les relations suivantes :

$$\begin{cases} \chi^2_{\text{total}} = \sum_i w_i \left(\log RP_i\right)^2 \\ \chi^2_{\text{assoc}} = \dfrac{\left[\sum_i w_i \log RP_i\right]^2}{\sum_i w_i} \\ \chi^2_{\text{homog}} = \sum_i w_i \left(\log RP_i - \overline{\log RP}\right)^2 \end{cases}$$

Par la méthode delta, on obtient $V_i = \dfrac{b_{1i}}{a_{1i}n_{1i}} + \dfrac{b_{0i}}{a_{0i}n_{0i}}$.

Rappelons que $\overline{\log RP} = \dfrac{\sum_i w_i \log(RP_i)}{\sum_i w_i}$ et que $V\left(\overline{\log RP}\right) = \dfrac{1}{\sum_i w_i}$.

La valeur $RP_V = e^{\overline{\log RP}}$ est une moyenne géométrique des RP_i.

10.2.2 TEST DE BRESLOW-DAY SUR L'HOMOGÉNÉITÉ DES MESURES RP_i

Sous l'hypothèse d'homogénéité, les mesures RP_i fluctuent aléatoirement autour d'une même mesure paramétrique ξ. Ce rapport ξ peut être estimé soit par le rapport RP correspondant à la transformation inverse de $\overline{\log RP}$, soit par le maximum de vraisemblance (RP_{RV}). C'est donc dire que les déviations $[a_{1i} - E(A_{1i} \mid \xi)]$ fluctuent aléatoirement autour de 0 avec une variance de $V(A_{1i} \mid \xi)$.

Le test se présente alors comme : $\chi^2_{k-1} = \sum_{i=1}^{k} \dfrac{\left(a_{1i} - E(A_{1i} \mid \xi)\right)^2}{V(A_{1i} \mid \xi)}$.

Pour la strate i, la variable A_{1i} obéit à une loi hypergéométrique de paramètres n_i, n_{1i} et m_{1i}. Sous l'hypothèse ξ, sa valeur attendue $E(A_{1i} \mid \xi)$ et sa variance $V(A_{1i} \mid \xi)$ sont respectivement estimées par :

$$E(A_{1i} \mid \xi) = m_{1i}\left(\frac{\xi n_{1i}}{\xi n_{1i} + n_{0i}}\right)$$

$$V(A_{1i} \mid \xi) = \left[\frac{1}{A_i} + \frac{1}{B_i} + \frac{1}{C_i} + \frac{1}{D_i}\right]^{-1}$$

où $A_i = E(A_{1i} \mid \xi)$, $B_i = m_{1i} - A_i$, $C_i = n_{1i} - A_i$ et $D_i = n_{0i} - B_i$.

10.2.3 INTERVALLE DE CONFIANCE EN APPROXIMATION NORMALE DE *RP* PONDÉRÉ

Dans le cadre d'une partition du khi-carré

On rappelle que la mesure RP_V est une moyenne géométrique des différentes mesures spécifiques RP_i (voir la section 10.2.1). De même que RP_V est la transformation inverse de la somme pondérée des $\log(RP_i)$, ses limites de confiance seront aussi obtenues par transformation inverse des limites de confiance de cette même somme pondérée :

Ainsi :

$$\xi_{\text{inf}} = RP_V \times e^{-z_{\alpha/2}\sqrt{\frac{1}{\sum_i w_i}}}$$

$$\xi_{\text{sup}} = RP_V \times e^{+z_{\alpha/2}\sqrt{\frac{1}{\sum_i w_i}}}$$

où $w_i = \dfrac{1}{V_i}$.

Dans le cadre d'une pondération arithmétique

On s'intéresse à l'intervalle de confiance d'un rapport de proportions *RP* tel que $RP = \sum_i \lambda_i RP_i$, où $\{\lambda_i\}$ est un système de poids. Puisqu'en bonne approximation $V\left[\log\left(\sum_i \lambda_i RP_i\right)\right] = \sum_i \lambda_i^2 V\left[\log RP_i\right]$, l'intervalle de confiance de *RP* se calcule simplement comme : $RP \times e^{\pm z_{\alpha/2}\sqrt{\sum_i \lambda_i^2 V_i}}$ où $V_i = \dfrac{b_{1i}}{a_{1i}n_{1i}} + \dfrac{b_{0i}}{a_{0i}n_{0i}}$. Ainsi,

$$\xi_{\text{inf}} = RP \times e^{-z_{\alpha/2}\sqrt{\sum_i \lambda_i^2 V_i}}$$

$$\xi_{\text{sup}} = RP \times e^{+z_{\alpha/2}\sqrt{\sum_i \lambda_i^2 V_i}}$$

Ci-dessous, nous proposons différents systèmes de poids $\{\lambda_i\}$ parmi les plus usités pour le *RP*. Pour que le lecteur s'y retrouve plus facilement, nous rappelons l'expression de ces différents poids en utilisant la notation du tableau 10.1.

Ainsi, en posant $\lambda_i = \dfrac{w_i}{\sum_i w_i}$, on a :

- le poids de Mantel-Haenszel, $w_i = \dfrac{a_{0i} \times n_{1i}}{n_i}$, qui conduit à la mesure pondérée RP_{MH} ;
- le poids de type *SMR*, $w_i = \dfrac{a_{0i} \times n_{1i}}{n_{0i}}$, qui conduit à la mesure pondérée RP_a ;
- le poids standardisé, proportionnel à la distribution du facteur F chez les cas non exposés, $w_i = a_{0i}$, qui conduit à la mesure pondérée RP_S.

10.2.4 TESTS STATISTIQUES ET INTERVALLE DE CONFIANCE PAR LA MÉTHODE DU RAPPORT DE VRAISEMBLANCE

Dans l'approche de la méthode du rapport de vraisemblance, le rapport de proportions est traité à l'aide de la transformation logarithmique de la proportion : la loi de distribution sous-jacente est la loi binomiale et la fonction de lien, la transformation logarithmique. Si $\pi(X,F)$ désigne la proportion comme une fonction des variables indépendantes X et F, alors les modèles considérés peuvent avoir la forme suivante :

Modèle 1 : $\log\ \pi(X,F) = \alpha_1 + \beta_{11}X + \beta_{12}F$

Modèle 2 : $\log\ \pi(X,F) = \alpha_2 + \beta_{21}X + \beta_{22}F + \beta_{23}XF$

Dans le premier modèle, le coefficient β_{11} représente le logarithme du rapport des proportions, ajusté pour le facteur F : $e^{\beta_{11}} = RP_{RV}$.

Dans le second modèle, le coefficient β_{23} représente le logarithme de l'interaction multiplicative entre X et F et marque ainsi l'hétérogénéité des RP_i à travers les strates de F. Si X et F sont dichotomiques, alors $e^{\beta_{23}} = \dfrac{RP_1}{RP_0}$. On peut écrire aussi $e^{\beta_{23}} = \dfrac{RP_{11}}{RP_{10} \times RP_{01}}$, où RP_{ij} représente le rapport de proportions issu de la comparaison entre la catégorie ($X = i$, $F = j$) et la catégorie de référence ($X = 0$, $F = 0$).

De même, pour porter un jugement sur la signification du rapport de proportions ajusté pour le facteur F, il suffit d'appliquer le test du rapport de vraisemblance au coefficient β_{11} du modèle 1.

Les intervalles de confiance s'obtiennent aussi en solutionnant pour β l'équation de vraisemblance $\left[L - L(\beta)\right] - 0,5\chi^2_{1,1-\alpha} = 0$.

Exemple 10.3

Une étude a été conduite sur la relation entre l'exposition au facteur *X* et la maladie *Y*. Les résultats stratifiés pour l'âge sont décrits au tableau 10.7.

Tableau 10.7

	Âge en années			
	20-49		50-79	
	$X = 1$	$X = 0$	$X = 1$	$X = 0$
$Y = 1$	80	40	760	90
$Y = 0$	1170	3710	1440	1710
Total	1250	3750	2200	1800

Tests en approximation normale

Méthode de la partition du khi-carré

Le tableau 10.8 présente les valeurs numériques nécessaires aux calculs des tests.

Tableau 10.8

Âge (années)	RP_i	Log (RP_i)	V_i	w_i	$w_i \log(RP_i)$	$w_i[\log RP_i]^2$
20-49	6,00	1,79	0,0364	27,45	49,18	88,12
50-79	6,91	1,93	0,0114	87,59	169,30	327,22
Total		–	–	115,04	218,48	415,34

La partition du khi-carré nous conduit aux statistiques décrites ci-dessous.

$$\begin{cases} \chi^2_{\text{total}} = 415,342 \\ \chi^2_{\text{assoc}} = \dfrac{218,48^2}{115,04} = 414,926 \quad p \approx 0 \\ \chi^2_{\text{homog}} = 0,416 \qquad\qquad p = 0,51897 \end{cases}$$

On peut donc conclure à une faible différence entre les deux mesures spécifiques. En effet, le khi-carré d'homogénéité est égal à 0,416 avec un degré de liberté ($p = 0,519$]. Par contre, la mesure globale log(*RP*) est significativement différente de 0, puisque le khi-carré d'association est de 414,926 ($p \approx 0$). Sans discuter du contexte de l'étude, il pourrait être justifié ici de ne présenter que la mesure globale *RP* correspondante. (**Pr10.11**)

Puisque $\log(RP) = \dfrac{218,48}{115,04} = 1,90$, alors $RP_V = e^{1,90} = 6,68$.

Intervalle de confiance normal du *RP* par transformation logarithmique

Nous décrivons alors, dans le tableau 10.9, la mesure pondérée *RP*, son logarithme, la variance du log et l'intervalle de confiance à 95 % du *RP* suivant différents systèmes de pondération.

Tableau 10.9

Mesure	*RP*	Log(*RP*)	*V*[log *RP*]	IC à 95 %	Programme
RP_V	6,6804	1,8992	0,008693	[5,56 ; 8,02]	(PR10.12)
RP_{MH}	6,7563	1,9091	0,008931	[5,61 ; 8,13]	
RP_a	6,8108	1,9185	0,009507	[5,63 ; 8.25]	(PR10.13)
RPs	6,6294	1,8915	0,008921	[5,51 ; 7,98]	

Nous remarquons que les *RP* obtenus diffèrent peu d'un système de poids à l'autre. Cela est dû principalement à la grande homogénéité entre les deux mesures spécifiques. La mesure *RP* ajustée, quelque soit le système de poids, doit obligatoirement se situer entre ces deux valeurs spécifiques : $RP_1 = 6{,}00$ et $RP_2 = 6{,}91$.

Test et intervalle de confiance du *RP* par la méthode du rapport de vraisemblance

On applique les modèles 1 et 2 de la section 10.2.4 aux données du tableau 10.7. Le facteur *X* correspond à l'exposition et *F* à l'âge. Les résultats nous conduisent aux constatations suivantes.

On remarque d'abord une faible hétérogénéité : $\beta_{23} \approx 0{,}1411$. Le test du rapport de vraisemblance appliqué à ce coefficient donne un khi-carré de 0,4128 pour une valeur-*p* de 0,5206. Par ailleurs, le coefficient β_{11} du premier modèle est égal à 1,9001 et le test du rapport de vraisemblance sur ce coefficient conduit à un khi-carré de 681,5657 pour un $p \approx 0$, ce qui suggère un rapport de proportions *RP* très significativement différent de 1. Ce rapport est estimé à $e^{1{,}9001} = 6{,}69$, avec un intervalle de confiance à 95 % de [5,6008 ; 8,0538]. Ces résultats sont similaires à ceux obtenus précédemment par la méthode de la partition du khi-carré.

(PR10.14)

Test de Breslow-Day sur l'homogénéité des mesures RP_i

Pour les données de l'exemple, nous résumons au tableau 10.10 les résultats du test d'homogénéité de Breslow-Day suivant trois estimateurs : RP_V, RP_{RV} (du rapport de vraisemblance) et RP_{MH}.

TABLEAU 10.10

Estimateur de ξ	Strate	a_i	A_i	$V(A_i)$	$(a_i - A_i)^2/V(A_i)$	Programme
RP_V $\xi = 6{,}68$	1	80	82,811	24,9428	0,31689	PR10.15
	2	760	757,255	74,7290	0,10082	
	Total	–	–	–	0,41771*	
RP_{RV} $\xi = 6{,}69$	1	80	82,834	24,9345	0,32220	PR10.16
	2	760	757,329	74,6816	0,09552	
	Total	–	–	–	0,41772**	
RP_{MH} $\xi = 6{,}76$	1	80	83,101	24,8381	0,38710	PR10.17
	2	760	758,185	74,1326	0,04446	
	Total			–	0,43156***	

* $p = 0{,}51808$ ** $p = 0{,}51808$ *** $p = 0{,}51123$

Nous remarquons ici une très bonne concordance entre les tests. Aucun test ne conduit au rejet de l'hypothèse de l'homogénéité.

◆

10.2.5 INTERVALLE DE CONFIANCE D'UNE INTERACTION MULTIPLICATIVE

Considérons les données du tableau 10.6, que nous reprenons dans le tableau 10.11.

TABLEAU 10.11

	X_1X_2				Total
	11	01	10	00	
$Y = 1$	a_3	a_2	a_1	a_0	m_1
$Y = 0$	b_3	b_2	b_1	b_0	m_0
Total	n_3	n_2	n_1	n_0	n

APPROXIMATION NORMALE

On rappelle que l'interaction multiplicative $I_\times$ se mesure par :

$$I_\times = \frac{RP_{11}}{RP_{10} \times RP_{01}}.$$

On peut montrer que la variance $V[\log(I_\times)]$ est donnée par : $V\left[\log(I_\times)\right] = \sum_{i=0}^{3} \frac{b_i}{a_i n_i}$.

Si on connaît la variance de $\log(I_\times)$ et qu'on utilise l'approximation normale, il est facile d'en déduire l'intervalle de confiance pour un niveau $100(1-\alpha)\,\%$:

$$(I_\times)_{\text{inf}} = (I_\times) \times \exp\left(-z_{\alpha/2} \sqrt{\sum_{i=0}^{3} \frac{b_i}{a_i n_i}} \right)$$

$$(I_\times)_{\text{sup}} = (I_\times) \times \exp\left(+z_{\alpha/2} \sqrt{\sum_{i=0}^{3} \frac{b_i}{a_i n_i}} \right)$$

Méthode du rapport de vraisemblance

Il suffit de résoudre l'équation $2\left[L - L(\beta^*)\right] = \chi^2_{1,1-\alpha}$.

où L correspond à la valeur maximale de la fonction du $\log FV(\alpha_2, \beta_{21}, \beta_{22}, \beta_{23})$ et $L(\beta^*)$ à la valeur maximale de la fonction du $\log FV(\alpha_2, \beta_{21}, \beta_{22}, \beta^*)$ pour une valeur fixe β^* de β_{23} (voir le modèle 2 de la section 10.2.4).

Exemple 10.4

Reprenons les données du tableau 10.7. Supposons que le groupe d'âge 20-49 constitue la référence pour la variable âge. Alors sur ces données, on obtient les mesures suivantes :

$$RP_{11} = \frac{760 \times 3750}{40 \times 2200} = 32,39,\ RP_{10} = \frac{80 \times 3750}{40 \times 1250} = 6,00 \text{ et}$$

$$RP_{01} = \frac{90 \times 3750}{40 \times 1800} = 4,69.$$

De ces mesures, on peut facilement obtenir celle de l'interaction multiplicative $I_\times$:

$$\begin{aligned} I_\times &= \frac{RP_{11}}{RP_{10} \times RP_{01}} \\ &= \frac{32,39}{6,00 \times 4,69} \\ &= 1,15 \end{aligned}$$

Méthode en approximation normale

La variance de $I_\times$ est estimée par :

$$V(\log I_\times) = \frac{1440}{760 \times 2200} + \frac{1170}{80 \times 1250} + \frac{1710}{90 \times 1800} + \frac{3710}{40 \times 3750} = 0,04785$$

Les limites de confiance à 95 % sont alors données par :

$$\begin{aligned}(I_\times)_{\text{inf}} &= 1,15 \times \exp\left(-1,96\sqrt{0,04785}\right) \\ &= 0,75 \\ (I_\times)_{\text{sup}} &= 1,15 \times \exp\left(+1,96\sqrt{0,04785}\right) \\ &= 1,77\end{aligned}$$

(**PR10.18**)

Méthode du rapport de vraisemblance

En utilisant GENMOD de SAS, on obtient les limites de confiance du rapport de vraisemblance pour l'interaction $I_\times$. Pour les données du tableau 10.7, ces limites de confiance sont sensiblement les mêmes que celles obtenues par approximation normale :

$$\begin{aligned}(I_\times)_{\text{inf}} &= 0,75 \\ (I_\times)_{\text{sup}} &= 1,76\end{aligned}$$

(**PR10.19**)

◆

10.3 MESURE DU *SMR* POUR LES PROPORTIONS EN ANALYSE STRATIFIÉE

Dans la section 6.4 du chapitre 6, nous avons considéré le *SMR* a été considéré dans sa plus simple expression. Les propriétés de la loi binomiale vont permettre ici d'étendre aux analyses stratifiées les principaux outils statistiques définis pour les analyses simples.

On considère une variable de stratification F comportant k catégories. Pour chaque strate i de la variable F, on désigne respectivement par n_i et π_{0i} les personnes observées dans l'échantillon et la proportion de décès estimée à partir de la population standard. La valeur attendue de cas X_{0i} correspond à $n_i\pi_{0i}$. Si, sur cette strate, le nombre de cas observés est de π_i, alors on a :

$$SMR_i = \frac{x_i}{X_{0i}}$$

On peut aussi considérer la définition théorique suivante. Pour la strate i, le nombre de cas observés x_i est la réalisation d'une variable binomiale X_i de paramètre $n_i\pi_i$ où π_i est une proportion inconnue. La valeur x_i est la meilleure estimation disponible pour $n_i\pi_i$. Le SMR_i correspond alors simplement à l'estimation du rapport ϕ_i des proportions π_i et π_{0i} :

$$SMR_i = \frac{n_i\pi_i}{n_{0i}\pi_{0i}} = \frac{\pi_i}{\pi_{0i}} = \phi_i$$

Ce SMR_i est l'estimation de ϕ_i (le *SMR* théorique pour la strate i), qui compare le risque π_i au risque π_{0i} de la population standard.

$$SMR_i = \frac{x_i}{X_{0i}} \approx \frac{n_i\pi_i}{n_i\pi_{0i}} = \frac{\pi_i}{\pi_{0i}} = \phi_i$$

Sous H_0, $\pi_i = \pi_{0i}$, ce qui correspond à un SMR_i de 1.

Le paramètre π_i de la variable binomiale X_i peut se décrire comme une fonction de ϕ_i :

$$\phi_i = \frac{n_i\pi_i}{n_i\pi_{0i}} = \frac{n_i\pi_i}{X_{0i}} \Rightarrow \pi_i = \frac{\phi_i X_{0i}}{n_i}$$

La fonction de vraisemblance construite sur les variables binomiales X_i en fonction des paramètres ϕ_i prend alors la forme suivante :

$$FV(\{\phi_i\}) = \prod_i C_{n_i}^{x_i} \left(\frac{\phi_i X_{0i}}{n_i}\right)^{x_i} \left(\frac{n_i - \phi_i X_{0i}}{n_i}\right)^{n_i - x_i}$$

Pour le *SMR* global ajusté, nous proposons deux estimateurs :

1. l'estimateur simplement défini comme $SMR = \frac{\sum_i x_i}{\sum_i X_{0i}} = \frac{x}{X_0}$ où x désigne le nombre total de cas observés et X_0 le nombre total de cas attendus sous l'hypothèse nulle.
2. l'estimateur du maximum de vraisemblance, correspondant au ϕ qui maximise la fonction de vraisemblance $FV(\phi) = \prod_i C_{n_i}^{x_i} \left(\frac{\phi X_{0i}}{n_i}\right)^{x_i} \left(\frac{n_i - \phi X_{0i}}{n_i}\right)^{n_i - x_i}$ ou, dans la transformation logarithmique, la fonction $L(\phi) = x_i \log\phi + (n_i - x_i)$

$\log(n_i - \phi X_{0i}) + K$. Cette valeur doit être une solution de l'équation de vraisemblance suivante :

$$\frac{\partial L(\phi)}{\partial \phi} = \sum_i \left(\frac{x_i}{\phi} - \frac{X_{0i}(n_i - x_i)}{n_i - \phi X_{0i}} \right) = 0$$

Cette solution peut être déterminée à l'aide de méthodes itératives.

L'estimateur simple

La mesure $SMR = \dfrac{x}{X_0}$ est de calcul simple. Dans cette forme, le *SMR* est la somme des SMR_i spécifiques pondérés par les valeurs attendues X_{0i}. Il s'interprète de façon analogue au *SMR* sur les taux. Le nombre total X est la somme des variables binomiales indépendantes X_i. La valeur attendue $E(X)$ de X est la somme des valeurs attendues $E(X_i)$ et sa variance $V(X)$, la somme des variances $V(X_i)$. Ainsi, sous l'hypothèse nulle, on a :

$$\begin{aligned} E(X) &= X_0 \\ V(X) &= \sum_i \frac{X_{0i}(n_i - X_{0i})}{n_i} \\ &= V_0 \end{aligned}$$

L'estimateur du maximum de vraisemblance

L'estimateur du maximum de vraisemblance, qui est une solution de la fonction de vraisemblance, n'a pas de forme explicite.

Par ailleurs, pour l'analyse du *SMR* par la méthode du maximum de vraisemblance, nous serons appelés à considérer trois fonctions de vraisemblance à partir desquelles seront définis les tests du rapport de vraisemblance pour la signification du *SMR* global et pour l'homogénéité des SMR_i spécifiques. Ces fonctions sont les suivantes :

Fonction pour le modèle de base $\phi = 1$

Ce modèle correspond à l'hypothèse nulle d'un $SMR = 1$. On rappelle que sous l'hypothèse nulle $\phi = 1$, on a $E(X) = X_0$ et $V(X) = V_0$.

Sous cette hypothèse, la fonction de vraisemblance prend la valeur L_0.

$$L_0 = \sum_{i=1}^{k} \left[(n_i - x_i) \log (n_i - X_{0i}) \right] + K$$

C'est la valeur de la fonction correspondant à l'hypothèse (nulle) d'un $SMR = 1$.

Fonction pour le modèle ϕ du maximum de vraisemblance

Ce modèle correspond à l'hypothèse du maximum de vraisemblance.

Sous cette hypothèse, la fonction de vraisemblance prend la valeur L_1 :

$$L_1 = \sum_{i=1}^{k} \left[x_i \log \phi + (n_i - x_i) \log(n_i - \phi X_{0i}) \right] + K$$

où ϕ correspond à l'estimateur du maximum de vraisemblance.

Fonction pour le modèle qui sature les données

Le modèle qui sature les données est celui qu'on obtient en remplaçant dans la fonction L chaque ϕ_i par $\frac{x_i}{X_{0i}}$, qui en est l'estimateur du maximum de vraisemblance.

$$L_S = \sum_{i=1}^{k} \left[x_i \log\left(\frac{x_i}{X_{0i}} \right) + (n_i - x_i) \log(n_i - x_i) \right] + K$$

À ce stade-ci de la présentation du *SMR* pour les proportions, il faut souligner certains faits.

1. La variable X, qui est une somme de variables binomiales, n'est pas en général une variable binomiale.

2. Le *SMR* défini par le simple rapport $\frac{x}{X_0}$ n'est pas toujours l'estimateur du maximum de vraisemblance.

3. Si ϕ est une mesure pondérée du *SMR*, alors le paramètre $\pi_i^* = \phi \pi_{0i}$ de la loi binomiale pour X_i est défini correctement seulement si $\phi \leq \frac{1}{\pi_{0i}}$. En conséquence, le *SMR* ne peut être traité dans le cadre de la loi binomiale que si $\phi \leq \min_i \left\{ \frac{1}{\pi_{0i}} \right\}$.

4. Si les SMR_i sont homogènes, alors toute pondération conduit à une même mesure ϕ (ça va de soi). Si les π_{0i} sont homogènes, alors l'estimateur $SMR = \dfrac{X}{X_0}$ est l'estimateur du maximum de vraisemblance (on peut le montrer facilement). Si les SMR_i et les π_{0i} sont variables, alors les mesures pondérées du *SMR* diffèrent en général.

10.3.1 TESTS SUR LE *SMR*

Puisqu'en général, l'estimateur simple n'est pas celui du maximum de vraisemblance (sauf si les π_{0i} sont homogènes), nous présentons deux tests en approximation normale pour ce premier estimateur. Par la suite, nous présentons le test du rapport de vraisemblance. Les approches exactes, qui s'avèrent laborieuses, sont exclues de cette présentation.

TESTS EN APPROXIMATION NORMALE POUR L'ESTIMATEUR SIMPLE

TEST SIMPLE

Le test se définit comme : $\dfrac{(x - X_0)^2}{V_0} = \chi_1^2$, où, rappelons-le, $V_0 = \sum_i \dfrac{X_{0i}(n_i - X_{0i})}{n_i}$.

TEST BASÉ SUR LOG(SMR)

Ce test se définit comme :

$$\chi_1^2 = \frac{(\log SMR)^2}{V(\log SMR)} = \frac{(x \log SMR)^2}{V}, \text{ où } V = \sum_i \frac{x_i(n_i - x_i)}{n_i}$$

TEST DU RAPPORT DE VRAISEMBLANCE

Le test du rapport de vraisemblance peut être considéré comme émanant d'une partition de la déviance totale à expliquer sur la base du modèle $\phi = 1$.

La déviance totale est mesurée par la différence $2[L_S - L_0]$. Elle peut être partitionnée suivant deux composantes : la déviance expliquée $2[L_1 - L_0]$ par le modèle $\phi = SMR$ (estimateur du maximum de vraisemblance) et la déviance résiduelle $2[L_S - L_1]$, marquée par la différence entre le modèle $\phi = SMR$ et le modèle saturé.

Nous présentons cette partition, et les tests qui en résultent, à l'aide du tableau de la déviance (tableau 10.12).

TABLEAU 10.12

Composante	Modèle	Degrés de liberté	Déviance	Test du khi-carré
Totale	$\phi = 1$	k*	$2[L_S - L_0]$	$2\sum_{i=1}^{k}\left[x_i \log\left(\frac{x_i}{X_{0i}}\right) + (n_i - x_i)\log\frac{n_i - x_i}{n_i - X_{0i}}\right]$
Expliquée	$\phi = SMR$	1	$2[L_1 - L_0]$	$2\sum_{i=1}^{k}\left[x_i \log\phi + (n_i - x_i)\log\frac{n_i - \phi X_{0i}}{n_i - X_{0i}}\right]$
Résiduelle	–	$k - 1$	$2[L_S - L_1]$	$2\sum_{i=1}^{k}\left[x_i \log\left(\frac{x_i}{\phi X_{0i}}\right) + (n_i - x_i)\log\frac{n_i - x_i}{n_i - \phi X_{0i}}\right]$

* Nombre de strates impliquées dans la définition du *SMR*.

Le test sur le modèle $\phi = SMR$ est le test d'association qui compare la valeur ϕ du *SMR* à la valeur 1 (valeur sous l'hypothèse nulle). Le test d'association se présente alors comme :

$$\chi_1^2 = 2\sum_i \left[x_i \log\ \phi + (n_i - x_i)\ \log\ \left(\frac{n_i - \phi X_{0i}}{n_i - X_{0i}}\right)\right]$$

Le test sur la déviance résiduelle permet de porter un jugement sur l'homogénéité des SMR_i spécifiques. Nous reviendrons sur ce test dans une section ultérieure.

En particulier, le test sur l'existence d'une association se confond avec le test sur la composante totale :

$$2\sum_{i=1}^{k}\left[x_i \log\left(\frac{x_i}{X_{0i}}\right) + (n_i - x_i)\log\frac{n_i - x_i}{n_i - X_{0i}}\right]$$

EXEMPLE 10.5

Chez les personnes âgées de 40 ans et plus d'une région donnée, on observe 114 cas incidents de la maladie *Y* dans une période d'un an. Dans le tableau 10.13 sont décrits, suivant les groupes d'âge chez les 40 ans et plus, la répar-

tition des cas (x_i) de maladie *Y* observée dans la région R pour une certaine période, les effectifs (en personnes n_i) de cette même région et les incidences cumulatives (proportions) (π_{0i}) pour la population standard (générale).

TABLEAU 10.13

Âge	x_i	n_i	p_{0i}	X_{0i}	SMR_i
40-49	15	1200	0,01	12	1,25
50-59	20	750	0,02	15	1,33
60-69	54	500	0,05	25	2,16
70 et +	25	100	0,10	10	2,50
Total	114	2550	0,051	62	1,84

À partir de ces données, peut-on dire qu'il y a un excès de cas incidents dans la région *R* ?

En d'autres termes, le nombre observé de cas est-il significativement supérieur à celui attendu sous l'hypothèse que, tenant compte de l'âge, cette région soit affectée des mêmes incidences cumulatives de la maladie *Y* que la population générale ?

TESTS EN APPROXIMATION NORMALE

Nous considérons alors le *SMR* global :

$$SMR = \frac{114}{62} = 1{,}84$$

TEST SIMPLE

Appliqué aux données du tableau 10.13, on a : $x = 114$, $X_0 = 62$ et la variance V_0 se calcule comme suit : $V_0 = \frac{12(1200-12)}{1200} + \frac{15(750-15)}{750} + \frac{25(500-25)}{500} + \frac{10(100-10)}{100} = 59{,}33$.

Alors, le test conduit à $\chi_1^2 = \frac{(114-62)^2}{59{,}33} = 45{,}58$, ce qui, pour 1 degré de liberté, correspond à une valeur $p \approx 0$.

TEST BASÉ SUR LOG(*SMR*)

Sur les données du tableau 10.13, la variance *V* de log *SMR* est estimée comme : $V = \frac{15(1200-15)}{1200} + \frac{20(750-20)}{750} + \frac{54(500-54)}{500} + \frac{25(100-25)}{100} = 101{,}20$

La valeur du test est donc de : $\chi_1^2 = \dfrac{\left[114 \log \dfrac{114}{62}\right]^2}{101,20} = 47,64$ ce qui donne aussi $p \approx 0$.

Test du rapport de vraisemblance (RV)

Le *SMR* estimé par le maximum de vraisemblance est de 1,867.

Appliqué aux données du tableau 10.13, le test du rapport de vraisemblance donne :

$$\chi_1^2 = 2\left[\begin{array}{l} 15 \log \dfrac{15}{12} + (1200-15)\log \dfrac{1200-15}{1200-12} + \\ 20 \log \dfrac{20}{15} + \ldots + (100-25)\log \dfrac{100-25}{100-10} \end{array}\right] = 47,69 .$$

Dans le tableau 10.14, nous rappelons les résultats des différents tests appliqués aux données de l'exemple.

Tableau 10.14

Méthode	Khi-carré	Valeur-*p*	Programme
Simple	45,58	$1,47 \times 10^{-11}$	Pr10.20
Log(*SMR*)	47,64	$5,12 \times 10^{-12}$	Pr10.21
RV	47,69	$4,99 \times 10^{-12}$	Pr10.22

◆

10.3.2 Intervalle de confiance pour le *SMR*

Intervalle de confiance par approximation normale pour l'estimateur simple

Nous nous intéressons à l'estimateur simple $\dfrac{x}{X_0}$ du *SMR*, dont la valeur théorique correspond à ϕ.

Méthode simple

En se rappelant que $SMR = \dfrac{\sum_i x_i}{\sum_i X_{0i}} = \dfrac{x}{X_0}$ et que X_i est une variable binomiale, on peut exprimer la variance $V(SMR)$ comme suit :

$$V(SMR) = \frac{1}{X_0^2} \sum_i \frac{x_i(n_i - x_i)}{n_i}$$

Les limites de confiance sont alors données par :

$$\phi_{\text{inf}} = SMR - z_{\alpha/2}\frac{1}{X_0}\sqrt{\sum_i \frac{x_i(n_i - x_i)}{n_i}}$$

$$\phi_{\text{sup}} = SMR + z_{\alpha/2}\frac{1}{X_0}\sqrt{\sum_i \frac{x_i(n_i - x_i)}{n_i}}$$

L'instabilité de la variance du *SMR* peut toutefois influer sur les résultats. La méthode suivante permet de pallier ce problème.

Utilisation de la transformation logarithmique du *SMR*

En se rappelant que $SMR = \Sigma\lambda_i SMR_i$, où les poids λ_i sont proportionnels aux valeurs attendues X_{0i}, et en appliquant la transformation logarithmique au *SMR*, on obtient les limites de confiance suivantes :

$$\phi_{\text{inf}} = SMR \times e^{-z_{\alpha/2}\sqrt{\sum_i \lambda_i^2 \frac{(n_i - x_i)}{n_i x_i}}}$$

$$\phi_{\text{sup}} = SMR \times e^{+z_{\alpha/2}\sqrt{\sum_i \lambda_i^2 \frac{(n_i - x_i)}{n_i x_i}}}$$

Intervalle de confiance par la méthode du rapport de vraisemblance

L'estimation de l'intervalle de confiance du *SMR* par la méthode du rapport de vraisemblance n'est valide que pour son estimateur du maximum de vraisemblance. Il suffirait de résoudre pour ϕ l'équation logarithmique $\left[L - L(\phi)\right] - 0,5\chi^2_{1,1-\alpha} = 0$, construite à partir de la comparaison des deux fonctions de vraisemblance :

$$L(\phi) = \sum_{i=1}^{k} x_i \log\left(\phi X_{0i}\right) + (n_i - x_i)\log(n_i - \phi X_{0i}) + K$$

et

$$L = \sum_{i=1}^{k} x_i \log\left(x_i\right) + (n_i - x_i)\log(n_i - x_i) + K$$

Pour de tels calculs, l'utilisation de la procédure GENMOD de SAS s'avère très utile.

Exemple 10.6

Dans le tableau 10.15, nous décrivons le *SMR* estimé sur les données du tableau 10.13 et les intervalles de confiance obtenus par les différentes méthodes décrites.

Tableau 10.15

Méthode	*SMR*	IC à 95 %	Programme
Simple	1,83871	[1,52069 ; 2,15673]	Pr10.23
Log(*SMR*)	1,83871	[1,53020 ; 2,20942]	Pr10.24
Du *RV*	1,86694	[1,55613 ; 2,21277]	Pr10.25

Les résultats des deux méthodes en approximation normale diffèrent légèrement de la méthode du maximum de vraisemblance. Rappelons que cette dernière porte sur le *SMR* du maximum de vraisemblance. C'est cette dernière méthode que nous préférons, en raison des propriétés mêmes de l'estimateur du maximum de vraisemblance.

◆

10.3.3 Tests pour l'homogénéité des *SMR* spécifiques en analyse stratifiée

Nous présentons dans un contexte simplifié deux tests statistiques qui permettent de porter un jugement sur l'homogénéité entre deux ou plusieurs *SMR*.

Ainsi, on suppose que F est un facteur pour lequel on pratique la stratification. Pour chaque strate i de F, on a l'estimation d'un SMR_i : $SMR_i = \dfrac{x_i}{X_{0i}}$, où X_i est une variable binomiale. La situation se présente alors comme au tableau 10.16.

Tableau 10.16

	F				Total
	1	2	…	*K*	
Observé (x_i)	x_1	x_2	…	x_k	X
Attendu (X_{0i})	X_{01}	X_{02}	…	X_{0k}	X_0
SMR_i	SMR_1	SMR_2	…	SMR_k	*SMR*

L'hypothèse à tester est la suivante :

$H_0 : SMR_1 = SMR_2 = \ldots. = SMR_k = SMR = \phi.$

Test en approximation normale (analogue au test de Breslow-Day)

On considère la statistique suivante : $\sum_{i=1}^{k} \frac{\left(x_i - E(X_i \mid \phi)\right)^2}{V(X_i \mid \phi)}$, qui mesure la variation totale des variables binomiales X_i autour de leurs valeurs attendues $E(X_i \mid \phi)$ sous l'hypothèse d'un *SMR* égal à ϕ. Puisque ϕ est en générale estimé à partir des données, cette statistique obéit en bonne approximation à un khi-carré avec $k - 1$ degrés de liberté.

L'estimation des variances est faite sous l'hypothèse testée :

$$V(X_i \mid \phi) = \frac{\phi X_{0i}(n_i - \phi X_{0i})}{n_i}$$

Le test devient alors :

$$\chi^2_{k-1} = \sum_{i=1}^{k} \frac{n_i \left(x_i - \phi X_{0i}\right)^2}{\phi X_{0i}(n_i - \phi X_{0i})}$$

L'avantage de ce test est qu'il peut s'adapter à une quelconque hypothèse sur ϕ.

Test du rapport de vraisemblance

Le test du rapport de vraisemblance pour l'homogénéité des *SMR* a déjà été défini à la section 10.3.1 (tableau 10.12). Il prend la forme suivante :

$2\sum_{i=1}^{k}\left[x_i \log\left(\frac{x_i}{\phi X_{0i}}\right) + (n_i - x_i)\log\frac{n_i - x_i}{n_i - \phi X_{0i}} \right]$ avec $k - 1$ degrés de liberté.

Exemple 10.7

Appliquons les deux tests définis à la section 10.3.3 pour juger de l'homogénéité des *SMR* spécifiques mesurés sur les données du tableau 10.13. Pour pouvoir comparer leurs résultats, nous utilisons pour ces deux tests la valeur de ϕ correspondant à celle du maximum de vraisemblance : $\phi = 1{,}867$.

Test de Breslow-Day

$$\chi^2_{k-1} = \frac{1200\left(15 - 1,867 \times 12\right)^2}{1,867 \times 12(1200 - 1,867 \times 12)} + \ldots$$
$$+ \frac{100\left(25 - 1,867 \times 10\right)^2}{1,867 \times 10(100 - 1,867 \times 10)} = 8,77$$

Pour 3 degrés de liberté, cette valeur du khi-carré correspond à la valeur-*p* de 0,0324 (tableau 10.17).

Test du rapport de vraisemblance

Le test du rapport de vraisemblance donne une valeur de khi-carré de 9,11, correspondant à une valeur-*p* de 0,0279 (tableau 10.17).

Tableau 10.17

Test	Khi-carré	Valeur-*p*	Programme
De Breslow	8,78	0,0324	Pr10.26
Du RV	9,11	0,0279	Pr10.27

Les résultats de ces deux tests, assez similaires, conduisent au rejet de l'hypothèse d'homogénéité des *SMR*.

◆

10.4 RAPPORT DE COTES EN ANALYSE STRATIFIÉE

10.4.1 TESTS STATISTIQUES SUR LES RAPPORTS DE COTES EN APPROXIMATION NORMALE

Reportons-nous au schéma décrit par le tableau 10.1. On s'intéresse alors au rapport de cotes (*RC*) mesurant l'association entre le facteur *X* et la maladie *Y* en analyse stratifiée pour le facteur *F*. Pour une strate *i*, on mesure le rapport de cotes RC_i. À partir d'un système de poids $\{\lambda_i\}$, on veut résumer ces mesures en une mesure globale *RC* telle que $RC = \sum_i \lambda_i RC_i$.

Pour répondre aux problèmes de l'homogénéité des mesures RC_i et de la signification statistique de la mesure globale *RC*, nous utilisons ici aussi la partition du khi-carré total en un khi-carré d'association et un khi-carré d'homogénéité. La partition sera pratiquée sur la transformation logarithmique de *RC*. Une fois les calculs faits, par transformation inverse, on obtient les expressions désirées pour le rapport de cotes.

Si on pose $w_i = \dfrac{1}{V\left[\log RC_i\right]}$, on a les relations suivantes :

$$\begin{cases} \chi^2_{\text{total}} = \sum_i w_i \left(\log RC_i\right)^2 \\ \chi^2_{\text{assoc}} = \dfrac{\left[\sum_i w_i \log RC_i\right]^2}{\sum_i w_i} \\ \chi^2_{\text{homog}} = \sum_i w_i \left(\log RC_i - \overline{\log RC}\right)^2 \end{cases}$$

La variance $V[\log RC_i]$ est celle déduite par la méthode delta :

$$V\left[\log RC_i\right] = V_i = \frac{1}{a_{1i}} + \frac{1}{b_{1i}} + \frac{1}{a_{0i}} + \frac{1}{b_{0i}}$$

De cette partition du khi-carré total, nous retenons deux tests,

- l'un pour l'association globale :

$$\chi^2_1(\text{assoc}) = \frac{\left[\sum_i w_i \log RC_i\right]^2}{\sum_i w_i}$$

- l'autre pour l'homogénéité des mesures spécifiques :

$$\chi^2_{k-1}(\text{homog}) = \sum_i \frac{\left(\log RC_i - \overline{\log RC}\right)^2}{V(\log RC_i)}$$

Comme pour le RP (voir section 10.2.1), on a $V\left(\overline{\log RC}\right) = \dfrac{1}{\sum_i w_i}$.

Aussi, $RC_V = e^{\overline{\log RC}}$ est une moyenne géométrique des RC_i.

Test de Mantel-Haenszel pour l'association

Le test se présente simplement comme $\chi^2_{MH} = \dfrac{\left[\sum_i [a_{1i} - E_0(A_{1i})]\right]^2}{\sum_i V_0(A_{1i})}$.

Sous l'hypothèse H_0, on a

$$E_0(A_{1i}) = \frac{n_{1i} \times m_{1i}}{n_I} \text{ et } V_0(A_{1i}) = \frac{m_{1i} m_{0i} n_{1i} n_{0i}}{n_i^2 (n_i - 1)}.$$

EXEMPLE 10.8

Une étude cas-témoins a été conduite sur la relation entre l'exposition au facteur *X* et la maladie *Y*. Les résultats stratifiés pour l'âge sont décrits au tableau 10.18.

TABLEAU 10.18

	Âge en années			
	20-49		50-79	
	$X = 1$	$X = 0$	$X = 1$	$X = 0$
Cas	80	40	325	55
Témoins	95	285	55	65

MÉTHODE DE PARTITION DU KHI-CARRÉ

Dans le tableau 10.19, nous présentons les valeurs numériques nécessaires aux calculs des tests.

TABLEAU 10.19

Âge (années)	RC_i	log (RC_i)	V_i	w_i	$w_i \log(RC_i)$	$w_i[\log RC_i]^2$
20-49	6,00	1,79	0,0515	19,40	34,77	62,29
50-79	6,98	1,94	0,0548	18,24	35,45	68,90
Total		–	–	37,64	70,22	131,19

La partition du khi-carré nous conduit aux statistiques suivantes :

$$\begin{cases} \chi^2_{\text{total}} = 131{,}19 \\ \chi^2_{\text{assoc}} = \dfrac{70{,}22^2}{37{,}64} = 130{,}98 \qquad p \approx 0 \\ \chi^2_{\text{homog}} = 0{,}21 \qquad\qquad p = 0{,}6416 \end{cases}$$

(PR10.28)

On peut conclure à une faible différence entre les deux mesures spécifiques [χ^2_1 (homog) = 0,2166, pour 1 degré de liberté]. Par contre, la mesure globale log(*RC*) est significativement différente de 0, puisque le χ^2_1 (assoc) est de 130,98. Sans discuter du contexte de l'étude, il pourrait être justifié ici de ne présenter que la mesure globale *RC* correspondante.

Puisque $RC = 1{,}8653$, le RC est donné par : $RC = e^{1{,}8653} = 6{,}4579$.

On peut facilement obtenir le test de Mantel-Haenszel par la procédure FREQ de SAS. Le résultat de ce test pour l'association entre le facteur X et la maladie Y, ajustée pour le facteur F, est de 146,9825. Ce résultat concorde tout à fait avec celui du χ^2_1 (assoc) précédent.

(**PR10.29**)

◆

10.4.2 INTERVALLE DE CONFIANCE DU *RC* PONDÉRÉ EN APPROXIMATION NORMALE

Dans le cadre d'une partition du khi-carré

On rappelle que la mesure RC_V est une moyenne géométrique des différentes mesures spécifiques RC_i (voir section 10.2.1). De même que RC_V est la transformation inverse de la somme pondérée des $\log(RC_i)$, de même ses limites de confiance seront aussi obtenues par transformation inverse des limites de confiance de cette même somme pondérée :

Ainsi :

$$\psi_{\text{inf}} = RC_V \times e^{-z_{\alpha/2}\sqrt{\frac{1}{\sum_i w_i}}}$$

$$\psi_{\text{sup}} = RC_V \times e^{+z_{\alpha/2}\sqrt{\frac{1}{\sum_i w_i}}}$$

où $w_i = \dfrac{1}{V_i}$.

Dans le cadre d'une pondération arithmétique

On s'intéresse à l'intervalle de confiance d'un rapport de cotes pondéré $RC = \sum_i \lambda_i RC_i$. Puisque $V\left[\log\left(\sum_i \lambda_i RC_i\right)\right] \approx \sum_i \lambda_i^2 V\left[\log RC_i\right]$, l'intervalle de confiance de RC se calcule simplement comme :

$$\psi_{\text{inf}} = RC \times e^{-z_{\alpha/2}\sqrt{\sum_i \lambda_i^2 V_i}}$$

$$\psi_{\text{sup}} = RC \times e^{+z_{\alpha/2}\sqrt{\sum_i \lambda_i^2 V_i}}$$

où $V_i = \dfrac{1}{a_{1i}} + \dfrac{1}{b_{1i}} + \dfrac{1}{a_{2i}} + \dfrac{1}{b_{2i}}$.

Si les poids λ_i sont proportionnels à l'inverse des variances V_i, $\lambda_i \propto 1/V_i$, alors la variance $V[\log RC]$ est estimée simplement par $\dfrac{1}{\sum_i \dfrac{1}{V_i}}$.

Nous décrivons ci-dessous, suivant les notations du tableau 10.1, les poids $\{\lambda_i\}$ les plus souvent utilisés. En posant $\lambda_i = \dfrac{w_i}{\sum_i w_i}$, on a

- le poids de Mantel-Haenszel, $w_i = \dfrac{b_{1i} \times a_{2i}}{n_i}$, qui donne le RC_{MH} ;
- le poids de type *SMR*, $w_i = \dfrac{b_{1i} a_{2i}}{b_{2i}}$, qui donne le RC_a ;
- le poids proportionnel à la distribution du facteur *F* chez les cas non exposés, $w_i = b_{1i}$, qui donne le RC_S.

Exemple 10.9

Considérons les données du tableau 10.18. Pour chacun des systèmes de poids décrits précédemment, nous présentons dans le tableau 10.20 la mesure *RC* pondérée, son logarithme, la variance du logarithme et l'intervalle de confiance normal à 95 % du *RC*.

Tableau 10.20

Pondération	*RC*	*V*[log *RC*]	IC à 95 %	Programme
RC_V	6,4579	0,026565	[4,69 ; 8,89]	Pr10.30
RC_{MH}	6,4359	0,026746	[4,67 ; 8,87]	Pr10.31
RC_a	6,7645	0,035681	[4,67 ; 9,80]	
RC_S	6,3606	0,028042	[4,58 ; 8,83]	

Nous remarquons que les *RC* obtenus diffèrent peu d'un système de poids à l'autre. Cela est dû principalement à la grande homogénéité entre les deux mesures spécifiques. La mesure *RC* ajustée, quel que soit le système de poids, doit obligatoirement se situer entre ces deux valeurs spécifiques : $RC_1 = 6{,}00$ et $RC_2 = 6{,}98$.

◆

10.4.3 TESTS STATISTIQUES ET INTERVALLE DE CONFIANCE PAR LA MÉTHODE DU RAPPORT DE VRAISEMBLANCE

Dans le cadre de la méthode du rapport de vraisemblance, le rapport de cotes est traité à l'aide de la transformation logit de la proportion. La loi de distribution sous-jacente est la loi binomiale et la fonction de lien, le logit (c'est la régression logistique). Si $\pi(X,F)$ désigne la proportion comme une fonction des variables indépendantes X et F, la transformation logit se définit comme : $\text{logit}\,\pi(X,F) = \log\dfrac{\pi(X,F)}{1-\pi(X,F)}$.

On considère alors les deux modèles suivants :

Modèle 1 : $\text{logit}\ \pi(X,F) = \alpha_1 + \beta_{11}X + \beta_{12}F$

Modèle 2 : $\text{logit}\ \pi(X,F) = \alpha_2 + \beta_{21}X + \beta_{22}F + \beta_{23}XF$

Dans le premier modèle, le coefficient β_{11} représente le logarithme du *RC*, ajusté pour le facteur F. Ainsi, $e^{\beta_{11}} = RC$ est ajusté pour le facteur F. Pour porter un jugement sur la signification statistique du rapport de cotes ajusté pour le facteur F, il suffit alors d'appliquer le test du rapport de vraisemblance au coefficient β_{11}.

Dans le second modèle, le coefficient β_{23} représente le terme d'interaction multiplicative. Il mesure alors l'hétérogénéité des RC_i à travers les strates de F. Si X et F sont dichotomiques, alors $e^{\beta_{23}} = \dfrac{RC_1}{RC_0}$. On peut écrire aussi $e^{\beta_{23}} = \dfrac{RC_{11}}{RC_{10} \times RC_{01}}$, où RC_{ij} représente le rapport de cotes issu de la comparaison entre la catégorie ($X = i$, $F = j$) et la catégorie de référence ($X = 0$, $F = 0$).

Les intervalles de confiance par la méthode du rapport de vraisemblance, pour le *RC* ou pour l'interaction, s'obtiennent en solutionnant une équation de vraisemblance de la forme $\left[L - L(\beta)\right] - 0,5\chi^2_{1,1-\alpha} = 0$.

EXEMPLE 10.10

Utilisons les données du tableau 10.18. Le facteur X correspond à l'exposition et F à l'âge. Les résultats nous conduisent aux constatations suivantes.

On remarque d'abord une faible hétérogénéité : $\beta_{23} \approx 0,1518$. Le test du rapport de vraisemblance sur ce coefficient donne un khi-carré de 0,22 pour une valeur-p de 0,6418. Ces résultats sont donc assez compatibles avec

l'hypothèse de l'homogénéité des RC_i. Par ailleurs, le coefficient β_{11} du premier modèle est égal à 1,8658 et le test du rapport de vraisemblance sur ce coefficient conduit à un khi-carré de 138,41 pour un $p \approx 0$. Ce résultat suggère un rapport de cotes *RC* très significativement différent de 1. On rappelle que le test sur l'association de l'exemple traité à l'exemple 10.6 donnait un khi-carré de 130,98.

Le rapport de cotes est estimé à $e^{1,8658} = 6,46$, avec un intervalle de confiance à 95 % de [4,705 ; 8,927]. Ces résultats sont similaires à ceux obtenus par la méthode des poids proportionnels à l'inverse des variances, à l'exemple 10.7. (**PR10.32**)

◆

10.4.4 TEST DE BRESLOW POUR L'HOMOGÉNÉITÉ DES *RC* EN ANALYSE STRATIFIÉE

Pour juger de l'homogénéité des mesures spécifiques, on peut aussi utiliser le test basé sur les variables hypergéométriques A_{1i} :

$\chi^2_{k-1} = \sum_i \frac{\left(a_{1i} - E(A_{1i} \mid \psi)\right)^2}{V(A_{1i} \mid \psi)}$ où $E(A_{1i} \mid \psi)$ et $V(A_{1i} \mid \psi)$ sont respectivement la valeur attendue et la variance de la variable A_{1i} sous l'hypothèse d'un $RC = \psi$. Pour une estimation ψ donnée, la valeur attendue de $E(A_{1i} \mid \psi)$ correspondante est une valeur de A_i telle que : $\frac{A_i D_i}{B_i C_i} = \psi$, où $A_i = E(A_{1i} \mid \psi)$, $B_i = m_{1i} - A_i$, $C_i = n_{1i} - A_i$ et $D_i = n_{0i} - m_{1i} + A_i$. Pour déterminer cette valeur, il suffit de résoudre pour A_i l'équation quadratique suivante : $(\psi - 1)A_i^2 - [(n_{0i} - m_{1i}) + \psi(n_{1i} + m_{1i})]A_i + \psi n_{1i} m_{1i} = 0.$

Des deux solutions pour A_i, on prend celle qui se situe dans le domaine de variation de la variable hypergéométrique A_{1i}. Comme on connaît alors les valeurs attendues A_i, B_i, C_i et D_i, la variance $V(A_{1i} \mid \psi)$ est donnée par

$$V(A_{1i} \mid \psi) = \left[\frac{1}{A_i} + \frac{1}{B_i} + \frac{1}{C_i} + \frac{1}{D_i} \right]^{-1}$$

La valeur de ψ à utiliser est théoriquement l'estimation du maximum de vraisemblance. En analyse stratifiée, cette estimation ne peut pas s'obtenir sans utiliser les méthodes itératives. Ces méthodes sont déjà inscrites dans certaines procédures de SAS, comme LOGISTIC ou GENMOD. Par ailleurs, l'estimation de Mantel-Haenszel, plus simple de calcul, donne aussi de bons résultats.

EXEMPLE 10.11

Dans le tableau 10.21, nous résumons les résultats du test d'homogénéité de Breslow-Day suivant les deux estimateurs : RC_{RV} (du rapport de vraisemblance) et RC_{MH}.

TABLEAU 10.21

Estimateur du *RC*	Strate	a_{1i}	$E(A_{1i})$	$V(A_{1i})$	$\frac{[a_{1i} - E(A_{1i})]^2}{V(A_{1i})}$
RC_{RV}	1	80	81,425	19,0902	0,10642
	2	325	323,575	18,4281	0,11024
	Total	–	–	–	0,21666*
RC_{MH}	1	80	81,351	19,1072	0,09546
	2	325	323,502	18,4371	0,12164
	Total	–	–	–	0,21709**

* $p = 0{,}6416$ ** $p = 0{,}6413$

Nous remarquons que ces deux tests ont des résultats très similaires, qui sont, par ailleurs, tout à fait compatibles avec l'hypothèse d'homogénéité des RC_i. Ces résultats sont tout à fait concordants avec les résultats des tests pratiqués dans le cadre de la partition du khi-carré ($\chi^2 = 0{,}2166$) et du test du rapport de vraisemblance ($\chi^2 = 0{,}217$). (**PR10.33**)

◆

CHAPITRE

11

LES MESURES FRACTIONNAIRES EN ANALYSE STRATIFIÉE

En utilisant les expressions développées au chapitre 7 pour les mesures fractionnaires, il est relativement facile de déterminer les intervalles de confiance de ces mesures en analyse stratifiée. Dans les sections qui suivent, nous présentons quelques méthodes simples de calcul des intervalles de confiance pour les fractions attribuables chez les exposés et de population et pour les fractions prévenues chez les exposés et de population, en analyse stratifiée.

11.1 FRACTION ATTRIBUABLE

Considérons les données d'une étude à visée étiologique. On veut alors estimer la fraction attribuable du facteur X pour la maladie Y. Le facteur F est une variable de stratification.

Pour la strate i, les données peuvent être décrites comme celles du tableau 11.1. Ces données peuvent être en unités de personnes-temps ou de personnes. On rappelle que sous les conditions de marges fixes, la variable A_{1i} est assimilable à une variable binomiale dans le premier cas, ou à une variable hypergéométrique dans le second cas.

TABLEAU 11.1

Facteur F		$X = 1$	$X = 0$	Total
Strate i	$Y = 1$	a_{1i}	a_{0i}	m_{1i}
	$Y = 0$*	b_{1i}	b_{0i}	m_{0i}
	Total	n_{1i}	n_{0i}	n_i

* Si les données sont en personnes-temps, cette ligne n'est pas définie.

Alors, pour la strate i, les fractions attribuables chez les exposés FA_{1i} et de population (ou totale) FA_{ti} sont décrites respectivement comme suit :

$$FA_{1i} = \frac{R_{1i} - R_{0i}}{R_{1i}} = \frac{RR_i - 1}{RR_i} \tag{11.1}$$

et

$$\begin{aligned} FA_{ti} &= \frac{R_{ti} - R_{0i}}{R_{ti}} \\ &= p_{ci} \times \frac{RR_i - 1}{RR_i} \\ &= p_{c_i} \times FA_{1i} \end{aligned} \tag{11.2}$$

où les RR_i sont des rapports de taux ou de risques et le $p_{ci} = \frac{a_{1i}}{m_{1i}}$ correspond à la proportion d'exposés parmi la totalité des cas sur la strate i.

11.1.1 FRACTION ATTRIBUABLE CHEZ LES EXPOSÉS

La fraction attribuable chez les exposés, FA_1, s'exprime comme la somme pondérée des fractions attribuables spécifiques chez les exposés FA_{1i}, les poids étant proportionnels aux effectifs des cas exposés :

$$FA_1 = \sum_i \lambda_i FA_{1i} \tag{11.3}$$

où $\lambda_i = \frac{a_{1i}}{\sum_i a_{1i}} = \frac{a_{1i}}{a_1}$.

L'estimation ponctuelle de FA_1 peut également être décrite à partir du RR_a ajusté par la méthode du *SMR* :

$$FA_1 = \frac{RR_a - 1}{RR_a} \tag{11.4}$$

Pour le calcul de l'intervalle de confiance de FA_1, nous proposons trois approches en approximation normale :

1. la méthode standard basée sur le RR_a ; cette approche est équivalente à l'application d'une transformation (standard) Φ_0 à FA_1, comme : $\Phi_0(FA_1) = \frac{1}{1 - FA_1} = RR_a$; on remarque que cette transformation est croissante ;
2. l'utilisation d'une transformation Φ_1 définie par $\Phi_1(FA_1) = \log\left(\frac{1}{1 - FA_1}\right) = \log(RR_a)$; cette transformation est aussi croissante ;
3. le calcul direct de l'intervalle de confiance dans le cas où l'étude est basée sur un échantillon aléatoire de la population.

(Pour la notion de transformation, voir au chapitre 2 les sections 2.10 et 2.11.)

Méthode standard basée sur le RR_a : transformation Φ_0

La méthode est relativement simple. En utilisant la relation inverse de celle décrite à la relation (11.4), on déduit assez facilement que $RR_a = \frac{1}{1 - FA_1}$.

On remarque alors que le RR_a est une transformation croissante de FA_1. Des limites de confiance du RR_a, on déduit celles de FA_1 par transformation inverse :

$$(FA_1)_{\text{inf}} = \frac{(RR_a)_{\text{inf}} - 1}{(RR_a)_{\text{inf}}}$$

$$(FA_1)_{\text{sup}} = \frac{(RR_a)_{\text{sup}} - 1}{(RR_a)_{\text{sup}}}$$

Méthode basée sur la transformation Φ_1

On considère la transformation Φ_1 suivante :

$$\Phi_1(FA_1) = \log\left(\frac{1}{1-FA_1}\right) \tag{11.5}$$

Ainsi, pour chaque mesure spécifique FA_{1i}, on a $\Phi_1(FA_{1i}) = \log RR_i$. De là, on déduit que

$$V\left[\Phi_1(FA_1)\right] = \sum_i \lambda_i^2 V\left[\Phi_1(FA_{1i})\right] = \sum_i \lambda_i^2 V\left[\log RR_i\right]$$

Les limites de confiance sont calculées pour la mesure transformée : $\Phi_{1\text{inf}}$ et $\Phi_{1\text{sup}}$.

$$\begin{aligned}\Phi_{1\text{inf}} &= \Phi_1(FA_1)_{\text{inf}} \\ &= \Phi_1(FA_1) - z_{\alpha/2}\sqrt{V\left[\Phi_1(FA_1)\right]} \\ \Phi_{1\text{sup}} &= \Phi_1(FA_1)_{\text{sup}} \\ &= \Phi_1(FA_1) + z_{\alpha/2}\sqrt{V\left[\Phi_1(FA_1)\right]}\end{aligned}$$

Par transformation inverse, on obtient celles de FA_1 :

$$\left(FA_1\right)_{\text{inf}} = \frac{e^{\Phi_{1\text{inf}}} - 1}{e^{\Phi_{1\text{inf}}}}$$

$$\left(FA_1\right)_{\text{sup}} = \frac{e^{\Phi_{1\text{sup}}} - 1}{e^{\Phi_{1\text{sup}}}}$$

Calcul direct pour une étude de population

Dans le cas d'une étude sur une population, pour une strate *i* particulière, en conditionnant sur le nombre de cas, la fraction attribuable chez les exposés peut être décrite comme :

$$FA_{1i} = \frac{n_i}{n_{0i}}\left[\frac{a_{1i} - E_0(A_{1i})}{a_{1i}}\right] = \frac{a_{1i} - E_0(A_{1i})}{p_{ci}\left[m_{1i} - E_0(A_{1i})\right]} \tag{11.6}$$

où $E_0(A_{1i}) = \dfrac{m_{1i}n_{1i}}{n_i}$.

La fraction attribuable pondérée chez les exposés est donc de la forme :

$$FA_1 = \sum_i \lambda_i FA_{1i} = \sum_i \frac{a_{1i}}{a_1}\left\{\frac{n_i}{n_{0i}}\left[\frac{a_{1i} - E_0(A_{1i})}{a_{1i}}\right]\right\}$$

$$= \frac{1}{a_1}\sum_i \frac{n_i}{n_{0i}}\left[a_{1i} - E_0(A_{1i})\right] \qquad (11.7)$$

La variable A_{1i}, dont a_{1i} est une observation, est une variable binomiale ou hypergéométrique suivant le contexte.

La variance de FA_1 est alors estimée simplement par :

$$V(FA_1) = \frac{1}{a_1^2}\sum_i \frac{n_i^2}{n_{0i}^2} V(A_{1i})$$

Pour le niveau $100(1 - \alpha)\,\%$, les limites de confiance de FA_1 sont alors facilement déduites comme :

$$(FA_1)_{\text{inf}} = FA_1 - z_{\alpha/2}\sqrt{V(FA_1)}$$

$$(FA_1)_{\text{sup}} = FA_1 + z_{\alpha/2}\sqrt{V(FA_1)}$$

Exemple 11.1

Les données utilisées sont celles de l'étude de Doll et Hill conduite chez les médecins britanniques, portant sur les effets de la cigarette sur la santé (tableau 11.2).

Tableau 11.2

Âge (en années)	Fumeurs		Non-fumeurs		Rapport de taux	Fraction FA_{1i}
	Décès	Pers-an.	Décès	Pers-an.		
35-44	32	52 407	2	18 790	5,74	0,8258
45-54	104	43 248	12	10 673	2,14	0,5327
55-64	206	28 612	28	5 710	1,47	0,3197
65-74	186	12 663	28	2 585	1,36	0,2647
75-84	102	5 317	31	1 462	0,90	–0,1111
Total	630	142 247	101	39 220		–

Sur ces données, la fraction attribuable chez les exposés est estimée comme :

$$FA_1 = \frac{\sum_i a_{1i} FA_{1i}}{\sum_i a_{1i}}$$
$$= \frac{32 \times 0,8258 + 104 \times 0,5327 + ... + 102 \times (-0,1111)}{630}$$
$$= 0,2946$$

Pour le calcul de cette fraction, on peut s'interroger sur la pertinence d'inclure la dernière catégorie des 75-84 ans. Nous laissons aux épidémiologistes le soin d'en décider.

Au tableau 11,3, nous comparons les résultats des trois méthodes de calcul de l'intervalle de confiance de la fraction attribuable FA_1.

TABLEAU 11.3

Méthode	Fraction attribuable	Intervalle de confiance	Programme
Transformation Φ_0	0,29459	[0,12799 ; 0,42935]	PR11.1
Transformation Φ_1	0,29459	[0,11851 ; 0,43549]	
Méthode directe	0,29459	[0,13748 ; 0,45169]	

Les résultats des deux premières méthodes sont assez concordants. Les écarts observés sont principalement dus à l'utilisation de transformations et de poids différents pour l'estimation de la variance. La méthode basée sur la transformation Φ_0, $\Phi_0(FA_1) = \frac{1}{1 - FA_1}$ implique directement l'utilisation des poids inhérents à la mesure pondérée RR_a. La seconde méthode, par contre, basée sur la transformation décrite par la relation (11.5), utilise dans le calcul de la variance d'autres poids qui nous apparaissent plus cohérents avec la définition même de la mesure FA_1.

Enfin, on remarque que la troisième méthode conduit à un intervalle centré sur la mesure. Bien qu'elle conduise à des résultats légèrement différents de la deuxième méthode, on note que leurs intervalles ont des étendues tout à fait similaires.

◆

11.1.2 FRACTION ATTRIBUABLE DE POPULATION

Pour le calcul d'un intervalle de confiance de la fraction attribuable de population, nous nous restreignons aux études conduites sur une population ou sur un échantillon de celle-là. Ce contexte permet de définir des calculs relativement simples : l'estimation des fractions et de leur variance est directe.

La fraction attribuable globale ou pondérée de population FA_t (ou plus simplement FA) s'exprime comme la somme pondérée des fractions attribuables spécifiques de population FA_i. Les poids sont proportionnels aux effectifs des cas :

$$FA = \sum_i \lambda_i FA_i \tag{11.8}$$

où $\lambda_i = \dfrac{m_{1i}}{\sum_i m_{1i}} = \dfrac{m_{1i}}{m_1}$.

Rappelons que cette fraction peut être décrite aussi en utilisant le RR_a :

$$FA = p_c \frac{RR_a - 1}{RR_a} \tag{11.9}$$

où $p_c = \dfrac{\sum_i a_{1i}}{\sum_i m_{1i}} = \dfrac{a_1}{m_1}$.

Sous les conditions de marges fixes, à partir de la relation (11.6), pour une strate particulière i, la fraction attribuable de population peut s'écrire comme :

$$FA_i = \frac{a_{1i} - E_0(A_{1i})}{m_{1i} - E_0(A_{1i})} \tag{11.10}$$

Cette condition des marges fixes permet de réduire la dépendance de la mesure FA_i à une simple variable A_{1i}, binomiale ou hypergéométrique suivant le type de données. Ainsi, la variance de FA_i, peut simplement être déduite comme :

$$\begin{aligned} V(FA_i) &= V\left[\frac{A_{1i} - E_0(A_{1i})}{m_{1i} - E_0(A_{1i})}\right] \\ &= \left[\frac{1}{m_{1i} - E_0(A_{1i})}\right]^2 V(A_{1i}) \end{aligned} \tag{11.11}$$

De cette dernière relation, on déduit la variance pour la fraction attribuable *FA* pondérée de population :

$$\begin{aligned} V(FA) &= V\left[\sum_i \lambda_i FA_i\right] \\ &= \sum_i \lambda_i^2 V(FA_i) \\ &= \sum_i \left[\frac{m_{1i}}{m_1}\right]^2 \left[\frac{1}{m_{1i} - E_0(A_{1i})}\right]^2 V(A_{1i}) \\ &= \frac{1}{m_1^2} \sum_i \left[\frac{n_i}{n_{0i}}\right]^2 V(A_{1i}) \end{aligned} \tag{11.12}$$

où la variance $V(A_{1i})$ est celle d'une variable binomiale ou hypergéométrique. Si la fraction attribuable est calculée à partir d'un rapport de taux, A_{1i} est une variable binomiale ; la variance $V(A_{1i})$ peut alors être estimée comme :

$$V(A_{1i}) = \frac{a_{1i}(m_{1i} - a_{1i})}{m_{1i}} \tag{11.13}$$

Si la fraction attribuable est calculée à partir d'un rapport de proportions, A_{1i} est une variable hypergéométrique. Dans ce cas, on peut estimer la variance $V(A_{1i})$ en utilisant l'expression

$$V(A_{1i} \mid \psi_i) = \left[\frac{1}{A_i} + \frac{1}{B_i} + \frac{1}{C_i} + \frac{1}{D_i}\right]^{-1} \tag{11.14}$$

où ψ_i est le rapport de cotes empirique ou estimé par la méthode du maximum de vraisemblance et $A_i = E(A_{1i} \mid \psi_i)$, $B_i = m_{1i} - A_i$, $C_i = n_{1i} - A_i$ et $D_i = n_{0i} - m_{1i} + A_i$ sont les valeurs attendues correspondant aux cellules du tableau 11.1 sous l'hypothèse ψ_i. Sous l'hypothèse de l'uniformité des rapports de cotes à travers les strates, ψ_i peut être estimé à partir du *RC* pondéré de Mantel-Haenszel ou du maximum de vraisemblance.

EXEMPLE 11.2

En utilisant les données du tableau 11.2, nous allons calculer l'intervalle de confiance de la fraction attribuable pondérée de population à partir de la méthode directe décrite précédemment. Dans le tableau 11.4, nous fournissons certains résultats utiles au calcul de l'intervalle de confiance de *FA*.

TABLEAU 11.4

Âge (années)	m_{1i}	FA_i	λ_i	$\left[\frac{n_i}{n_{0i}}\right]^2$	$V(A_{1i})$	$\left[\frac{n_i}{n_{0i}}\right]^2 \times V(A_{1i})$
35-44	34	0,777	0,05	14,3572	1,88	27,025
45-54	116	0,477	0,16	25,5237	10,76	274,599
55-64	234	0,281	0,32	36,1304	24,65	890,599
65-74	214	0,228	0,29	34,7940	24,34	846,763
75-84	133	– 0,081	0,18	21,5000	23,77	511,148
Total	731	0,254*	1,00	–	–	2550,134

* Fonction attribuable pondérée.

La fraction attribuable *FA* est estimée à 0,254 :

$$FA = \frac{34 \times 0,777 + 116 \times 0,477 + 234 \times 0,281 + 214 \times 0,228 + 133 \times (-0,081)}{731} = 0,254$$

En utilisant l'expression 11.12, on calcule la variance de la fraction attribuable globale comme :

$$\begin{aligned} V(FA) &= \frac{1}{m_1^2} \sum_i \left[\frac{n_i}{n_{0i}}\right]^2 V(A_{1i}) \\ &= \frac{1}{731^2} \times 2550,134 \\ &= 0,004772 \end{aligned}$$

Les limites de confiance à 95 % pour la fraction attribuable pondérée *FA* sont alors données par :

$$FA_{\text{inf}} = 0,254 - 1,96 \times \sqrt{0,004772} = 0,118$$
$$FA_{\text{sup}} = 0,254 + 1,96 \times \sqrt{0,004772} = 0,389$$

(**PR11.2**)

Les limites de confiance obtenues sont légèrement différentes de celles présentées dans Rothman et Greenland[1] pour le même exemple : 0,108 et 0,377. Celles que nous avons calculées, contrairement aux secondes, sont centrées sur la mesure. Par ailleurs, il est intéressant de noter que les étendues des intervalles par l'une et l'autre des méthodes sont très similaires : (0,389 – 0,118) = 0,271 versus (0,377 – 0,108) = 0,269.

Nous soulignons de nouveau que la fraction attribuable est légèrement négative sur la strate des 75-84 ans. Cette observation peut être conforme à l'hypothèse d'un effet nul du facteur sur la strate. Si on restreint les calculs de la

[1] Rothman, K.J. et S. Greenland, *Modern Epidemiology,* 2e éd., Philadelphie, Lippincott-Raven,1998, p. 296.

fraction attribuable aux personnes dont l'âge est inférieur à 75 ans, la fraction *FA* prendra une valeur légèrement supérieure à la précédente : $FA = 0{,}328$ (tableau 11.5).

TABLEAU 11.5

Âge (années)	m_{1i}	FA_i	λ_i	$\left[\frac{n_i}{n_{0i}}\right]^2$	$V(A_{1i})$	$\left[\frac{n_i}{n_{0i}}\right]^2 \times V(A_{1i})$
35-44	34	0,777	0,06	14,3572	1,88	27,025
45-54	116	0,477	0,19	25,5237	10,76	274,599
55-64	234	0,281	0,39	36,1304	24,65	890,599
65-74	214	0,228	0,36	34,7940	24,34	846,763
Pondérée	598	0,328	1,00	–	–	2038,986

La variance de *FA* est estimée à 0,0057. Les limites de confiance à 95 % de cette fraction attribuable sont :

$$FA_{\text{inf}} = 0{,}178$$
$$FA_{\text{sup}} = 0{,}476$$

◆

11.2 FRACTION PRÉVENUE

On s'intéresse à un facteur *X* protecteur de la maladie *Y*. On veut alors estimer la fraction prévenue de ce facteur en tenant compte d'une variable de stratification *F*.

Pour la strate *i*, les données peuvent être décrites comme celles du tableau 11.6 (reprise du tableau 11.1). Il peut s'agir de données de personnes-temps ou de personnes. On souligne que sous les conditions de marges fixes, la variable A_{0i} est assimilable à une variable binomiale dans le premier cas, ou à une variable hypergéométrique dans le second cas.

TABLEAU 11.6

Facteur *F*		*X* = 1	*X* = 0	Total
Strate *i*	*Y* = 1	a_{1i}	a_{0i}	m_{1i}
	Y = 0*	b_{1i}	b_{0i}	m_{0i}
	Total	n_{1i}	n_{0i}	n_i

* Si les données sont en personnes-temps, cette ligne n'est pas définie.

Pour la strate *i*, les fractions prévenues chez les exposés FP_{1i} et de population FP_{ti} sont décrites respectivement comme suit :

$$FP_{1i} = \frac{R_{0i} - R_{1i}}{R_{0i}} = \frac{Ef_i - 1}{Ef_i}$$

et

$$FP_{Pi} = \frac{R_{0i} - R_{ti}}{R_{0i}} = \frac{p_{ci}[Ef_i - 1]}{p_{ci}[Ef_i - 1] + 1}$$

où $Ef_i = \frac{1}{RR_i}$ et $p_{ci} = \frac{a_{1i}}{m_{1i}}$ correspondent respectivement à l'effet protecteur de X et à la proportion de cas exposés parmi tous les cas sur la strate i.

11.2.1 FRACTION PRÉVENUE CHEZ LES EXPOSÉS

La fraction prévenue chez les exposés FP_1, pondérée pour le facteur F, s'exprime comme la somme pondérée des fractions attribuables spécifiques chez les exposés FP_{1i}, les poids étant proportionnels aux effectifs des cas potentiels exposés : $FP_1 = \sum_i \lambda_i FP_{1i}$

où $\lambda_i = \frac{a_{1i}Ef_i}{\sum_i a_{1i}Ef_i} = \frac{a_{0i}n_{1i}/n_{0i}}{\sum_i a_{0i}n_{1i}/n_{0i}} = \frac{a^*_{1i}}{\sum_i a^*_{1i}} = \frac{a^*_{1i}}{a^*_1}$.

La valeur a^*_{1i} représente le nombre de cas potentiels ou attendus parmi les exposés sous l'hypothèse que le risque de la maladie dans ce groupe soit le même que dans celui des non-exposés. Le nombre a^*_1 représente donc la totalité des cas potentiels parmi les exposés, sous cette même hypothèse. Pour simplifier, on désigne ce nombre de cas potentiels par $E(A_{1i})$.

L'estimation ponctuelle de FP_1 peut également être décrite à partir de Ef_a ajustée par la méthode du RR_a^{-1} (c'est-à-dire $Ef_a = \frac{1}{RR_a}$). Dans ce cas, $FP_1 = \frac{Ef_a - 1}{Ef_a}$.

Pour le calcul de l'intervalle de confiance de FP_1, nous proposons trois approches en approximation normale :

1. La méthode standard basée sur la mesure Ef_a. Cette approche est équivalente à l'application de la transformation (standard) Φ_0 à FP_1, comme $\Phi_0(FP_1) = \frac{1}{1 - FP_1} = Ef_a$.

2. L'utilisation de la transformation Φ_2 définie comme $\Phi_2(FP_1) = \log(1 - FP_1) = \log(RR_a)$, sur la fraction prévenue exprimée comme la somme pondérée des fractions prévenues spécifiques, $FP_{1i} = \dfrac{Ef_i - 1}{Ef_i}$. Cette transformation est décroissante.
3. Le calcul direct de l'intervalle de confiance dans le cas où l'étude est basée sur un échantillon aléatoire de la population.

Méthode basée sur le E_{fa} : transformation Φ_0

La méthode est ici analogue à celle utilisée pour la fraction attribuable basée sur le RR_a (section 11.1.1). On déduit assez facilement que $Ef_a = \dfrac{1}{1 - FP_1}$. Alors, si les limites de confiance $(Ef_a)_{\text{inf}}$ et $(Ef_a)_{\text{sup}}$ de Ef_a sont connues, celles de FP_1 sont alors déduites par transformation inverse :

$$(FP_1)_{\text{inf}} = \frac{(EF_a)_{\text{inf}} - 1}{(Ef_a)_{\text{inf}}}$$

$$(FP_1)_{\text{sup}} = \frac{(Ef_a)_{\text{sup}} - 1}{(Ef_a)_{\text{sup}}}$$

De façon équivalente, ces limites peuvent être décrites en fonction du RR_a. En effet, puisque $FP_1 = 1 - RR_a$, les limites de confiance s'expriment comme

$$(FP_1)_{\text{inf}} = 1 - \left(RR_a\right)_{\text{sup}}$$

$$(FP_1)_{\text{sup}} = 1 - \left(RR_a\right)_{\text{inf}}$$

Méthode basée sur la transformation Φ_2

On veut calculer l'intervalle de confiance de la fraction prévenue globale chez les exposés : FP_1.

Aussi, pour chaque mesure spécifique FP_{1i}, on a $\Phi_2(FP_{1i}) = \log RR_i$.

Ainsi, $V\left[\Phi_2(FP_1)\right] = \sum_i \lambda_i^2 V\left[\Phi_2(FP_{1i})\right] = \sum_i \lambda_i^2 V\left[\log RR_i\right]$.

Les limites de confiance sont d'abord calculées sur la transformation $\Phi_2(FP_1)$ de la fraction étiologique.

$$\left(\Phi_2\right)_{\text{inf}} = \Phi_2(FP_1) - z_{\alpha/2}\sqrt{V[\Phi_2(FP_1)]}$$
$$\left(\Phi_2\right)_{\text{sup}} = \Phi_2(FP_1) + z_{\alpha/2}\sqrt{V[\Phi_2(FP_1)]}$$

Par transformation inverse Φ_2^{-1} de Φ_2, on obtient les limites de confiance de FP_1.

$$\left(FP_1\right)_{\text{inf}} = 1 - e^{(\Phi_2)_{\text{sup}}}$$
$$\left(FP_1\right)_{\text{sup}} = 1 - e^{(\Phi_2)_{\text{inf}}}$$

Calcul direct pour une étude de population

Dans le cas d'une étude sur une population, pour une strate *i* particulière, la fraction prévenue chez les exposés peut être décrite comme : $FP_{1i} = \dfrac{n_i}{n_{1i}}\left[\dfrac{a_{0i} - E_0(A_{0i})}{a_{0i}}\right]$ (voir la relation 7.10 du chapitre 7), où $E_0(A_{0i}) = \dfrac{n_{0i}m_{1i}}{n_i}$ est assimilable à la valeur attendue de la variable A_{0i} sous H_0. Cette variable, pour laquelle a_{0i} est une observation, est une variable binomiale ou hypergéométrique suivant le contexte.

La fraction prévenue chez les exposés pondérée pour le facteur *F* est donc de la forme :

$$FP_1 = \sum_i \lambda_i FP_{1i} = \left\{\sum_i E(A_{1i})\left[\frac{n_i}{n_{1i}}\left(\frac{a_{0i} - E_0(A_{0i})}{a_{0i}}\right)\right]\right\} \Big/ E(A_1)$$
$$= \left\{\sum_i \frac{n_i}{n_{0i}}\left(a_{0i} - E_0(A_{0i})\right)\right\} \Big/ E(A_1)$$

où $E(A_1) = \sum_i E(A_{1i})$.

La variance de FP_1 est alors estimée simplement par :

$$V(FP_1) = \frac{1}{\left[E(A_1)\right]^2}\sum_i \frac{n_i^2}{n_{0i}^2}V(A_{0i})$$

Si la fraction prévenue est calculée à partir d'un rapport de taux, A_{0i} est une variable binomiale et sa variance est estimée par $V(A_{0i}) = \dfrac{a_{0i} \times a_{1i}}{m_{1i}}$.

Si la fraction prévenue est calculée à partir d'un rapport de proportions, A_{0i} est une variable hypergéométrique. Dans ce cas, la variance $V(A_{0i})$ peut être estimée suivant la relation (11.14).

$$V(A_{0i} \mid \psi_i) = \left[\frac{1}{A_i} + \frac{1}{B_i} + \frac{1}{C_i} + \frac{1}{D_i} \right]^{-1}$$

On peut aussi tout simplement estimer la variance comme :

$$V(A_{0i}) = \left[\frac{1}{a_{0i}} + \frac{1}{a_{1i}} + \frac{1}{b_{0i}} + \frac{1}{b_{1i}} \right]^{-1}$$

L'intervalle de confiance de FP_1 de niveau $100(1 - \alpha)\,\%$ est alors facilement déduit comme :

$$FP_1 \pm z_{\alpha/2} \sqrt{V(FP_1)}$$

De façon explicite, on a :

$$FP_{1\,\text{inf}} = FP_1 - z_{\alpha/2} \sqrt{V(FP_1)}$$

$$FP_{1\,\text{sup}} = FP_1 + z_{\alpha/2} \sqrt{V(FP_1)}$$

EXEMPLE 11.3

Soit une étude (fictive) portant sur l'effet protecteur d'un facteur *X* contre la maladie *Y*. Un groupe de 2000 sujets exposés à ce facteur ($X = 1$) a été comparé à un groupe de 1000 sujets non exposés ($X = 0$). Les groupes ont été observés pour une période précise de risque de la maladie *Y*. Sur cette période, pour chaque groupe, on a recensé le nombre de nouveaux cas et calculé l'incidence cumulative de la maladie *Y*. Le tableau 11.7 décrit les données stratifiées pour le facteur *F* (en l'occurrence l'âge) qu'il faut contrôler.

TABLEAU 11.7

	$X = 1$			$X = 0$		
F **(en années)**	**Décès**	**Personnes à risque**	**Incidence cumulative**	**Décès**	**Personnes à risque**	**Incidence cumulative**
35-44	3	900	0,003	4	300	0,013
45-54	8	500	0,048	24	200	0,120
55-64	12	250	0,048	24	200	0,120
65-74	24	200	0,120	36	150	0,240
75-84	30	150	0,200	30	100	0,300
Total	77	2000	0,0385	106	1000	0,106

Dans le tableau 11.8, nous présentons les mesures RR_i, leurs mesures inverses Ef_i, les fractions prévenues FP_{1i} et les cas potentiels $E(A_{1i})$ chez les exposés à X.

TABLEAU 11.8

F (en années)	RR_i	Ef_i	FP_{1i}	$E(A_{1i})$	$V(A_{0i})$
35-44	0,25	4,0	0,75	12	1,71
45-54	0,33	3,0	0,67	24	4,80
55-64	0,40	2,5	0,60	30	8,00
65-74	0,50	2,0	0,50	48	14,40
75-84	0,67	1,5	0,33	45	15,00
Total	0,4843[a]	2,0649[b]	0,5157[c]	159	–

[a] Valeur de RR_a ; [b] valeur de Ef_a ; [c] valeur de FP_1.

La fraction prévenue chez les exposés pondérée pour le facteur F est donnée par : $FP_1 = \dfrac{Ef_a - 1}{Ef_a} = \dfrac{2,0649 - 1}{2,0649} = 0,5157$, ce qui correspond effectivement à :

$$\frac{\sum_i E(A_{1i})FP_{1i}}{E(A_1)}$$

$$FP_1 = \frac{12}{159} \times 0,75 + \frac{24}{159} \times 0,67 + \frac{30}{159} \times 0,60 + \frac{48}{159} \times 0,50 + \frac{45}{159} \times 0,33$$

Dans le tableau 11.9, nous présentons les intervalles de confiance obtenus par les différentes méthodes précédemment décrites.

TABLEAU 11.9

Méthode	Fraction prévenue	Intervalle de confiance à 95 %	Programme
Transformation Φ_0	0,5157	[0,3554 ; 0,6362]	PR11.3
Transformation Φ_2	0,5157	[0,3553 ; 0,6362]	
Méthode directe	0,5157	[0,3241 ; 0,7073]	

Les résultats des deux premières méthodes sont très concordants, pour ne pas dire identiques. Par contre, on remarque que la troisième méthode conduit à un intervalle centré sur la mesure, d'étendue légèrement plus grande que celle des deux autres intervalles.

◆

11.2.2 FRACTION PRÉVENUE DE POPULATION

Pour le calcul d'un intervalle de confiance de la fraction prévenue de population, nous nous restreignons aux études conduites sur une population ou sur un échantillon de celle-là.

Ce contexte permet de définir des calculs relativement simples : l'estimation des fractions et de leur variance est directe.

La fraction prévenue de population pondérée pour le facteur F, désignée par FP_t (ou plus simplement FP) s'exprime comme la somme pondérée des fractions prévenues spécifiques de population FP_i. Les poids sont proportionnels aux effectifs des cas potentiels. Ainsi,

$$FP = \sum_i \lambda_i FP_i$$

où $\lambda_i = \dfrac{a_{1i}[Ef_i - 1] + m_{1i}}{\sum_i a_{1i}[Ef_i - 1] + m_{1i}} = \dfrac{[E(A_{1i}) - a_{1i}] + m_{1i}}{\sum_i [E(A_{1i}) - a_{1i}] + m_{1i}}$.

L'estimation ponctuelle de FP_t peut être décrite à partir de Ef_a ($Ef_a = 1/RR_a$) : $FP = p_1 \dfrac{Ef_a - 1}{Ef_a}$

où $p_1 = \dfrac{\sum_i n_{1i}}{\sum_i n_i} = \dfrac{n_1}{n}$.

Sous les conditions de marges fixes, pour une strate particulière i, la fraction prévenue de population peut s'écrire comme : $FP_i = \dfrac{a_{0i} - E_0(A_{0i})}{a_{0i}}$

où $E_0(A_{0i}) = \dfrac{m_{1i} n_{0i}}{n_i}$.

Cette condition des marges fixes permet de réduire la dépendance de la mesure FP_i à une simple variable A_{0i}, binomiale ou hypergéométrique suivant le type de données.

Pour simplifier le calcul de la variance de FP, nous suggérons d'utiliser la transformation standard Φ_0 : $\Phi_0(FP) = (1 - FP)^{-1}$. Dans ce cas, $\Phi_0(FP_i) = \dfrac{a_{0i}}{E_0(A_{0i})}$ et $V[\Phi_0(FP_i)] = \dfrac{1}{E_0(A_{0i})^2} V(A_{0i})$. La variance $V[\Phi_0(FP)]$ est alors calculée comme : $V[\Phi_0(FP)] = \sum_i \lambda_i^2 V[\Phi_0(FP_i)]$.

Les limites de confiance de $\Phi(FP)$ sont :

$$\left(\Phi_0\right)_{\text{inf}} = \Phi_0(FP) - z_{\alpha/2}\sqrt{\sum\nolimits_i \lambda_i^2 \frac{V(A_{0i})}{E_0(A_{0i})^2}}$$

$$\left(\Phi_0\right)_{\text{sup}} = \Phi_0(FP) + z_{\alpha/2}\sqrt{\sum\nolimits_i \lambda_i^2 \frac{V(A_{0i})}{E_0(A_{0i})^2}}$$

Par transformation inverse, on déduit celles de FP :

$$FP_{\text{inf}} = \frac{\left(\Phi_0\right)_{\text{inf}} - 1}{\left(\Phi_0\right)_{\text{inf}}}$$

$$FP_{\text{sup}} = \frac{\left(\Phi_0\right)_{\text{sup}} - 1}{\left(\Phi_0\right)_{\text{sup}}}$$

Exemple 11.4

En utilisant les données du tableau 11.7, nous allons calculer l'intervalle de confiance de la fraction attribuable pondérée de population à partir de la méthode présentée.

Dans un premier tableau (tableau 11.10), nous présentons les données de base nécessaires au calcul des limites de confiance.

Tableau 11.10

Âge (années)	m_{1i}	FP_i	λ_i	$E_0(A_{0i})$	$V(A_{0i})$	$\left[\lambda_i / E_0(A_{0i})\right]^2 \times V(A_{0i})$
35-44	7	0,5625	0,1509	1,7500	1,7012	0,01265
45-54	20	0,4444	0,2264	6,6667	4,6605	0,00537
55-64	36	0,3333	0,1132	16,0000	7,4138	0,00037
65-74	60	0,2857	0,2264	25,7143	11,9192	0,00092
75-84	60	0,2000	0,2830	24,0000	11,2000	0,00156
Total	177	0,3446	1,00	–	–	0,02088

En utilisant ces données du tableau 11.10, on peut d'abord déduire la valeur de la fonction prévenue de population pondérée : $FP = 0{,}3446$.

La transformation $\Phi_0(FP)$ donne la valeur de 1,5258 :

$$\Phi_0(FP) = \frac{1}{1 - 0{,}3446} = 1{,}5258$$

La variance de V[Φ(*FP*)] étant de 0,02088, les limites de confiance de Φ(*FP*) sont :

$$\begin{aligned}\left(\Phi_0\right)_{\text{inf}} &= 1,5258 - 1,96\sqrt{0,02088} \\ &= 1,2426 \\ \left(\Phi_0\right)_{\text{sup}} &= 1,5258 + 1,96\sqrt{0,02088} \\ &= 1,8090\end{aligned}$$

Celles de *FP* sont alors de :

$$FP_{\text{inf}} = \frac{\left(\Phi_0\right)_{\text{inf}} - 1}{\left(\Phi_0\right)_{\text{inf}}} = \frac{1,2426 - 1}{1,2426} = 0,1951$$

$$FP_{\text{sup}} = \frac{\left(\Phi_0\right)_{\text{sup}} - 1}{\left(\Phi_0\right)_{\text{sup}}} = \frac{1,8090 - 1}{1,8090} = 0,4472$$

(PR11.4)

◆

PARTIE

5

SUJETS COMPLÉMENTAIRES

CHAPITRE

12

ANALYSE DE PLUSIEURS TAUX

Dans ce chapitre, nous présentons les tests statistiques les plus courants qui permettent de comparer plusieurs taux observés. Une telle situation se présente lorsque, par exemple, on veut décrire le taux d'incidence d'une maladie Y en fonction d'une variable d'exposition X comptant plusieurs catégories ou niveaux. Pour chaque catégorie x_j de la variable X, les données peuvent être décrites suivant le schéma présenté au tableau 12.1. Les rapports a_j/n_j représentent des taux. Les effectifs n_j sont en personnes-temps à risque.

TABLEAU 12.1

	Catégories de X			Total
	x_1	…	x_J	
$Y = 1$	a_1	…	a_J	a
Total*	n_1	…	n_J	n

* Les données sont en personnes-temps.

Nous allons considérer deux situations :

1. La comparaison des taux est faite entre les catégories d'une variable X nominale (ou traitée comme telle). Chaque catégorie de la variable X représente un groupe, un échantillon sans plus. L'hypothèse nulle est alors tout simplement *: les taux sont identiques,* et sa contre-hypothèse tout aussi simple *: les taux ne sont pas tous identiques,* ce qui veut dire qu*'il existe au moins deux taux qui se distinguent l'un de l'autre*.
2. On s'intéresse au comportement linéaire du taux suivant les différents niveaux d'une variable X quantitative. Le taux croît-il (ou décroît-il) linéairement suivant les niveaux de X ? C'est l'étude de la tendance linéaire.

Nous verrons que, pour une variable quantitative, l'étude de la tendance se rapporte assez naturellement à la comparaison de plusieurs taux.

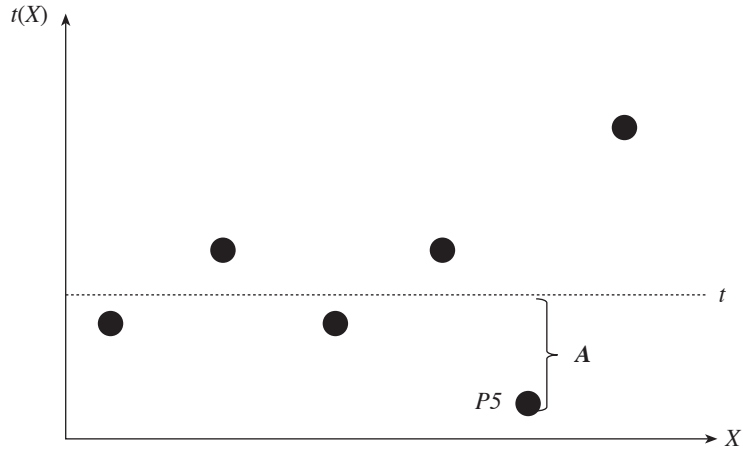
12.1 COMPARAISON DE PLUSIEURS TAUX POUR UNE VARIABLE *X* NOMINALE

Supposons que l'incidence de la maladie Y est décrite pour les différentes catégories de la variable X, suivant les notations du tableau 12.1. Dans ce tableau, les taux observés sont : $t_1 = \frac{a_1}{n_1}, ..., t_J = \frac{a_J}{n_J}$. Le taux marginal est de $t = \frac{a}{n}$.

Pour la comparaison des J taux, nous présentons les trois tests statistiques : un test exact, le test de Pearson et le test du rapport de vraisemblance (RV). Ces deux derniers sont les plus utilisés. Pour ces tests, l'hypothèse nulle suppose que les taux sont égaux ou homogènes. Pratiquement, cela veut dire que les taux observés fluctuent autour d'un taux commun suivant les lois du hasard. Des écarts trop grands entre certains taux et le taux moyen (ou marginal) pourraient s'expliquer en partie par un

phénomène autre que le hasard. Dans la figure 12.1, on illustre différents écarts de taux par rapport à leur moyenne commune *t*. Par exemple, l'écart *A* représente la déviation du point *P5* par rapport au taux *t*.

FIGURE 12.1



12.1.1 TEST EXACT POUR LA COMPARAISON DE PLUSIEURS TAUX

La probabilité exacte liée aux données du tableau 12.1 se calcule dans le cadre de la loi multinomiale (section 1.7.2 du chapitre 1).

Sous l'hypothèse nulle, on a :

$$P(A_1 = a_1,\ A_2 = a_2,\ \ldots,\ A_J = a_J \mid \mathrm{H}_0) = \frac{a!}{n^a} \prod_{j=1}^{J} \frac{n_j^{a_j}}{a_j!}$$

Le test exact pour la comparaison de plusieurs taux est d'emblée défini comme un test bilatéral. Pour la définition de ce test, on utilise la notion de tableau (ou configuration) extrême. Ainsi, sous la condition des marges fixes et sous l'hypothèse nulle, un tableau est dit également ou plus extrême que le tableau observé si sa probabilité est au plus égale à celle du tableau observé. Si on désigne par *H* l'ensemble des configurations (ou tableaux) *u* également ou plus extrêmes que celle observée, y compris cette dernière, sous les conditions des marges fixes, alors

$$\begin{aligned} p &= \sum\nolimits_{u \in H} P(u) \\ &= \sum\nolimits_{u \in H} P(A_1 = u_1,\ A_2 = u_2,\ \ldots,\ A_J = u_J \mid H_0) \end{aligned}$$

où u correspond à une configuration (u_1, u_2, …, u_J) de la ligne $Y = 1$, telle que $\Sigma_j u_j = a$.

Remarquons que la procédure exacte atteint assez rapidement les limites de calcul des logiciels. Le nombre n de configurations (u_1, u_2, …, u_J) est déterminé par $C^a_{a+J-1} = \dfrac{(a+J-1)!}{a!(J-1)!}$, croissant beaucoup plus rapidement avec J qu'avec a (voir le tableau 12.2).

TABLEAU 12.2

		a					
		5	**10**	**15**	**20**	**25**	**30**
J	2	*n* = 6	11	16	21	26	31
	3	21	66	136	231	351	496
	4	56	286	816	1771	3276	5456
	5	126	1001	3876	10626	23751	46376
	6	252	3003	15504	53130	142506	324632
	7	462	8008	54264	230230	736281	1947792

EXEMPLE 12.1

Considérons les données du tableau 12.3.

TABLEAU 12.3

	Catégorie de *X*			Total
	x_1	x_2	x_3	
$Y = 1$	0	2	3	5
Total*	300	400	500	1200

* Les données sont en personnes-temps.

Sous l'hypothèse nulle, la probabilité de ce tableau est donnée par :

$$P(A_1 = 0,\ A_2 = 2,\ A_3 = 3 \mid H_0\ ) = \frac{5!}{1200^5} \times \frac{300^0}{0!} \times \frac{400^2}{2!} \times \frac{500^3}{3!} = 0{,}0804$$

Nous présentons au tableau 12.4 l'ensemble des réalisations (u_1, u_2, u_3) qui conduisent à la somme $a = 5$. Pour chacune de ces réalisations (ou tableaux), nous en présentons la probabilité. Ces réalisations sont présentées en ordre de probabilité décroissant. La valeur-p bilatérale est alors calculée en sommant les probabilités des réalisations au moins aussi extrêmes que celle observée. **(PR12.1)**

TABLEAU 12.4

Réalisation (u_1,u_2,u_3)	Probabilité sous H_0	Valeur-p bilatérale
(1, 2, 2)	0,14468	
(1, 1, 3)	0,12056	–
(2, 1, 2)	0,10851	
(2, 2, 1)	0,08681	
(0, 2, 3) ←	0,08038	
(1, 3, 1)	0,07716	
(0, 3, 2)	0,06430	
(0, 1, 4)	0,05023	
(2, 0, 3)	0,04521	
(3, 1, 1)	0,04340	
(1, 0, 4)	0,03768	
(3, 0, 2)	0,02713	= 0,53946 (convention intégrale)
(0, 4, 1)	0,02572	= 0,49926 (convention mi-p)
(2, 3, 0)	0,02315	
(3, 2, 0)	0,01736	
(1, 4, 0)	0,01543	
(0, 0, 5)	0,01256	
(4, 0, 1)	0,00814	
(4, 1, 0)	0,00651	
(0, 5, 0)	0,00412	
(5, 0, 0)	0,00098	

◆

12.1.2 TEST DU KHI-CARRÉ DE PEARSON

Sous l'hypothèse nulle, on admet que chaque taux t_j fluctue autour de la moyenne commune t, avec une variance $V(t_j)$. Ainsi, par approximation normale, on a :

$$\frac{(t_j - t)}{\sqrt{V(t_j)}} = \frac{\sqrt{n_j}(t_j - t)}{\sqrt{t}} = z_j$$

En élevant au carré le z_j, on obtient le test spécifique au groupe j :

$$\chi^2_{1j} = \frac{n_j(t_j - t)^2}{t}$$

La somme de ces valeurs à travers les J catégories indépendantes de X donne le test du khi-carré désiré : $\chi^2_{J-1} = \sum_{j=1}^{J} \frac{n_j(t_j - t)^2}{t}$. Ce khi-carré a $(J - 1)$ degrés de liberté ; les taux étant liés par leur moyenne t, il y a perte d'un degré de liberté parmi ces J catégories indépendantes.

Ce test (le plus usité) permet de juger de l'homogénéité des mesures t_j entre elles.

Dans une forme mieux connue et plus adaptée aux calculs, nous le présentons comme :

$$\chi^2_{J-1} = \sum \frac{(O - A)^2}{A} \qquad (12.1)$$

où pour chaque cellule, O représente la valeur observée et A la valeur attendue sous l'hypothèse nulle. La sommation se fait sur les J cellules.

12.1.3 TEST DU RAPPORT DE VRAISEMBLANCE

On suppose que les données du tableau 12.1 décrivent les distributions de J variables de Poisson indépendantes, chacune de paramètre $n_j\tau_j$. Alors, sous l'hypothèse nulle que les taux τ_j sont tous égaux ($\tau_1 = \tau_2 = \ldots = \tau_J = \tau$), la fonction de vraisemblance FV_0 de ces données peut être décrite comme :

$$FV_0 = \prod_{j=1}^{J} \left(\frac{e^{-n_j\tau}(n_j\tau)^{a_j}}{a_j!} \right)$$

L'hypothèse nulle se résume à un paramètre unique qui sera estimé par a/n.

Par ailleurs, sous une hypothèse spécifiant des valeurs uniques τ_j pour chacun des J taux, la fonction de vraisemblance FV_1 de ces données se présente comme :

$$FV_1 = \prod_{j=1}^{J} \left(\frac{e^{-n_j\tau_j}(n_j\tau_j)^{a_j}}{a_j!} \right)$$

Pour le maximum de la fonction FV_1, les J paramètres sont fixés aux valeurs observées :

$$\tau_1 = \frac{a_1}{n_1}, \tau_2 = \frac{a_2}{n_2}, ..., \tau_J = \frac{a_J}{n_J}$$

Le test du rapport de vraisemblance se construit ici aussi par une comparaison des logarithmes des fonctions de vraisemblance : $-2\log\left(\frac{FV_0}{FV_1}\right)$. Cette statistique obéit, en bonne approximation, à une loi du khi-carré avec $(J - 1)$ degrés de liberté. Traduit en formule, ce test se présente comme :

$$\chi^2_{J-1} = 2\sum_{j=1}^{J}\left[a_j \log\left(\frac{t_j}{t}\right) + (tn_j - t_j n_j)\right]$$
$$= 2\sum_{j=1}^{J}\left[O \log\left(\frac{O}{A}\right)\right]$$

puisque $\sum tn_j = \sum t_j n_j$. Pour chaque cellule, O correspond à la valeur observée et A à la valeur attendue sous l'hypothèse nulle. La sommation se fait sur toutes les cellules.

EXEMPLE 12.2

Considérons l'ensemble de données du tableau 12.5 qui décrit le nombre de cas de Y suivant les quatre catégories de la variable catégorielle X. Pour chaque niveau x_j de X, le taux de Y est mesuré par le nombre de cas sur les personnes-années à risque.

TABLEAU 12.5

	Catégorie de X				Total
	x_1	x_2	x_3	x_4	
$Y = 1$	11	2	8	1	22
Personnes-années	2200	2400	2600	2800	10000

On veut tester l'hypothèse de l'uniformité de ces taux à travers les catégories de X, et ainsi porter un jugement sur l'association entre X et les taux de la maladie Y.

Test de Pearson

Pour examiner l'association de X avec le taux de la maladie Y, nous comparons les taux des différentes catégories de X à l'aide du test du khi-carré.

Pour chaque cellule, on calcule la valeur $\frac{(O-A)^2}{A}$. Ainsi pour la première cellule, $O = 11$, $A = 2200 \times \frac{22}{10000} = 4,84$, et, donc, $\frac{(O-A)^2}{A} = \frac{(11-4,84)^2}{4,84} = 7,84$. Le total est alors $\chi^1_3 = 15,1087$, pour un $p = 0,00173$.

La variable X est donc significativement associée au taux de la maladie Y. En d'autres termes, il existe au moins une catégorie de X pour laquelle le taux de la maladie Y est significativement différent des taux de certaines autres catégories. (**PR12.2**)

Remarquons que le test exact donne ici une valeur-p semblable : $p = 0,00119$ dans la convention intégrale, et $p = 0,00118$ dans la convention mi-p.

Test du rapport de vraisemblance

Le test appliqué aux données du tableau 12.5 donne $\chi^1_3 = 15,9099$:

$$\chi^2_{4-1} = 2\left[11 \times \log\left(\frac{11}{4,84}\right) + ... + 1 \times \log\left(\frac{1}{6,16}\right)\right]$$
$$= 15,9099$$

La valeur-p correspondante est de 0,001183.

Cette valeur est sensiblement la même que celle donnée par le test exact dans la convention mi-p. (**PR12.3**)

◆

12.2 COMPARAISON DE PLUSIEURS TAUX : ANALYSE D'UNE TENDANCE

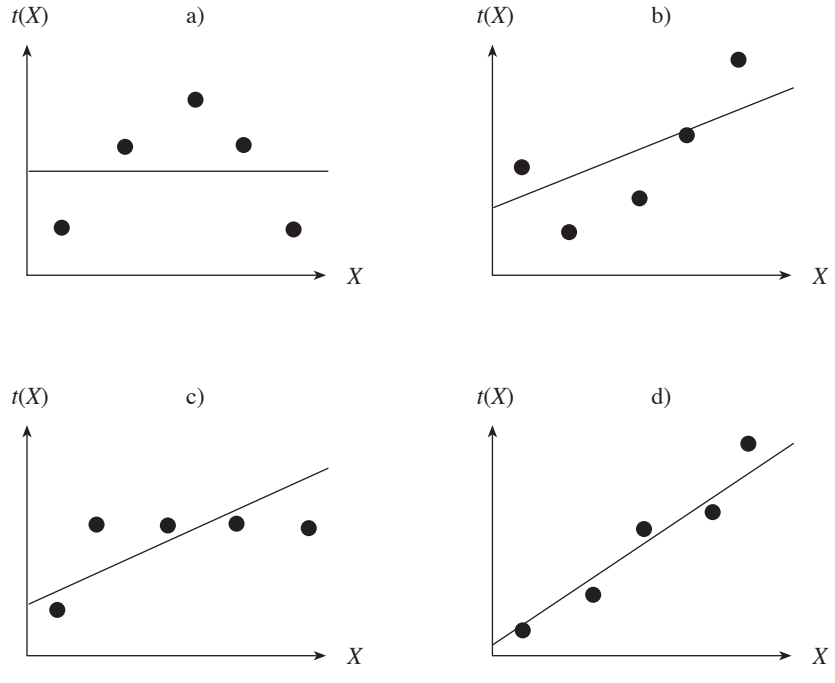
Considérons un ensemble de données, disposées suivant le schéma décrit au tableau 12.1 et décrivant le nombre de cas de Y suivant les différents niveaux de X (variable indépendante quantitative). Pour chaque niveau x_j de X, le taux de Y est désigné par t_j ($t_j = a_j/n_j$). Nous nous intéressons à la croissance (ou décroissance) des t_j en fonction de X.

Le jugement sur la tendance passe généralement par la modélisation linéaire.

On construit la droite de régression pour l'ensemble des points (x_j, t_j) : $t_j = \alpha + \beta x_j + e$.

Une pente positive de la droite ($\beta > 0$) indique que les taux t_j croissent à travers les niveaux de X, et une pente négative ($\beta < 0$), qu'ils décroissent. Si $\beta = 0$, on peut conclure qu'il n'y a pas de tendance linéaire. Dans ce dernier cas, il serait erroné de conclure qu'il n'y a aucune tendance. Absence de tendance linéaire ne veut pas dire absence de toute tendance. Une tendance peut être curvilinéaire sans être linéaire (figure 12.2a). On peut aussi détecter une tendance ($\beta > 0$), sans que le modèle linéaire ne soit le plus adéquat (figure 12.2b). Par contre, il est possible de détecter une tendance linéaire qui n'en soit pas vraiment une (figure 12.2c). Enfin, une tendance peut vraiment être qualifiée de linéaire lorsque $\beta \neq 0$ et que les points $t_j(X)$ se collent assez bien à la droite de régression (figure 12.2d).

FIGURE 12.2



L'analyse des tendances séculaires passe généralement par des tests de tendance. Comprise dans un sens restrictif, l'expression « tendance séculaire » veut dire croissance ou décroissance monotone de la mesure avec le temps. Dans un sens plus large, l'analyse de tendance vise à déterminer le modèle qui colle le mieux à la description de la mesure en fonction du temps.

De par le modèle linéaire utilisé, les tests de tendance présentés ici portent précisément sur la croissance (ou décroissance) monotone des taux.

12.2.1 TEST D'ARMITAGE-COCHRAN

Ce test est défini dans le cadre des variables de Poisson indépendantes. Le jugement sur la tendance des taux passe par le modèle linéaire additif : $t(X) = \alpha + \beta X$. Les coefficients α et β de la droite de régression sont alors estimés respectivement par les expressions (12.2) et (12.3) :

$$\beta = \frac{\sum_{j=1}^{J} n_j \left(t_j - t\right)\left(x_j - \bar{X}\right)}{\sum_{j=1}^{J} n_j \left(x_j - \bar{X}\right)^2} \tag{12.2}$$

$$\alpha = t - \beta \bar{X} \tag{12.3}$$

où $t = \frac{a}{n}$ et $\bar{X} = \frac{\sum_{j=1}^{J} n_j x_j}{n}$.

Le coefficient α représente l'ordonnée à l'origine et correspond à la valeur du taux estimé au niveau $X = 0$. Le coefficient β représente la pente de la droite de régression et correspond à l'effet qu'a, sur le taux, l'accroissement d'une unité en X. En d'autres termes, pour chaque unité d'accroissement en X, le taux croît d'une quantité égale à β. Ainsi, le coefficient β s'interprète simplement comme la différence $[t(x + 1) - t(x)]$ des taux pour les niveaux $x + 1$ et x de X.

Le test de tendance peut facilement être construit sous l'hypothèse nulle $\beta = 0$. Il prend alors la forme suivante :

$$\chi^2_1(\text{tend}) = \frac{\beta^2}{V_0(\beta)}$$

La variance $V_0(\beta)$ est celle de β sous l'hypothèse nulle. Il est facile de montrer que

$$V_0(\beta) = \frac{t}{\sum_{j=1}^{J} n_j \left(x_j - \bar{X}\right)^2}$$

À partir de la régression de $t(X)$ sur X, il peut être intéressant d'examiner si le modèle linéaire est adéquat. On peut montrer que le khi-carré total sur les données du tableau 12.1 peut être partitionné en deux composantes : l'une décrivant la variation liée à la tendance linaire de $t(X)$ sur X, l'autre décrivant les résidus, c'est-à-dire la variation liée aux déviations des observations par rapport au modèle linéaire. Cette partition est décrite à l'expression (12.4) ; les expressions (12.5) et (12.6) en explicitent les composantes.

$$\chi^2_{J-1}(\text{total}) = \chi^2_1(\text{tend}) + \chi^2_{J-2}(\text{res}) \tag{12.4}$$

Composante tendance (tend) :

$$\chi^2_1(\text{tend}) = \frac{\beta^2 \sum_{j=1}^{J} n_j \left(x_j - \bar{X}\right)^2}{t} \tag{12.5}$$

Composante résiduelle (res) :

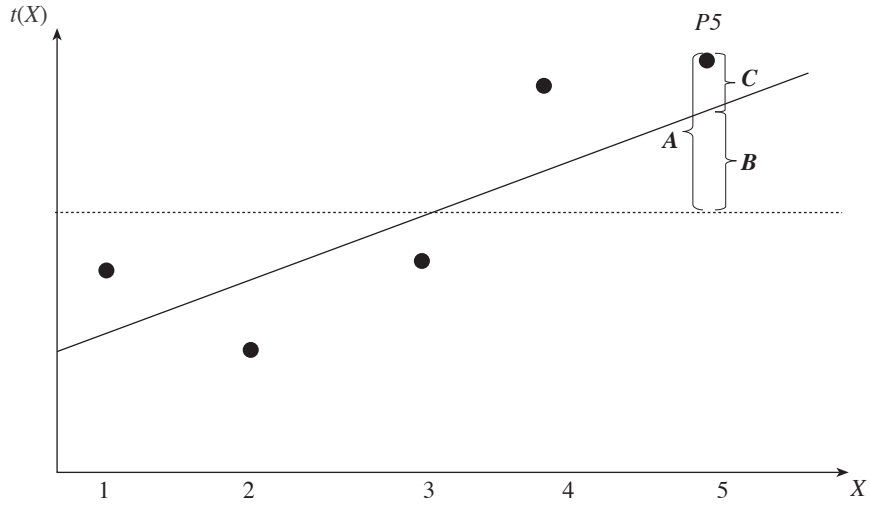
$$\chi^2_{J-2}(\text{res}) = \frac{\sum_{j=1}^{J} n_j \left(t_j - \hat{t}_j\right)^2}{t} \tag{12.6}$$

où $\hat{t}_j$ représente la valeur prédite par le modèle pour le niveau j.

Alors que le χ^2_1 (tend) permet de porter un jugement sur la signification statistique de la pente β, le χ^2_{J-2} (res) permet de porter un jugement sur la tendance qu'ont les points à se coller à la droite. Si cette statistique est nulle ou faible, c'est l'indication que les points se situent sur la droite ou s'y collent de très près. Le modèle est alors adéquat. De façon plus générale, on peut dire que le modèle n'est pas adéquat si χ^2_{J-2} (res) est significatif. Ce test est conduit sous l'hypothèse que les taux s'alignent suivant la droite ou le plan de régression. Si la dispersion des points autour de la droite est importante, au-delà de ce que peut expliquer le hasard, alors on rejette l'hypothèse de la linéarité. Sinon, le modèle linéaire est considéré adéquat.

On peut facilement visualiser la partition à l'aide du schéma de la figure 12.3. Considérons le point *P5*. La déviation totale *A* (du taux *P5* par rapport à la moyenne *t*) est partitionnée suivant les deux composantes : la variation *B* liée à la régression (tendance) et la variation *C* liée à la déviation de *P5* par rapport au modèle linéaire.

FIGURE 12.3



12.2.2 TEST DE MANTEL-HAENSZEL

On peut définir un test analogue à celui de Mantel-Haenszel dans le cadre d'une distribution multinomiale, en conditionnant sur le nombre total de cas de *Y*. Il est intéressant de souligner que ce test de Mantel-Haenszel pour les taux est identique à celui d'Armitage-Cochran défini précédemment.

Reportons-nous au schéma décrit au tableau 12.1.

Le test de Mantel-Haenszel est ici aussi basé sur le score $S = \sum_j A_j x_j$, où A_j représente la variable « nombre de cas » pour le niveau x_j ($X = j$). La valeur observée s de S est $\sum_j a_j x_j$. Si $E(S)$ et $V(S)$ décrivent respectivement la valeur attendue et la variance de S, alors le test est décrit comme :

$$\chi_1^2 = \frac{[s - E(S)]^2}{V(S)}$$

Si on comprend que, dans les conditions de marges fixes, les variables A_j obéissent à une loi J-multinomiale telle que

$$P(A_1 = a_1, A_2 = a_2, \ldots, A_J = a_J \mid H_0) = \frac{a!}{n^a} \prod_{j=1}^{J} \frac{n_j^{a_j}}{a_j!}$$

(voir la section 1.7.1 du chapitre 1), alors, dans le cadre de cette distribution de probabilités, on peut établir que :

$$E(A_j) = \frac{a\,n_j}{n}, V(A_j) = \frac{a n_j (n - n_j)}{n^2} \text{ et } Cov(A_u, A_v) = -\frac{a n_u n_v}{n^2} \text{ où } u \neq v.$$

Ainsi, la variance $V(S)$ se présente comme :

$$V(S) = \frac{a}{n^2}\left[\sum_j X_j^2 n_j (n - n_j) + 2\sum_{u,v} X_u X_v n_u n_v\right]$$

$$= \frac{a}{n^2}\left[n\sum_j X_j^2 n_j - \left(\sum_j X_j n_j\right)^2\right]$$

et le test :

$$\chi_1^2 = \frac{\left[n\sum_j a_j x_j - a\sum_j n_j x_j\right]^2}{a\left[n\sum_j n_j x_j^2 - \left(\sum_j n_j x_j\right)^2\right]}$$

Il est facile de montrer algébriquement que cette expression est équivalente à celle du test de tendance défini précédemment :

$$\chi_1^2(\text{tend}) = \frac{\beta^2 \sum_{j=1}^{J} n_j \left(x_j - \bar{X}\right)^2}{t}$$

12.2.3 TEST DU RAPPORT DE VRAISEMBLANCE

Pour le test du rapport de vraisemblance, nous considérons le modèle linéaire suivant : $\Phi[t(X)] = \alpha + \beta X$, où $\Phi = I$ pour le modèle additif et $\Phi = \log$ pour le modèle multiplicatif. Dans le modèle linéaire additif, le coefficient β s'interprète comme la différence entre les taux de deux niveaux successifs de X : $\beta = [t(x + 1) - t(x)]$, et dans le modèle linéaire multiplicatif comme le rapport de ces taux : $\beta = \log\left(\frac{t(x+1)}{t(x)}\right)$.

Il suffit de comparer les valeurs des deux fonctions de vraisemblance calculées sur les données : la valeur de la fonction $FV(\alpha)$ correspondant à l'hypothèse nulle (soit le modèle linéaire de base, $\Phi[t(X)] = \alpha$) et la valeur de la fonction $FV(\alpha,\beta)$ correspondant au modèle linéaire $\Phi[t(X)] = \alpha + \beta X$. La statistique $-2\log\left(\frac{FV(\alpha)}{FV(\alpha,\beta)}\right)$ obéit, en bonne approximation, à une loi du khi-carré avec 1 degré de liberté. Les deux fonctions $FV(\alpha)$ et $FV(\alpha,\beta)$ (respectivement FV_0 et FV_1) ont déjà été décrites à la section 12.1.3.

Ainsi, la fonction $FV(\alpha,\beta)$ se présente comme $FV_1 = \prod_{j=1}^{J}\left(\frac{e^{-n_j\tau_j}(n_j\tau_j)^{a_j}}{a_j!}\right)$ où $\tau_j = \alpha + \beta x_j$ pour le modèle additif, et $\tau_j = e^{\alpha+\beta x_j}$ pour le modèle multiplicatif.

L'estimation des coefficients α et β et celle de leurs erreurs-types peuvent être obtenues par la procédure GENMOD dans le cadre de la loi de Poisson.

EXEMPLE 12.3

Considérons les données du tableau 12.6 où X est considéré comme une variable quantitative ayant les niveaux $X = 1, 2, 3$ et 4.

TABLEAU 12.6

	Niveau de X				
	1	2	3	4	Total
$Y = 1$	50	40	30	20	140
Personnes-années	2200	2400	2600	2800	10000

TEST D'ARMITAGE-COCHRAN

Examinons la tendance des taux de Y à travers les différents niveaux de X. La droite de régression obtenue pour la modélisation des taux est de la forme :

$$t(X) = 0,027419 - 0,0051613X$$

La valeur du khi-carré total est de 23,7172 avec 3 degrés de liberté. Celle de χ_1^2 (tend) est de 23,5945. Par soustraction, on peut déduire celle de χ_2^2 (res) est de 23,7172 – 23,5945 = 0,1277.

La tendance est très significative ($p < 0,0001$). Ainsi, à partir de ce modèle, on peut estimer à 0,005 la décroissance dans le taux de Y pour chaque unité d'accroissement dans X. On remarque que la valeur de χ_1^2 (tend) occupe presque toute la place dans la partition du khi-carré total. En conséquence, le χ_2^2 (res) est presque nul, indiquant ainsi que les taux se comportent de façon quasi linéaire avec X. (**PR12.4**)

TEST DE MANTEL-HAENSZEL

Comme on peut s'y attendre, le test de Mantel-Haenszel appliqué aux données du tableau 12.6 donne une valeur identique à celle calculée par le test d'Armitage-Cochran. (**PR12.5**)

TESTS DU RAPPORT DE VRAISEMBLANCE

Appliquée aux données du tableau 12.6, la méthode du rapport de vraisemblance conduit aux résultats suivants. Indiquons d'abord que le khi-carré total est de 23,88, avec 3 degrés de liberté. Puis :

1. Pour le modèle additif, la différence des taux est estimée à $\beta = -0{,}005$; le χ_1^2 (tend) donne une valeur de 23,76 et le χ_2^2 (res), une valeur de 0,1210. Ces résultats se comparent assez bien à ceux obtenus dans l'approche Armitage-Cochran.

2. Pour le modèle multiplicatif, le rapport des taux est estimé à $e^{\beta} = e^{-0{,}3737} = 0{,}69$ et le χ_1^2 (tend) donne une valeur de 23,65 et le χ_2^2 (res), une valeur de 0,2250. (**PR12.6**)

◆

12.3 EXTENSION DES TESTS DE TENDANCE AU CONTRÔLE D'UNE VARIABLE

Nous présentons deux extensions possibles du test de tendance pour le contrôle d'une variable : le test basé sur la somme pondérée des pentes β_i et l'extension de Mantel adaptée aux taux. (Pour l'extension de Mantel, voir ci-après, au chapitre 13, la section 13.3.2.) Le premier test découle assez naturellement du test d'Armitage-Cochran et le second, du test de Mantel. Encore ici, nous verrons qu'ils conduisent à des résultats assez similaires.

12.3.1 TEST SUR LA SOMME PONDÉRÉE DES PENTES

Si, pour la strate i, β_i désigne la pente de la droite de régression et $V(\beta_i)$ sa variance, alors la pente moyenne sur les strates peut être définie comme :

$$\bar{\beta} = \frac{\sum_i w_i \beta_i}{\sum_i w_i}$$

où $w_i = \dfrac{1}{V(\beta_i)}$.

Sous l'hypothèse nulle, le test statistique prend la forme

$$\chi_1^2 = \frac{(\bar{\beta})^2}{V(\bar{\beta})} = \frac{(\sum_i w_i \beta_i)^2}{\sum_i w_i}$$

12.3.2 EXTENSION DU TEST DE MANTEL

Pour les taux, l'extension de Mantel peut s'obtenir en faisant appel à la loi multinomiale. Sous cette loi, et compte tenu que les strates sont indépendantes les unes des autres, le test prend la forme :

$$\chi_I^2 = \frac{\left\{\sum_i \left[\sum_j a_{ij}x_j - \frac{a_{i.}}{n_{i.}}\sum_j n_{ij}x_j\right]\right\}^2}{\sum_i \left\{\frac{a_{i.}}{n_{i.}^2}\left[n_{i.}\sum_j n_{ij}x_j^2 - \left(\sum_j n_{ij}x_j\right)^2\right]\right\}},$$

où $a_{i.}, b_{i.}, n_{i.} = \sum_j a_{ij}, \sum_j b_{ij}, \sum_j n_{ij}$.

12.3.3 TEST DU RAPPORT DE VRAISEMBLANCE

Supposons que l'on s'intéresse à la tendance des taux suivant la variable *X*, contrôlée pour un vecteur *Z* de variables. Alors, l'analyse de cette tendance repose sur la construction du modèle multivarié. Il suffit de considérer le modèle $\Phi[\tau(X,Z)] = \alpha + \beta_1 X + \beta_2 Z$, additif ou multiplicatif. L'extension des tests du rapport de vraisemblance au contrôle de plusieurs variables se fait naturellement à l'intérieur de la procédure GENMOD de SAS.

EXEMPLE 12.4

Considérons les données du tableau 12.6 stratifiées pour une variable *F*. Supposons que nous obtenions le tableau 12.7.

TABLEAU 12.7

F = 1	Niveau de *X*				
	1	2	3	4	Total
Y = 1	20	15	10	5	50
Personnes-années	600	700	800	900	3000
F = 2	Niveau de *X*				
	1	2	3	4	Total
Y = 1	30	25	20	15	90
Personnes-années	1600	1700	1800	1900	7000

Nous voulons décrire la tendance des taux en ajustant pour la variable *F*.

Test sur la somme pondérée des pentes

Le tableau 12.8 décrit quelques résultats importants en analyse stratifiée et globale ajustée. Nous rappelons qu'en analyse brute, la pente de la droite était de ?0,0051613 (exemple 12.3).

Tableau 12.8

Strate	β	$V(\beta)$	w	χ^2_1 (tend)	IC à 95 %
$F = 1$	–0,090909	0,00045455	2200,00	18,1818	[–0,13270 ; 0,04912]
$F = 2$	–0,036066	0,00014754	6777,78	8,8160	[–0,05987 ; 0,01226]
Pente ajustée	–0,049505	0,00011139	8977,78	22,0022	[–0,07019 ; 0,02882]

Nous remarquons que la pente est fortement modifiée par le facteur F : elle est de –0,091 pour la strate 1 et de –0,036 pour la strate 2. Nous pouvons ici appliquer un test sur l'homogénéité des pentes en utilisant les propriétés de la partition du khi-carré total en analyse stratifiée.

Le $\chi^2(\text{total}) = \sum_{i=1}^{I} w_i \beta_i^2$ est la somme des χ^2_1 (tend) spécifiques. Sa valeur est de (18,1818 + 8,8160) = 26,9978 pour 2 degrés de liberté. Ce khi-carré total peut être partitionné en un khi-carré qui porte sur la pente ajustée [χ^2_1 (tend) = 22,0022] et en un second qui porte sur l'homogénéité des pentes spécifiques [χ^2_1 (homog) = 26,9978 – 22,0022 = 4,9956]. Le test sur l'homogénéité est ici significatif au niveau 5 %. Ainsi, les pentes sont significativement différentes d'une strate à l'autre. Par contre, les pentes brutes et ajustées sont assez similaires (–0,051613 versus –0,049505). Il en va de même pour les tests correspondants (23,5945 versus 22,0022). (**Pr12.7**)

Extension de Mantel

Appliqué aux données du tableau 12.7, le test a comme valeur 24,2221, alors que celle du test d'Armitage-Cochran est de 22,0022.

Si, en analyse simple, les tests d'Armitage-Cochran et de Mantel-Haenszel coïncident, en analyse stratifiée ils diffèrent légèrement. (**Pr12.8**)

Test du rapport de vraisemblance

En appliquant ce test aux données du tableau 12.7, on obtient les résultats suivants.

1. Pour le modèle additif, la différence des taux, ajustée pour la variable F, est estimée à $\beta = –0,0049$; le χ^2_1 (tend) a comme valeur 22,33.
2. Pour le modèle multiplicatif, le rapport des taux, ajusté pour la variable F, est estimé à 0,6846 et le χ^2_1 (tend) est de 24,27. (**Pr12.9**)

◆

12.4 TEST DE TENDANCE EXACT POUR LES TAUX EN ANALYSE UNIVARIÉE

Considérons le tableau 12.9, qui décrit les données d'une étude de cohorte conduite dans une population ouverte. Pour le niveau x_j de X, a_j nouveaux cas de maladie ont été recensés pour les n_j personnes-temps observées. Les variables A_j correspondantes sont des variables de Poisson indépendantes, chacune de paramètre $\mu_j = \tau_j n_j$. Les taux d'incidence observés sont les estimations des taux paramétriques τ_j.

TABLEAU 12.9

	Niveau de X				Total
	x_1	x_2	…	x_J	
$Y = 1$	a_1	a_2	…	a_J	a
$Y = 0$					
Personnes-temps	n_1	n_2	…	n_J	n

On s'interroge sur l'existence d'une tendance linéaire pour les taux d'incidence t_j suivant les différents niveaux x_j de X. On suppose que $x_1 < x_2 < \ldots < x_J$.

Sous la condition des marges fixes, à savoir que $\Sigma_j A_j = a$, les variables $(A_1, A_2, \ldots, A_J)$ obéissent à une loi multinomiale de paramètres $\{\pi_j\}$ et a, où $\Sigma_j \pi_j = 1$. Les paramètres π_j peuvent être décrits à l'aide des taux τ_j comme :

$$\pi_j = \frac{\tau_j n_j}{\sum_{k=1}^{J} \tau_k n_k}.$$

Dans ces conditions, la probabilité d'observer les données du tableau est décrite comme :

$$P(A_1 = a_1,\ A_2 = a_2,\ \ldots,\ A_J = a_J \mid \{\pi_j\}) = a! \times \prod_{j=1}^{J} \frac{\pi_j^{a_j}}{a_j!}$$

Sous l'hypothèse d'une croissance linéaire des taux τ_j avec les niveaux de x_j de X, on a : $\tau_j = \alpha + \beta x_j$. Il peut être plus avantageux de considérer la croissance linéaire de $\log(\tau_j)$: $\log(\tau_j) = \alpha + \beta x_j$, sans que s'en trouve modifié l'essentiel de la méthode. Suivant cette transformation, la probabilité multinomiale correspond alors à :

$$P(A_1 = a_1,\ A_2 = a_2,\ \ldots\ , A_J = a_J \mid \alpha, \beta)$$
$$= \frac{a!e^{\alpha a}}{\left(\sum_{j=1}^{J} \tau_j n_j\right)^a} \times e^{\sum_j a_j x_j \beta} \times \prod_{j=1}^{J} \frac{n_j^{a_j}}{a_j!} \quad .$$

Soit $\{u_j\}$ une réalisation quelconque de la variable multinomiale $\{A_j\}$, telle que $\Sigma_j u_j = a$.

Alors, on peut écrire :

$$P(A_1 = u_1,\ A_2 = u_2,\ \ldots\ , A_J = u_J \mid \alpha, \beta)$$
$$= \frac{a!e^{\alpha a}}{\left(\sum_{j=1}^{J} \tau_j n_j\right)^a} \times e^{\sum_j u_j x_j \beta} \times \prod_{j=1}^{J} \frac{n_j^{u_j}}{u_j!} \quad .$$

Considérons la variable-score $S = \sum_j A_j x_j$ qui désigne le score moyen sur X pour les variables A_j. Les valeurs de S varient entre ax_1 et ax_J. À chaque valeur s de S correspond un ensemble $R(s)$ de réalisations de $\{A_j\}$ qui donnent ce score s. On dira que $R(s)$ regroupe les réalisations équipotentes des variables A_j pour le score s : pour une réalisation $u = \{u_j\}$ de $R(s)$, on a $s_u = \sum_j u_j x_j = s$. En particulier, le score $t = \sum_j a_j x_j$ observé peut aussi être obtenu par d'autres réalisations équipotentes de l'ensemble $R(t)$.

Suivant cette convention, la probabilité du score s de S se décrit comme :

$$P(S = s \mid \alpha, \beta) = \sum_{u \in R(s)} P(A_1 = u_1,\ A_2 = u_2, \ldots\ , A_J = u_J \mid \alpha, \beta)$$
$$= \frac{a!e^{a\alpha}}{\left(\sum_j \tau_j n_j\right)^a} \times e^{s\beta} \times \sum_{u \in R(s)} \left(\prod_{j=1}^{J} \frac{n_j^{u_j}}{u_j!}\right)$$

Pour le calcul de cette probabilité sous une certaine hypothèse de la tendance β, le coefficient α est un paramètre de nuisance : en effet, on ne saurait déterminer la probabilité sans émettre aussi une hypothèse sur α.

Remarquons que la probabilité non conditionnelle $P(S = s \mid \alpha, \beta)$ est factorisée en trois termes. Le premier terme, fonction de α et des valeurs marginales, est constant quelle que soit la valeur s de S ; il ne comporte donc aucune information qui permette de distinguer les différentes probabilités des valeurs du score S. Le second est fonction du score S et de β. Le troisième, indépendant de α et β, correspond à un poids w_s proportionnel à la probabilité de la valeur s du score S.

Pour éliminer le paramètre de nuisance α, il suffit d'éliminer le premier terme de la factorisation, ce que l'on réalise en établissant le rapport de la probabilité $P(S = t \mid \alpha, \beta)$ à la somme des probabilités sur s, $\Sigma_s P(S = s \mid \alpha, \beta)$. On obtient alors ce que l'on peut considérer comme la probabilité conditionnelle exacte sur S. Cette probabilité est celle d'observer $S = t$ relativement à toutes les valeurs possibles s de S, pondérées par le facteur w_s :

$$\frac{P(S = t \mid \alpha, \beta)}{\sum_s P(S = s \mid \alpha, \beta)} = \frac{w_t e^{t\beta}}{\sum_s w_s e^{s\beta}} = P(S = t \mid \beta)$$

où $w_s = \sum_{u \in R(s)} \left(\prod_{j=1}^{J} \frac{n_j^{u_j}}{u_j!} \right)$ u parcourant l'ensemble $R(s)$ des réalisations équipotentes de $\{A_j\}$ pour s (et $R(t)$ pour le score t).

La probabilité conditionnelle exacte de S, $P(S = t \mid \beta)$, s'avère identique à la probabilité non conditionnelle $P(S = t \mid \alpha, \beta)$, puisque le dénominateur du premier membre de l'équation est identiquement égal à 1. Par ailleurs, cette probabilité conditionnelle a l'avantage d'être indépendante du paramètre de nuisance α.

On peut estimer le paramètre β par le maximum de vraisemblance en utilisant la méthode itérative de Newton-Raphson.

Pour le cas particulier de $\beta = 0$ (l'hypothèse nulle), cette probabilité se réduit à la distribution multinomiale. Pour le montrer, il suffit d'établir la relation suivante :

$$\sum_{u \in R(.)} \left(\prod_{j=1}^{J} \frac{n_j^{u_j}}{u_j!} \right) = \frac{n^a}{a!}$$

où u parcourt l'ensemble $R(.)$ de toutes les réalisations possibles de $\{A_j\}$, tel que $\Sigma_j u_j = a$.

Pour établir le test exact (unilatéral à droite) sous H_0, il suffit alors de calculer la probabilité : $P(S \geq t \mid \beta = 0)$.

Sous l'hypothèse nulle, le score attendu $E(S)$ de S est donné par :

$$E(S) = x_1 \times \frac{an_1}{n} + x_2 \times \frac{an_2}{n} + x_3 \times \frac{an_3}{n} = \frac{a}{n} \sum_j x_j n_j$$

et la variance $V(S)$ par :

$$V(S) = \sum_j x_j^2 V(A_j) + 2 \sum_{i \neq j} x_i x_j \operatorname{cov}(A_i, A_j)$$

$$= \frac{a}{n^2} \left[n \sum_j n_j x_j^2 - \left(\sum_j n_j x_j \right)^2 \right]$$

où les variances et covariances sont calculées dans le cadre de la distribution multinomiale.

EXEMPLE 12.5

Considérons les données du tableau 12.10. Elles reprennent essentiellement celles du tableau 12.3.

TABLEAU 12.10

	Niveau de X			Total
	1	2	3	
$Y = 1$	0	2	3	5
Personnes-temps	300	400	500	1200

Sous la condition que $A = A_1 + A_2 + A_3 = 5$, on relève 21 réalisations possibles du vecteur (A_1, A_2, A_3) (tableau 12.11). Parmi celles-là, on observe le résultat (0,2,3). Pour les valeurs de X suivant les trois niveaux considérés : $X_1 = 1$, $X_2 = 2$ et $X_3 = 3$, le score t calculé est de $1 \times 0 + 2 \times 2 + 3 \times 3 = 13$.

L'ensemble $R(13)$ des réalisations équipotentes comprend les deux réalisations suivantes : le résultat observé (0,2,3) et (1,0,4), chacune donnant un score de 13. Ainsi, la probabilité d'observer un score de 13 revient à celle d'observer l'une ou l'autre des deux réalisations de $R(13)$: $P(S = 13) = P[(1,0,4)] + P[(0,2,3)]$. Ces dernières probabilités se calculent dans le cadre de la loi multinomiale.

◆

Au tableau 12.11, nous décrivons les différents scores de S, leurs ensembles équipotents, ainsi que les probabilités des réalisations et des scores calculées sous l'hypothèse nulle.

TABLEAU 12.11

Score s*	Classe $R(s)$ des réalisations (a_1,a_2,a_3) équipotentes pour le score s		Probabilité sous H_0	
			Des réalisations (a_1,a_2,a_3)	Des scores s
5	$R(5)$:	(5,0,0)	0,00098	0,00098
6	$R(6)$:	(4,1,0)	0,00651	0,00651
7	$R(7)$:	(3,2,0)	0,01736	
		(4,0,1)	0,00814	0,02550
8	$R(8)$:	(2,3,0)	0,02315	
		(3,1,1)	0,04340	0,06655
9	$R(9)$:	(1,4,0)	0,01543	
		(2,2,1)	0,08681	
		(3,0,2)	0,02713	0,12937
10	$R(10)$:	(0,5,0)	0,00412	
		(1,3,1)	0,07716	
		(2,1,2)	0,10851	0,18979
11	$R(11)$:	(0,4,1)	0,02572	
		(1,2,2)	0,14468	
		(2,0,3)	0,04521	0,21561
12	$R(12)$:	(0,3,2)	0,06430	
		(1,1,3)	0,12056	0,18486
13	$R(13)$:	(0,2,3)	0,08038	
		(1,0,4)	0,03768	0,11806
14	$R(14)$:	(0,1,4)	0,05023	0,05023
15	$R(15)$:	(0,0,5)	0,01256	0,01256

* $s = X_1a_1 + X_2a_2 + X_3a_3$.

Sous l'hypothèse nulle, la valeur-p unilatérale à droite, $P(S \geq 13)$, est calculée sur les données du tableau 12.11.

Dans la convention intégrale :

$$p = 0{,}11806 + 0{,}05023 + 0{,}01256 = 0{,}18085$$

Dans la convention mi-p,

$$p = 1/2\,30{,}11806 + 0{,}05023 + 0{,}01256 = 0{,}12182$$

(PR12.10)

La valeur attendue $E(S)$ et la variance $V(S)$ du score S sous l'hypothèse nulle sont :

$$E(S) = \frac{5}{12}\left[3\times 1 + 4\times 2 + 5\times 3\right] = 10,83$$

$$V(S) = \frac{5}{12^2}\left[\begin{array}{l} 12\times\left(3\times 1^2 + 4\times 2^2 + 5\times 3^2\right) \\ -\left(3\times 1 + 4\times 2 + 5\times 3\right)^2 \end{array}\right] = 3,19$$

Ainsi, à titre de comparaison, le test de Mantel (unilatéral à droite) donne :

$$\chi_2 = \frac{(13-10,83)^2}{3,19} = 1,4761$$

pour une valeur-*p* unilatérale de 0,1122.

Dans le tableau 12.12, nous reproduisons les données du tableau 12.11 sous l'hypothèse d'une pente $\beta = 1$.

TABLEAU 12.12

Score *s*			Probabilité sous l'hypothèse $\beta = 1$	
			Des réalisations (a_1,a_2,a_3)	Des scores *s*
5	*R*(5) :	(5,0,0)	0,00000	0,00000
6	*R*(6) :	(4,1,0)	0,00001	0,00001
7	*R*(7) :	(4,0,1)	0,00004	
		(3,2,0)	0,00009	0,00013
8	*R*(8) :	(2,3,0)	0,00034	
		(3,1,1)	0,00064	0,00098
9	*R*(9) :	(1,4,0)	0,00062	
		(3,0,2)	0,00109	
		(2,2,1)	0,00348	0,00519
10	*R*(10) :	(0,5,0)	0,00045	
		(1,3,1)	0,00841	
		(2,1,2)	0,01182	0,02068
11	*R*(11) :	(0,4,1)	0,00762	
		(2,0,3)	0,01339	
		(1,2,2)	0,04285	0,06386
12	*R*(12) :	(0,3,2)	0,05177	
		(1,1,3)	0,09707	0,14884
13	*R*(13) :	(1,0,4)	0,08246	
		(0,2,3)	0,17591	0,25837
14	*R*(14) :	(0,1,4)	0,29885	0,29885
15	*R*(15) :	(0,0,5)	0,20309	0,20309

La valeur-*p* calculée dans la convention intégrale sous l'hypothèse $\beta = 1$ est de 0,76 (valeur mi-*p* de 0,63), ce qui indique une excellente compatibilité des données avec cette hypothèse.

Dans la convention intégrale :

$p = 0{,}25837 + 0{,}29885 + 0{,}20309 = 0{,}76031$

Dans la convention mi-*p*,

$p = 1/2\,30{,}25837 + 0{,}29885 + 0{,}20309 = 0{,}631125$

(PR12.11)

CHAPITRE 13

ANALYSE DE PLUSIEURS PROPORTIONS

Dans ce chapitre, nous présentons les tests statistiques les plus courants qui permettent de comparer plusieurs proportions. Une telle situation se présente lorsque, par exemple, on veut décrire le risque d'une maladie Y en fonction d'une variable d'exposition X comptant plusieurs catégories ou niveaux. Pour chaque catégorie x_j de la variable X, les données peuvent être décrites suivant le schéma présenté au tableau 13.1. Les rapports a_j/n_j représentent des proportions. Les effectifs n_j sont en personnes.

TABLEAU 13.1

	Catégories de X			Total
	x_1	...	x_J	
$Y = 1$	a_1	...	a_J	a
$Y = 0$	b_1	...	b_J	b
Total*	n_1	...	n_J	n

Nous allons considérer deux situations :

1. On compare les proportions de catégories d'une variable X nominale (ou traitée comme telle). Chaque catégorie de la variable X représente un groupe, un échantillon sans plus. L'hypothèse nulle est alors tout simplement *: les proportions sont identiques,* et sa contre-hypothèse, tout aussi simple : *les proportions ne sont pas toutes identiques,* ce qui veut dire qu'*il existe au moins deux proportions qui se distinguent l'une de l'autre*.
2. On s'intéresse au comportement linéaire de la proportion suivant les différents niveaux d'une variable X quantitative. La proportion croît-elle (ou décroît-elle) linéairement suivant les niveaux de X ? C'est l'étude de la tendance linéaire.

Nous verrons que, pour une variable quantitative, l'étude de la tendance se rapporte assez naturellement à la comparaison de plusieurs proportions.

13.1 COMPARAISON DE PLUSIEURS PROPORTIONS POUR UNE VALEUR NOMILALE DE *X*

Supposons que le risque de la maladie Y est décrit pour les différentes catégories de la variable X, suivant le schéma du tableau 13.1. Dans ce tableau, les proportions observées sont

$$p_1 = \frac{a_1}{n_1}, ..., p_J = \frac{a_J}{n_J}$$

La proportion marginale est de $p = \frac{a}{n}$.

Pour la comparaison des J proportions, nous présentons les trois tests statistiques : un test exact, le test de Pearson et le test du rapport de vraisemblance (RV). Ces deux derniers sont les plus utilisés. Pour ces tests,

l'hypothèse nulle suppose que les proportions sont égales ou homogènes. Pratiquement, cela veut dire que les proportions observées fluctuent autour d'une proportion commune suivant les lois du hasard. Des écarts trop grands entre certaines proportions et la proportion moyenne (ou marginale) pourraient s'expliquer en partie par un phénomène autre que le hasard. (Voir à cet égard la figure 12.1 du chapitre 12.)

13.1.1 TEST EXACT POUR LA COMPARAISON DE PLUSIEURS PROPORTIONS

Le test exact pour la comparaison de plusieurs proportions est une généralisation du test de Fisher. Établissons d'abord que la probabilité exacte liée au tableau 13.1 se calcule dans le cadre de la loi hypergéométrique multiple (voir la section 1.7.4, chapitre 1). Sous l'hypothèse nulle, on a :

$$P(A_1 = a_1,\ A_2 = a_2,\ \ldots,\ A_J = a_J \mid H_0) = \frac{\prod_{j=1}^{J} C_{n_j}^{a_j}}{C_n^a}$$

Si la notion de valeur extrême pour une variable A_1 hypergéométrique simple dans le cadre d'un tableau 2×2 est naturelle, elle ne l'est plus pour des variables hypergéométriques multiples $\{A_j\}$. En d'autres termes, les tests unilatéraux (à droite ou à gauche) ne sont pas définis. Par ailleurs, il est possible d'étendre à la comparaison de plusieurs proportions le test de Fisher dans l'approche bilatérale. À cette fin, nous devons établir la notion de tableau extrême. Sous les conditions des marges fixes et sous l'hypothèse nulle, un tableau est dit également ou plus extrême que le tableau observé si sa probabilité est au plus égale à celle du tableau observé. En désignant par H l'ensemble des tableaux u également ou plus extrêmes que le tableau observé, y compris ce dernier, sous les conditions des marges fixes, alors

$$\begin{aligned} p &= \sum\nolimits_{u \in H} P(u) \\ &= \sum\nolimits_{u \in H} P(A_1 = u_1,\ A_2 = u_2,\ \ldots,\ A_J = u_J \mid H_0) \end{aligned}$$

Un tableau u correspond à une configuration (u_1, u_2, …, u_J) de la ligne $Y = 1$, telle que $\Sigma_j\, u_j = a$.

EXEMPLE 13.1

Considérons les données du tableau 13.2.

TABLEAU 13.2

	Catégorie de X			Total
	x_1	x_2	x_3	
$Y = 1$	0	2	3	5
$Y = 0$	3	2	2	7
Personnes	3	4	5	12

La probabilité de ce tableau est donnée par :

$$P(A_1 = 0,\ A_2 = 2,\ A_3 = 3 \mid H_0\ ) = \frac{C_3^0 C_4^2 C_5^3}{C_{12}^5} = 0{,}07576$$

Nous présentons au tableau 13.3 l'ensemble des réalisations (u_1,u_2,u_3) qui conduisent à la somme $a = 5$. Pour chacune de ces réalisations (ou tableaux) nous en présentons la probabilité. Ces réalisations sont présentées en ordre décroissant de leurs probabilités.

TABLEAU 13.3

Réalisation (u_1,u_2,u_3)	Probabilité sous H_0	Valeur-p bilatérale
(1, 2, 2)	0,22727	
(2, 1, 2)	0,15152	–
(1, 1, 3)	0,15152	
(2, 2, 1)	0,11364	
(0, 2, 3)	0,07576	
(1, 3, 1)	0,07576	
(0, 3, 2)	0,05051	
(2, 0, 3)	0,03788	
(3, 1, 1)	0,02525	
(0, 1, 4)	0,02525	= 0,**35605** (convention intégrale)
(1, 0, 4)	0,01894	= 0,**28029** (convention mi-p)
(2, 3, 0)	0,01515	
(3, 0, 2)	0,01263	
(3, 2, 0)	0,00758	
(0, 4, 1)	0,00631	
(1, 4, 0)	0,00379	
(0, 0, 5)	0,00126	

(PR13.1)

◆

13.1.2 TEST DU KHI-CARRÉ DE PEARSON

Sous l'hypothèse nulle, chaque proportion p_j fluctue autour de la moyenne commune p, avec une variance $V(p_j)$. Ainsi, par approximation normale, on a :

$$\frac{(p_j - p)}{\sqrt{V(p_j)}} = \frac{(p_j - p)}{\sqrt{\dfrac{p(1-p)}{n_j}}} = Z_j \mapsto \mathrm{N}(0,1)$$

Le carré Z_j^2 de cette variable obéit à une loi du khi-carré :

$$\chi_1^2 = \frac{n_j(p_j - p)^2}{p(1-p)}$$

La somme $\sum_j Z_j^2$ obéit à une loi du khi-carré avec J degrés de liberté, si les Z_j sont des variables indépendantes les unes des autres (voir la section 1.6 du chapitre 1). L'indépendance de ces J variables n'est cependant pas tout à fait respectée puisque la définition du paramètre commun p est liée à chacune des mesures p_j : $p = \sum_{j=1}^{J} \frac{n_j}{n} p_j$.

Cette situation crée une dépendance linéaire : pour $J-1$ proportions arbitrairement fixées, la J^e est automatiquement déterminée. La conséquence de cette dépendance entre les p_j, et donc entre les Z_j, est la perte d'un degré de liberté pour le khi-carré.

Ainsi : $\chi_{J-1}^2 = \sum_{j=1}^{J} \frac{n_j(p_j - p)^2}{p(1-p)}$.

C'est un test (le plus usité) qui permet de juger de l'homogénéité des mesures entre elles. Dans une forme mieux connue et plus adaptée aux calculs, nous le présentons comme

$$\chi_{J-1}^2 = \sum \frac{(O - A)^2}{A} \tag{13.1}$$

où, pour chaque cellule, O représente la valeur observée et A la valeur attendue sous l'hypothèse nulle. La sommation est faite sur les $2 \times J$ cellules.

13.1.3 TEST DU RAPPORT DE VRAISEMBLANCE

Pour les données du tableau 13.1 où on compare entre elles J variables binomiales, la fonction de vraisemblance a la forme générale suivante :

$$FV = \pi_1^{a_1}(1-\pi_1)^{b_1}\pi_2^{a_2}(1-\pi_2)^{b_2}\ldots\pi_J^{a_J}(1-\pi_J)^{b_J}$$

Sous l'hypothèse H_0 de l'égalité des proportions ($\pi_1 = \pi_2 = \ldots = \pi_J = \pi$), la fonction de vraisemblance FV_0 de ces données peut se décrire comme suit :

$$FV_0 = \pi^{a_1}(1-\pi)^{b_1}\pi^{a_2}(1-\pi)^{b_2}\ldots\pi^{a_J}(1-\pi)^{b_J}$$

où π correspond à a/n, valeur suggérée par les données.

Sous l'hypothèse la plus vraisemblable, les proportions π_j sont celles correspondant aux différentes proportions mesurées sur les données :

$$\pi_1 \approx p_1 = \frac{a_1}{n_1},\quad \pi_2 \approx p_2 = \frac{a_2}{n_2},\quad \ldots,\quad \pi_J \approx p_J = \frac{a_J}{n_J}$$

On comprend bien ici que l'hypothèse (ou le modèle) qui explique le mieux les données est celle dont la valeur de la FV est la plus grande.

La fonction de vraisemblance FV_1 est alors décrite comme :

$$FV_1 = p_1^{a_1}\left(1-p_1\right)^{b_1}\ p_2^{a_2}\left(1-p_2\right)^{b_2}\ldots\ p_J^{a_J}\left(1-p_J\right)^{b_J}$$

Pour construire le test, il suffit alors de comparer les valeurs des deux fonctions de vraisemblance calculées sur les données : FV_0 et FV_1. Le test du rapport de vraisemblance se présente donc comme : $\chi^2 = -2\log\left(\frac{FV_0}{FV_1}\right)$, avec $J-1$ degrés de liberté.

Traduit en formule, ce test se décrit comme :

$$\chi^2 = 2\sum_{j=1}^{J}\left[a_j\log\left(\frac{p_j}{p}\right) + b_j\log\left(\frac{(1-p_j)}{(1-p)}\right)\right]$$
$$= 2\sum O\log\left(\frac{O}{A}\right)$$

EXEMPLE 13.2

Considérons la variable d'exposition X dont certaines catégories peuvent être associées au risque de la maladie Y (tableau 13.4).

TABLEAU 13.4

	Niveau de X				
	x_1	x_2	x_3	x_4	Total
$Y = 1$	50	40	30	20	140
$Y = 0$	170	200	230	260	860
Total	220	240	260	280	1000

Pour examiner l'association entre le facteur X et la maladie Y, nous comparons les risques de Y entre les différentes catégories de X, à l'aide du test du khi-carré.

TEST DU KHI-CARRÉ DE PEARSON

Pour chaque cellule du tableau, on calcule la valeur $\frac{(O-A)^2}{A}$. Par exemple, pour la première cellule, $O = 50$ et $A = 220 \times \frac{140}{1000} = 30,8$, donc $\frac{(O-A)^2}{A} = \frac{(50-30,8)^2}{30,8} = 11,97$. Il en va de même pour les autres cellules. Puis on fait la somme des huit cellules.

$$\chi^2_3 = \frac{(50-30,80)^2}{30,80} + \ldots + \frac{(260-240,80)^2}{240,80}$$
$$= 27,578$$

Pour $\chi^2_3 = 27,578$, on a une valeur-$p < 0,001$. La variable X est donc significativement associée au risque de la maladie Y. En d'autres termes, il existe au moins une catégorie de X pour laquelle le risque de maladie Y est significativement différent des risques pour certaines autres catégories.

TEST DU RAPPORT DE VRAISEMBLANCE

Le test appliqué aux données du tableau 13.4 donne $\chi^2_3 = 27,771$, pour une valeur-$p < 0,001$.

$$\chi^2 = 2\sum O \log\left(\frac{O}{A}\right) = 2\left[\begin{array}{l} 50\log\left(\frac{50}{30,8}\right) + 40\log\left(\frac{40}{33,6}\right) + \\ \ldots + 260\log\left(\frac{260}{240,8}\right) \end{array}\right]$$
$$= 27,771$$

Le résultat de ce test est similaire à celui du test précédent. L'hypothèse d'homogénéité des proportions est rejetée. (**PR13.2**)

◆

13.2 COMPARAISON DE PLUSIEURS PROPORTIONS : ANALYSE D'UNE TENDANCE

Considérons un ensemble de données, disposées suivant le schéma décrit au tableau 13.5 et décrivant la distribution de Y (variable dépendante) suivant les différents niveaux de X maintenant considérée comme variable indépendante quantitative. Pour chaque niveau x_j de X, la proportion de $Y = 1$ est désignée par p_j ($= a_j/n_j$). Nous nous intéressons à la croissance (ou décroissance) des p_j en fonction de X.

TABLEAU 13.5

	Niveau de X				Total
	x_1	x_2	...	x_J	
$Y = 1$	a_1	a_2	...	a_J	a
$Y = 0$	b_1	b_2	...	b_J	b
Total	n_1	n_2	...	n_J	n

L'analyse de tendance linéaire pour les proportions est analogue à celle pour les taux (voir la section 12.2 du chapitre 12).

Trois tests sont disponibles pour juger directement de la tendance linéaire des proportions sur X : le test d'Armitage-Cochran, le test de Mantel-Haenszel et le test du rapport de vraisemblance. Le premier est construit dans le cadre des variables binomiales indépendantes. Le deuxième est un test conditionnel conduit dans le cadre d'une distribution multiple hypergéométrique. Enfin, le test du rapport de vraisemblance est basé sur la comparaison de la fonction de vraisemblance du modèle linéaire ($Y = \alpha + \beta X$) à celle du modèle de base ($Y = \alpha$). Ce test se trouve directement dans le prolongement du test du rapport de vraisemblance décrit pour la comparaison de plusieurs proportions.

13.2.1 TEST D'ARMITAGE-COCHRAN

À l'aide de la méthode des moindres carrés, nous déterminons la droite de régression qui permet de décrire la tendance des proportions suivant les niveaux de X (tableau 13.5).

Les coefficients α et β de la droite de régression $p(X) = \alpha + \beta X$ sont alors estimés respectivement par les expressions (13.2) et (13.3) :

$$\beta = \frac{\sum_{j=1}^{J} n_j \left(p_j - p\right)\left(x_j - \bar{X}\right)}{\sum_{j=1}^{J} n_j \left(x_j - \bar{X}\right)^2} \tag{13.2}$$

$$\alpha = p - \beta\bar{X} \tag{13.3}$$

où $p = \frac{a}{n}$ et $\bar{X} = \frac{\sum_{j=1}^{J} n_j x_j}{n}$. Le coefficient α représente l'ordonnée à l'origine et correspond à la proportion sur le niveau $X = 0$. Il peut arriver que cette proportion ne corresponde à aucune réalité. Par exemple, si le modèle $p(X) = \alpha + \beta X$ décrit le risque de faible poids à la naissance $p(X)$ chez le bébé en fonction de l'âge X de la mère, le coefficient α décrit le risque de faible poids chez le bébé né d'une mère d'âge 0 année. On peut raisonnablement penser qu'aucune mère n'a accouché, n'accouche ou n'accouchera à un tel âge. Pour chaque niveau d'âge $X = x$, le risque n'est réellement décrit que par l'expression $\alpha + \beta x$.

Le coefficient β représente la pente de la droite de régression et correspond à l'effet qu'a sur le risque l'accroissement d'une unité en X. En d'autres termes, pour chaque unité d'accroissement en X, le risque croît d'une quantité égale à β. La tendance linéaire est donc marquée par ce coefficient β.

Le test de tendance peut facilement être construit sous l'hypothèse nulle $\beta = 0$. Il se présente alors comme :

$$\chi_1^2(\text{tend}) = \frac{\beta^2}{V_0(\beta)}$$

La variance $V_0(\beta)$ est celle de β sous l'hypothèse nulle. Il est facile de montrer que

$$V_0(\beta) = \frac{pq}{\sum_{j=1}^{J} n_j \left(x_j - \bar{X}\right)^2}$$

À partir de la régression de $p(X)$ sur X, il peut être intéressant d'examiner si le modèle linéaire est adéquat.

Le khi-carré total sur les données du tableau 13.5 peut être partitionné en deux composantes : l'une décrivant la variation liée à la tendance de $p(X)$ sur X, l'autre décrivant les résidus, c'est-à-dire la variation liée aux déviations des observations par rapport au modèle linéaire. Cette partition est décrite à l'expression (13.4) ; les expressions (13.5) et (13.6) en explicitent les composantes.

$$\chi^2_{J-1}(\text{total}) = \chi^2_1(\text{tend}) + \chi^2_{J-2}(\text{res}) \tag{13.4}$$

Composante tendance (tend) :

$$\chi^2_1(\text{tend}) = \frac{\beta^2 \sum_{j=1}^{J} n_j \left(x_j - \bar{X}\right)^2}{pq} \tag{13.5}$$

Composante résiduelle (res) :

$$\chi^2_{J-2}(\text{res}) = \frac{\sum_{j=1}^{J} n_j \left(p_j - \hat{p}_j\right)^2}{pq} \tag{13.6}$$

Dans l'expression (13.6), $\hat{p}_j$ désigne la valeur prédite par le modèle pour la valeur x_j.

Comme pour les taux, le χ^2_1 (tend) permet de porter un jugement sur la signification statistique de la pente β alors que le χ^2_{J-2} (res) permet de porter un jugement sur la tendance qu'ont les points à se coller à la droite. Si cette statistique est nulle ou faible, c'est l'indication que les points se situent sur la droite ou sont très près d'elle. Le modèle est alors adéquat. De façon plus générale, on peut dire que le modèle n'est pas adéquat si χ^2_{J-2} (res) est significatif.

13.2.2 TEST DE MANTEL-HAENSZEL

Examinons le tableau 13.5.

Le test de Mantel-Haneszel est basé sur la statistique ou le score $S = \sum_j A_j x_j$, où A_j représente la variable « nombre de cas » pour le niveau x_j. La valeur observée s de S correspond à $\sum_j a_j x_j$.

Si on désigne par $E(S)$ et par $V(S)$ respectivement la valeur attendue et la variance de S, alors le test est simplement défini comme :

$$\chi^2_1 = \frac{\left[s - E(S)\right]^2}{V(S)}$$

Dans les conditions de marges fixes, la valeur attendue et la variance de S sont respectivement données par :

$$\begin{aligned} E(S) &= E\left(\sum_j A_j x_j\right) \\ &= \sum_j \left[E\left(A_j x_j\right)\right] \\ &= \sum_j x_j E\left(A_j\right) \\ V(S) &= V(\textstyle\sum_j A_j x_j) \\ &= \textstyle\sum_j V(A_j x_j) + 2\sum_{u,v} \operatorname{cov}(A_u x_u, A_v x_v) \\ &= \textstyle\sum_j x_j^2 V(A_j) + 2\sum_{u,v} x_u x_v \operatorname{cov}(A_u, A_v) \text{ où } 1 \le u,v \le J \text{ et } u \ne v \end{aligned}$$

On comprend ici que les valeurs x_j de X sont fixes.

Par ailleurs, en se plaçant dans les conditions de marges fixes et sous l'hypothèse nulle, on peut considérer le a_j comme la réalisation d'une variable A_j hypergéométrique multiple de dimension J, telle que

$$P(A_1 = a_1, A_2 = a_2, \ldots, A_J = a_J \mid H_0) = \frac{\displaystyle\prod_{j=1}^{J} C_{n_j}^{a_j}}{C_n^a}$$

Alors, dans le cadre de cette distribution de probabilités, on peut établir que :

$$E(A_j) = \frac{a\, n_j}{n}, V(A_j) = \frac{a\, b\, n_j (n - n_j)}{n^2 (n-1)} \text{ et } \operatorname{cov}(A_u, A_v) = -\frac{a\, b\, a_u b_v}{n^2 (n-1)}$$

où $u \ne v$.

Ainsi, l'expression de la variance $V(S)$ devient :

$$\begin{aligned} V(S) &= \frac{a\, b}{n^2 (n-1)} \left[\textstyle\sum_j x_j^2 n_j (n - n_j) + 2\sum_{u,v} x_u x_v n_u n_v\right] \\ &= \frac{a\, b}{n^2 (n-1)} \left[n \textstyle\sum_j x_j^2 n_j - \left(\sum_j x_j n_j\right)^2\right] \end{aligned}$$

Le test prend alors la forme :

$$\chi_1^2 = \frac{\left[\displaystyle\sum_j a_j x_j - \frac{a}{n} \sum_j n_j x_j\right]^2}{\dfrac{ab}{n^2 (n-1)} \left[n \displaystyle\sum_j n_j x_j^2 - \left(\sum_j n_j x_j\right)^2\right]}$$

13.2.3 TEST DU RAPPORT DE VRAISEMBLANCE

Ce test est analogue à celui décrit pour les taux (voir section 12.2.3 du chapitre 12).

Pour la construction du test du rapport de vraisemblance, nous considérons le modèle linéaire suivant : $\Phi[\pi(X)] = \alpha + \beta$ où Φ est une transformation (fonction de lien) de $\pi(X)$. Nous nous intéressons particulièrement à trois transformations Φ : $\Phi = \mathrm{I}$, log et logit, correspondant respectivement au modèle linéaire additif $\pi(X) = \alpha + \beta X$, au modèle linéaire multiplicatif $\log[\pi(X)] = \alpha + \beta X$ pour le rapport des proportions et au modèle linéaire multiplicatif $\log it[\pi(X)] = \log \dfrac{\pi(X)}{1-\pi(X)} = \alpha + \beta X$ pour le rapport de cotes.

Dans chacun de ces cas, il suffit de comparer les valeurs des deux fonctions de vraisemblance calculées sur les données : la fonction $FV(\alpha)$ correspond au modèle linéaire $\Phi[\pi(X)] = \alpha$ et la fonction $FV(\alpha,\beta)$ correspond au modèle linéaire $\Phi[\pi(X)] = \alpha + \beta X$.

La statistique $-2\log\left(\dfrac{FV(\alpha)}{FV(\alpha,\beta)}\right)$ obéit, en bonne approximation, à une loi du khi-carré avec 1 degré de liberté. La fonction $FV(\alpha)$ a déjà été décrite à la section 13.1.3 comme FV_0. La fonction $FV(\alpha,\beta)$ se présente comme $FV_1 = \prod_{j=1}^{J} \pi_j^{a_j}(1-\pi_j)^{b_j}$ où $\pi_j = \alpha + \beta x_j$ dans le modèle additif, $\pi_j = e^{\alpha+\beta x_j}$ dans le modèle multiplicatif pour le rapport de proportions, ou $\pi_j = \dfrac{e^{\alpha+\beta x_j}}{1+e^{\alpha+\beta x_j}}$ dans le modèle multiplicatif pour le rapport de cotes.

Dans GENMOD de SAS, la valeur des tests de tendance est également disponible suivant la transformation choisie : IDENTITY, LOG ou LOGIT.

EXEMPLE 13.3

Considérons la variable X d'exposition dont certaines catégories peuvent être associées au risque de la maladie Y. Nous reproduisons les données du tableau 13.4 dans le tableau 13.6, où sont maintenant fixés des niveaux de X : 1, 2, 3 et 4.

TABLEAU 13.6

	Niveau de X				
	1	**2**	**3**	**4**	**Total**
$Y = 1$	50	40	30	20	140
$Y = 0$	170	200	230	260	860
Total	220	240	260	280	1000

Pour l'analyse de la tendance du risque de Y à travers les différents niveaux de X, nous présentons les résultats des trois tests décrits précédemment.

TEST D'ARMITAGE-COCHRAN

Pour la modélisation, la variable X est traitée comme une variable quantitative où les valeurs sont respectivement 1, 2, 3 et 4. La droite de régression ainsi obtenue est de la forme :

$$\pi(X) = 0{,}27419 - 0{,}051613X$$

La valeur de χ^2_3 (total) est de 27,578 (voir exemple 13.2). Celle du χ^2_1 (tend) est de 27,435. Par soustraction, on peut déduire celle du χ^2_2 (res) : 27,5782 – 27,4354 = 0,14275.

Le χ^2_1 (tend) conduit à une valeur-p très faible ($p < 0{,}001$), ce qui marque une tendance fortement significative au plan statistique. Par ailleurs, le χ^2_2 (res) est relativement faible, indiquant ainsi que les proportions se comportent de façon quasi linéaire avec X.

À partir de ce modèle, on peut estimer à 0,0516 la décroissance dans le risque de Y pour chaque unité d'accroissement dans X. (**PR13.3** et **PR13.4**)

TEST DE MANTEL-HAENSZEL

Appliqué aux données du tableau 13.6, le résultat du test de Mantel-Haenszel est χ^2_1 (tend) = 27,408, résultat très similaire au χ^2_1 (tend) calculé par la méthode d'Armitage-Cochran. (**PR13.5**)

TEST DU RAPPORT DE VRAISEMBLANCE

Appliqués aux données du tableau 13.6, les résultats des tests du rapport de vraisemblance sont

1. pour le modèle additif :
   χ^2_1 (tend) = 27,63 et χ^2_2 (res) = 0,1437
2. pour le modèle multiplicatif du RP :
   χ^2_1 (tend) = 27,51 et χ^2_2 (res) = 0,2590
3. pour le modèle multiplicatif du RC :
   χ^2_1 (tend) = 27,63 et χ^2_2 (res) = 0,1380

On remarque que, dans les trois cas, χ^2_1 (tend) + χ^2_2 (res) = 27,77, compte tenu des valeurs arrondies. Cette valeur est précisément celle du χ^2_3 (total) du rapport de vraisemblance (voir l'exemple 13.2). (**PR13.6**)

◆

13.3 EXTENSION DES TESTS DE TENDANCE SUR LES PROPORTIONS AU CONTRÔLE D'UNE VARIABLE

Nous présentons ici deux extensions possibles du test de tendance pour le contrôle d'une variable : le test basé sur la somme pondérée des pentes β_i et l'extension de Mantel. Le premier découle assez naturellement du test d'Armitage-Cochran et le second, du test de Mantel-Haenszel. Encore ici, nous verrons qu'ils conduisent à des résultats assez similaires.

13.3.1 EXTENSION DU TEST D'ARMITAGE-COCHRAN

On peut facilement définir une extension du test de tendance en considérant la somme des pentes β_i, pondérée chacune par l'inverse de sa variance. La moyenne des pentes ainsi obtenue constitue, comme on le sait, une statistique qui se prête facilement aux outils d'inférence statistique.

Si, pour la strate i, β_i désigne la pente de la droite de régression et $V(\beta_i)$ sa variance, alors la pente moyenne sur les strates est définie comme :

$$\bar{\beta} = \frac{\sum_i w_i \beta_i}{\sum_i w_i} \text{ où } w_i = 1/V(\beta_i).$$

Sous l'hypothèse de l'indépendance des strates, on a $V\left(\bar{\beta}\right) = \dfrac{1}{\sum_i w_i}$.

Sous l'hypothèse nulle, le test statistique se présente comme :

$$\chi_1^2 = \frac{\left(\bar{\beta}\right)^2}{V\left(\bar{\beta}\right)} = \frac{\left(\sum_i w_i \beta_i\right)^2}{\sum_i w_i}$$

13.3.2 EXTENSION DU TEST DE MANTEL-HAENSZEL

On doit à Mantel un test de tendance qui peut tenir compte d'une tierce variable.

Considérons une variable F à contrôler pour une étude de tendance du risque de la maladie Y suivant les niveaux d'exposition à un facteur X. On suppose que la variable F a I catégories et le facteur X a J niveaux. Pour une strate i de la variable F, le tableau de données peut se présenter comme :

TABLEAU 13.7

Strate i	Niveau de X				Total
	x_1	x_2	...	x_J	
$Y = 1$	a_{i1}	a_{i2}	...	a_{iJ}	a_i
$Y = 0$	b_{i1}	b_{i2}	...	b_{iJ}	b_i
Total	n_{i1}	n_{i2}	...	n_{iJ}	n_i

Pour la strate i, le score S_i correspond à celui défini précédemment : $S_i = \sum_i A_{ij} x_j$. La valeur attendue et la variance de ce score ont déjà été décrites précédemment, à la section 13.2.2.

La valeur observée s du score S global correspond à $\sum_i \sum_j a_{ij} x_j$. Le test est alors défini comme : $\chi_1^2 = \dfrac{[s - E(S)]^2}{V(S)}$.

Dans sa forme opérationnelle, le test peut être décrit comme :

$$\chi_1^2 = \frac{\left\{ \sum_i \left[\sum_j a_{ij} x_j - \frac{a_{i.}}{n_{i.}} \sum_j n_{ij} x_j \right] \right\}^2}{\sum_i \left\{ \frac{a_{i.} b_{i.}}{n_{i.}^2 (n_{i.} - 1)} \left[n_{i.} \sum_j n_{ij} x_j^2 - \left(\sum_j n_{ij} x_j \right)^2 \right] \right\}}$$

où $a_{i.}$, $b_{i.}$, $n_{i.} = \sum_j a_{ij}$, $\sum_j b_{ij}$, $\sum_j n_{ij}$.

13.3.3 EXTENSION DU TEST DU RAPPORT DE VRAISEMBLANCE

Supposons que l'on veuille déterminer l'effet du facteur X en contrôlant pour celui de F. Le test du rapport de vraisemblance est alors basé simplement sur la comparaison des valeurs des deux fonctions de vraisemblance calculées sur les données : la valeur de la fonction $FV(\alpha,\beta_2)$ du modèle $Y = \alpha + \beta_2 F$ ne comprenant pas X et la valeur de la fonction $FV(\alpha,\beta_1,\beta_2)$ du modèle linéaire $Y = \alpha + \beta_1 X + \beta_2 F$. La statistique $-2\log\left(\dfrac{FV(\alpha,\beta_2)}{FV(\alpha,\beta_1,\beta_2)}\right)$ obéit, en bonne approximation, à une loi du khi-carré avec 1 degré de liberté. L'estimation du maximum de vraisemblance des coefficients α, β_1 et β_2 des modèles linéaires est obtenue dans la procédure GENMOD.

Exemple 13.4

Considérons les données du tableau 13.6. L'information sur le facteur *F* permet de stratifier les données pour ce facteur. Les données stratifiées sont présentées au tableau 13.8.

Tableau 13.8

F = 1	Niveau de *X*				Total
	1	2	3	4	
Y = 1	20	15	10	5	50
Y = 0					
Total	60	70	80	90	300
F = 2	Niveau de *X*				Total
	1	2	3	4	
Y = 1	30	25	20	15	90
Y = 0					
Total	160	170	180	190	700

Nous voulons présenter la tendance des risques de *Y* sur les niveaux de *X*, en ajustant pour la variable *F*.

Extension du test d'Armitage-Cochran

Dans le tableau 13.9, nous présentons certains résultats importants de l'analyse stratifiée et globale ajustée des données du tableau 13.8. Nous rappelons qu'en analyse brute, la pente de la droite était de –0,051613 (voir exemple 13.3).

Tableau 13.9

Strate de *F*	β	*V*(β)	*w*	χ^2(tend)	IC à 95 %
1	–0,090909	0,00037879	2640,00	21,8182	[–0,12906 ; 0,052763]
2	–0,036066	0,00012857	7777,78	10,1168	[–0,05829 ; 0,013841]
Pente ajustée	–0,049964	0,00009599	10417,78	26,0066	[–0,06917 ; 0,030761]

Nous remarquons que la pente est fortement modifiée par le facteur *F* : elle est de –0,091 pour la strate 1 et de –0,036 pour la strate 2. Nous pouvons ici appliquer un test sur l'homogénéité des pentes en utilisant les propriétés de la partition du khi-carré total en analyse stratifiée. Le $\chi^2(\text{total})$ $\left(=\sum_{i=1}^{I} w_i\beta_i^2\right)$ a comme composantes les χ_i^2 (tend)des strates. Sa valeur est donc de (21,8182 + 10,1168) = 31,9350 pour 2 degrés de liberté. Ce khi-carré total peut être partitionné en un khi-carré qui porte sur la pente ajustée [χ_i^2 (tend) = 26,0066]

avec 1 degré de liberté et en un khi-carré d'homogénéité entre les pentes [χ^2_1 (homog) = 31,9350 – 26,0066 = 5,9284]. La pente ajustée est significativement différente de 0 et les pentes sur les strates sont significativement différentes l'une de l'autre.

Certes, en présence d'une forte modification, il peut être moins pertinent de présenter une mesure globale ajustée. En tout état de cause, nous remarquons ici que la pente ajustée est sensiblement la même que celle décrite en analyse brute à l'exemple 13.3, test d'Armitage-Cochran : $\beta = -0{,}051613$ versus $-0{,}049964$, ce qui laisse supposer que le facteur F est faiblement confondant. Les tests correspondants, en analyse brute et ajustée, sont respectivement de 27,4354 et 26,0066 ; ils sont aussi très similaires. (**PR13.7**)

Extension du test de Mantel-Haenszel

Appliqué aux données du tableau 13-9, le χ^2(tend) a comme valeur 28,1726. (**PR13.8** ou **PR13.9**)

Test du rapport de vraisemblance

Appliqué aux données du tableau 13.8, le χ^2_1 (tend) a comme valeur 26,2310 pour le modèle additif, 28,85 pour le modèle multiplicatif avec le rapport de proportions et 28,44 pour le modèle multiplicatif avec le rapport de cotes. (**PR13.10**)

On remarque que le test du rapport de vraisemblance dans le modèle additif [χ^2_1 (tend) = 26,2310] et le test d'Armitage-Cochran [χ^2_1 (tend) = 26,0066] sont assez similaires ; de même, on retrouve une bonne similitude entre le test du rapport de vraisemblance dans le modèle multiplicatif avec rapport de cotes et l'extension du test de Mantel.

◆

13.4 TEST DE TENDANCE EXACT POUR LES PROPORTIONS EN ANALYSE UNIVARIÉE

Considérons le tableau 13.10, qui décrit les données d'une étude de cohorte conduite dans une population fermée. Pour le niveau x_j de la variable quantitative X, a_j cas de maladie ont été recensés pour n_j personnes observées. Nous rappelons que les variables A_j correspondantes sont des variables binomiales indépendantes, chacune de paramètre π_j et n_j. Les proportions

$p_j = \dfrac{a_j}{n_j}$ observées sont les estimations des proportions π_j.

TABLEAU 13.10

	Niveau de X				Total
	x_1	x_2	...	x_J	
$Y = 1$	a_1	a_2	...	a_J	a
$Y = 0$	b_1	b_2	...	b_J	b
Total	n_1	n_2	...	n_J	n

On veut alors déterminer s'il existe une tendance linéaire pour les proportions p_j suivant les différents niveaux x_j de X. On suppose que $x_1 < x_2 < \ldots < x_J$.

Sous la condition des marges fixes, à savoir que $\sum_j A_j = a$ (et que $\sum_j B_j = b$), les variables $(A_1, A_2, \ldots, A_J)$ obéissent à une loi hypergéométrique multiple de paramètres $\{\{\kappa_j\},\{n_j\},a\}$ où κ_j désigne la cote de risque pour le niveau x_j. Les paramètres κ_j peuvent être décrits à l'aide des proportions π_j comme : $\kappa_j = \dfrac{\pi_j}{1-\pi_j}$. Dans ces conditions, la probabilité d'observer les données du tableau 13.10 est décrite par :

$$P(A_1 = a_1,\ A_2 = a_2,\ \ldots,\ A_J = a_J \mid \{\kappa_j\},\ a) = \frac{\prod_{j=1}^{J} C_{n_j}^{a_j} \kappa_j^{a_j}}{\sum_{u \in R} \prod_{j=1}^{J} C_{n_j}^{u_j} \kappa_j^{u_j}}$$

où R désigne l'ensemble des réalisations (ou tableaux) $u = \{u_j\}$ de $\{A_j\}$ telles que $\sum_j u_j = a$.

Sous l'hypothèse d'une croissante linéaire des proportions p_j avec les niveaux de x_j de X, on a : $\pi_j = \alpha + \beta x_j$. Il peut être plus avantageux de considérer la croissance linéaire de $\log(\kappa_j)$: $\log(\kappa_j) = \alpha + \beta x_j$, sans que s'en trouve modifié l'essentiel de la méthode. Suivant cette transformation logit des proportions, la probabilité décrite dans le modèle hypergéométrique correspond alors à :

$$P(A_1 = a_1,\ A_2 = a_2,\ \ldots,\ A_J = a_J \mid \beta) = \frac{e^{a\alpha} \times e^{\sum_j a_j x_j \beta} \times \prod_{j=1}^{J} \dfrac{n_j!}{a_j!\, b_j!}}{e^{a\alpha} \sum_{u \in R} e^{\sum_j u_j x_j \beta} \times \prod_{j=1}^{J} \dfrac{n_j!}{u_j!\, v_j!}}$$

où $b_j = n_j - a_j$ et $v_j = n_j - u_j$.

En simplifiant les termes semblables, on obtient :

$$P(A_1 = a_1,\ A_2 = a_2,\ \ldots,\ A_J = a_J \mid \beta) = \frac{e^{\sum_j a_j x_j \beta} \times \prod_{j=1}^{J} \frac{1}{a_j!\, b_j!}}{\sum_{u \in R} e^{\sum_j u_j x_j \beta} \times \prod_{j=1}^{J} \frac{1}{u_j!\, v_j!}}$$

Considérons la variable-score $S = \sum_j A_j x_j$ qui désigne le score moyen sur X pour les variables A_j. Les valeurs de S varient entre ax_1 et ax_J. Chaque valeur s de S correspond à un ensemble $R(s)$ de réalisations équipotentes des variables A_j pour cette valeur du score. Ainsi, pour une réalisation $u = \{u_j\}$ de R(s), on a $s_u = \sum_j u_j x_j = s$. En particulier, le score $t = \sum_j a_j x_j$ observé peut aussi être obtenu par d'autres réalisations équipotentes de l'ensemble $R(t)$.

Suivant cette convention, on peut décrire la probabilité du score t de S comme :

$$P(S = t \mid \beta) = \frac{w_t e^{t\beta}}{\sum_s w_s e^{s\beta}} \text{ où } w_s = \sum_{u \in R(s)} \left(\prod_{j=1}^{J} \frac{1}{u_j!\ v_j!} \right)$$

u parcourant l'ensemble $R(s)$ des réalisations équipotentes de $\{A_j\}$ pour s [et $R(t)$ pour le score t]. Cette probabilité conditionnelle exacte de S est fonction de β seulement. Le paramètre β peut être estimé par le maximum de vraisemblance en utilisant la méthode itérative de Newton-Raphson.

Pour le cas particulier de $\beta = 0$ (l'hypothèse nulle), cette probabilité se réduit à la distribution hypergéométrique centrée. Pour le montrer, il suffit d'établir la relation suivante :

$$\sum_{u \in R} \left(\prod_{j=1}^{J} \frac{n_j!}{u_j!\, v_j!} \right) = \frac{n!}{a!b!}$$

où u parcourt l'ensemble R de toutes les réalisations possibles de $\{A_j\}$, telle que $\Sigma_j u_j = a$ et $v_j = n_j - u_j$.

Pour la conduite du test sous H_0 (unilatéral à droite), il suffit alors de calculer la probabilité $P(S \geq t \mid \beta = 0)$.

Sous l'hypothèse nulle, la moyenne $E(S)$ et la variance $V(S)$ de S sont respectivement données par :

$$E(S) = x_1 \times \frac{an_1}{n} + x_2 \times \frac{an_2}{n} + \ldots + x_J \times \frac{an_J}{n} = \frac{a}{n}\sum_j x_j n_j$$

$$V(S) = \sum_j x_j^2 V(A_j) + 2\sum_{i \neq j} x_i x_j \operatorname{cov}(A_i, A_j)$$

$$= \frac{ab}{n^2(n-1)}\left[n\sum_j n_j x_j^2 - \left(\sum_j n_j x_j\right)^2 \right]$$

EXEMPLE 13.5

Considérons les données d'une étude de cohorte, comme au tableau 13.11.

TABLEAU 13.11

	Niveau de X			Total
	1	2	3	
$Y = 1$	0	2	3	5
$Y = 0$	3	2	2	7
Personnes	3	4	5	12

Sous la condition que $A = A_1 + A_2 + A_3 = 5$, on relève 17 réalisations possibles du vecteur (A_1, A_2, A_3) (tableau 13.12). Parmi celles-là, on identifie la réalisation (0, 2, 3). Pour les valeurs de X suivant les trois niveaux considérés : $x_1 = 1$, $x_2 = 2$ et $x_3 = 3$, le score t calculé est de $1\times0 + 2\times2 + 3\times3 = 13$.

Dans le tableau 13.12, on présente le calcul de la valeur-p unilatérale à droite sous l'hypothèse nulle $\beta = 0$:

$$p = P(S \geq 13) = 0{,}09470 + 0{,}02525 + 0{,}00126 = 0{,}12121$$

Dans la convention mi-p, on a $p = 1/2\times0{,}09470 + 0{,}02525 + 0{,}00126 = 0{,}07386$. (**PR13.11**)

La valeur attendue $E(S)$ et la variance $V(S)$ du score S sous l'hypothèse nulle sont :

$$E(S) = \frac{5}{12}\left[3\times1 + 4\times2 + 5\times3\right] = 10{,}83$$

$$V(S) = \frac{5\times7}{12^2(12-1)}\left[\begin{matrix} 12\times\left(3\times1^2 + 4\times2^2 + 5\times3^2\right) \\ -\left(3\times1 + 4\times2 + 5\times3\right)^2 \end{matrix}\right] = 2{,}03$$

Ainsi, à titre de comparaison, le test de Mantel-Haenszel donne :

$$\chi^2 = \frac{(13-10{,}83)^2}{2{,}03} = 2{,}32 \text{ pour une valeur-}p \text{ unilatérale de } 0{,}06386.$$

TABLEAU 13.12

Score s*	Classe $R(s)$ des réalisations (a_1, a_2, a_3) équipotentes pour le score s		Probabilité sous H_0	
			Des réalisations (a_1, a_2, a_3)	Des scores s
7	$R(7)$:	(3, 2, 0)	0,00758	0,00758
8	$R(8)$:	(2, 3, 0)	0,01515	
		(3, 1, 1)	0,02525	0,04040
9	$R(9)$:	(1, 4, 0)	0,00379	
		(2, 2, 1)	0,11364	
		(3, 0, 2)	0,01263	0,13006
10	$R(10)$:	(1, 3, 1)	0,07576	
		(2, 1, 2)	0,15152	0,22728
11	$R(11)$:	(0, 4, 1)	0,00631	
		(1, 2, 2)	0,22727	
		(2, 0, 3)	0,03788	0,27166
12	$R(12)$:	(0, 3, 2)	0,05051	
		(1, 1, 3)	0,15152	0,20203
13	$R(13)$:	(0, 2, 3) ←	0,07576	
		(1, 0, 4)	0,01894	0,09470
14	$R(14)$:	(0, 1, 4)	0,02525	0,02525
15	$R(15)$:	(0, 0, 5)	0,00126	0,00126

* $s = x_1a_1 + x_2a_2 + x_3a_3$, ← Valeurs observées.

Dans le tableau 13.13, nous présentons les résultats calculés sous l'hypothèse $\beta = 1$. La valeur-p unilatérale à droite est de 0,58 et, dans la convention mi-p, de 0,42. Les données sont relativement compatibles avec l'hypothèse d'une pente $\beta = 1$. (**PR13.12**)

TABLEAU 13.13

Score s	Classe $R(s)$ des réalisations (a_1, a_2, a_3) équipotentes pour le score s		Probabilité sous l'hypothèse $\beta = 1$	
			Des réalisations (a_1, a_2, a_3)	Des scores s
7	$R(7)$:	(3, 2, 0)	0,00006	0,00006
8	$R(8)$:	(2, 3, 0)	0,00034	
		(3, 1, 1)	0,00057	0,00091
9	$R(9)$:	(1, 4, 0)	0,00023	
		(2, 2, 1)	0,00699	
		(3, 0, 2)	0,00078	0,00800
10	$R(10)$:	(1, 3, 1)	0,01267	
		(2, 1, 2)	0,02534	0,03801
11	$R(11)$:	(0, 4, 1)	0,00287	
		(1, 2, 2)	0,10331	
		(2, 0, 3)	0,01722	0,12340
12	$R(12)$:	(0, 3, 2)	0,06241	
		(1, 1, 3)	0,18722	0,24963
13	$R(13)$:	(0, 2, 3) ←	0,25446	
		(1, 0, 4)	0,06362	0,31808
14	$R(14)$:	(0, 1, 4)	0,23057	0,23057
15	$R(15)$:	(0, 0, 5)	0,03134	0,03134

CHAPITRE 14

ANALYSE APPARIÉE POUR LES ÉTUDES CAS-TÉMOINS

Considérons une étude de type cas-témoins qui porte sur l'association entre une maladie Y et un facteur dichotomique d'exposition X. Chacun des cas a été apparié à un certain nombre de témoins, pour un ou plusieurs facteurs à contrôler. Cet appariement définit K strates homogènes sur lesquelles on retrouve un cas et un certain nombre de témoins. Les sujets de chacune des strates sont alors classés comme exposés ($X = 1$) ou non exposés ($X = 0$), suivant ce que révèle l'observation.

Sur les données de cette étude, on décide de pratiquer une analyse appariée. Ce type d'analyse permet alors de mesurer l'association, d'établir un intervalle de confiance de cette mesure ou de tester cette association entre X et Y, en assurant un ajustement pour les facteurs d'appariement.

L'analyse appariée dépend du type d'appariement pratiqué. L'appariement le plus simple est celui où chaque sujet d'un groupe est assorti à un sujet spécifique de l'autre groupe, par exemple, un cas à un témoin. C'est l'appariement 1 à 1. On peut imaginer des situations plus complexes : chaque cas est apparié à 2 témoins (appariement 1 à 2), 1 cas à 3 témoins (appariement 1 à 3), et ainsi de suite. Pour chacune de ces situations, l'appariement est uniforme, au sens où chaque cas est apparié au même nombre de témoins. L'appariement peut être non uniforme si le nombre de témoins d'appariement varie d'un cas à l'autre.

Nous présentons dans les sections suivantes les méthodes d'analyse appariée pour la mesure du *RC* dans les études cas-témoins. La présentation couvrira quelques concepts fondamentaux, la description du *RC* dans le contexte des analyses appariées, les tests statistiques et les intervalles de confiance correspondants pour le *RC*. Elle se fera suivant différents niveaux d'appariement uniforme : 1 à 1, 1 à 2 et 1 à *r*.

14.1 APPARIEMENT 1 À 1

Chaque cas ($Y = 1$) est apparié à un témoin unique ($Y = 0$) pour un ou plusieurs facteurs de confusion. Suivant le facteur d'exposition considéré ($X = 1$ ou $X = 0$), les paires (Cas, Témoin) constituées peuvent être réparties en quatre groupes :

Pour chaque paire, nombre de sujets exposés à X	Paire (Cas, Témoin)	Nombre de paires f_{ij}
• les deux sujets	($X = 1, X = 1$)	f_{11}
• le cas seul	($X = 1, X = 0$)	f_{10}
• le témoin seul	($X = 0, X = 1$)	f_{01}
• aucun sujet	($X = 0, X = 0$)	f_{00}

Les deux cellules, f_{10} et f_{01}, sont dites des paires discordantes : elles regroupent les paires où le cas et le témoin ont subi des expositions différentes, ($X = 1, X = 0$) ou ($X = 0, X = 1$). On peut reprendre dans un tableau de fréquences la description de ces données (tableau 14.1). Chaque cellule du tableau décrit le nombre de paires d'une configuration donnée.

TABLEAU 14.1

		Nombre de témoins exposés	
		1	0
Cas	$X = 1$	f_{11}	f_{10}
	$X = 0$	f_{01}	f_{00}

Le rapport de cotes mesurant l'association entre X et Y est estimé simplement par le rapport du nombre de paires où seul le cas a été exposé au nombre de paires où seul le témoin a été exposé :

$$RC = \frac{f_{10}}{f_{01}} \tag{14.1}$$

On observe ici que seules les cellules discordantes comportent de l'information pertinente pour l'estimation de cette mesure d'association.

On peut montrer que ce rapport de cotes est une mesure ajustée par la méthode Mantel-Haenszel. À cette fin, considérons les *K* paires différentes comme autant de strates, chacune contenant exactement deux sujets. Les données d'une strate particulière *i* peuvent se présenter suivant un schéma standard (tableau 14.2).

TABLEAU 14.2

	Strate *i*		
	$X = 1$	$X = 0$	Total
$Y = 1$	a_i	b_i	m_{1i}
$Y = 0$	c_i	d_i	m_{0i}
Total	n_{1i}	n_{0i}	n_i

Rappelons que l'ajustement de Mantel-Haenszel (MH) pour le rapport de cotes dans les études cas-témoins (mais aussi dans les études de cohortes qui utilisent cette mesure) se présente comme :

$$RC_{MH} = \frac{\sum_i a_i d_i / n_i}{\sum_i b_i c_i / n_i}$$

Les paires se regroupent suivant quatre types de strates que l'on peut décrire à l'aide des tableaux 2×2 suivants :

TABLEAU 14.3

Type I : f_{11}				**Type II : f_{10}**			
	$X = 1$	$X = 0$	Total		$X = 1$	$X = 0$	Total
$Y = 1$	1	0	1	$Y = 1$	1	0	1
$Y = 0$	1	0	1	$Y = 0$	0	1	1
Total	2	0	2	Total	1	1	2
Type III : f_{01}				**Type IV : f_{00}**			
	$X = 1$	$X = 0$	Total		$X = 1$	$X = 0$	Total
$Y = 1$	0	1	1	$Y = 1$	0	1	1
$Y = 0$	1	0	1	$Y = 0$	0	1	1
Total	1	1	2	Total	0	2	2

En appliquant l'ajustement MH aux données du tableau 14.3, on obtient :

$$RC_{MH} = \frac{f_{11}(1\times 0)/2 + f_{10}(1\times 1)/2 + f_{01}(0\times 0)/2 + f_{00}(0\times 1)/2}{f_{11}(0\times 1)/2 + f_{10}(0\times 0)/2 + f_{01}(1\times 1)/2 + f_{00}(1\times 0)/2}$$
$$= \frac{f_{10}}{f_{01}}$$

14.1.1 TESTS STATISTIQUES

Les tests statistiques qui sont présentés ici sont ceux rattachés au *RC*. L'hypothèse testée est celle d'un *RC* égal à 1 (hypothèse nulle). La définition de ces tests n'implique que les cellules discordantes. Nous présentons d'abord le test exact défini dans le cadre de la loi binomiale, puis le test en approximation normale (test de McNemar) et, enfin, le test du rapport de vraisemblance (RV).

TEST EXACT

Le test exact est basé sur la loi binomiale définie sur la diagonale des paires discordantes.

Sous l'hypothèse nulle ($RC = 1$), la variable F_{10} (correspondant à la cellule f_{10}) obéit à une loi binomiale de paramètre p ($\pi = 1/2$) et $n = f_{10} + f_{01}$. Le test exact unilatéral à droite se définit, dans la convention mi-p, par :

$$p = 1/2\ C_n^{f_{10}}(1/2)^n + \sum_{u=f_{10}+1}^{n} C_n^u (1/2)^n$$

On peut définir de façon analogue le test unilatéral à gauche.

Le test exact bilatéral se définit par convention comme :

$$p = (1/2)^{1-\delta} C_n^{f_{10}} (1/2)^n + \sum_u C_n^u (1/2)^n$$

où la sommation est faite sur toutes les valeurs u plus extrêmes que f_{10}.

S'il existe une valeur u également extrême à f_{10} (dans ce cas $\delta = 1$; autrement $\delta = 0$), sa probabilité est traitée de même façon que celle de f_{10} avant d'être cumulée dans la valeur-p.

On peut définir un test binomial pour une hypothèse quelconque sur le RC. Sous l'hypothèse $RC = \psi$, la variable F_{10} obéit à une variable binomiale de paramètres $\pi = \dfrac{\psi}{\psi+1}$ et n. Le test se présente alors comme :

$$p = (1/2)^{1-\delta} C_n^{f_{10}} \left(\frac{\psi}{\psi+1}\right)^{f_{10}} \left(\frac{1}{\psi+1}\right)^{n-f_{10}} + \sum_u C_n^u \left(\frac{\psi}{\psi+1}\right)^u \left(\frac{1}{\psi+1}\right)^{n-u}$$

Tests en approximation normale

Test de McNemar

Le test approximatif approprié est celui de McNemar, défini à partir de la variable binomiale F_{10}. La variable binomiale F_{10} a comme moyenne $n/2$ et comme variance $n/4$. On peut alors définir la statistique Z comme

$$Z = \frac{F_{10} - \dfrac{n}{2}}{\sqrt{\dfrac{n}{4}}} = \frac{2F_{10} - n}{\sqrt{n}}$$

Le carré de la variable Z obéit à une loi du khi-carré. En remplaçant la variable F_{10} par sa valeur observée f_{10}, on obtient alors une valeur du khi-carré à partir de laquelle on peut facilement déduire la valeur-p :

$$\chi_1^2 = \frac{(2f_{10} - n)^2}{n}$$

ce qui s'écrit aussi comme :

$$\chi_1^2 = \frac{(f_{10} - f_{01})^2}{f_{10} + f_{01}} \tag{14.2}$$

Le test de McNemar peut aussi s'exprimer en fonction du *RC* :

$$\chi_1^2 = \frac{f_{01}\left[RC - 1\right]^2}{RC + 1} \tag{14.3}$$

Test de Wald (basé sur le log du RC)

Ce test se présente simplement comme :

$$\chi_1^2 = \frac{\left[\log(\psi)\right]^2}{V[\log(\psi)]}$$

où ψ est estimé par la méthode du maximum de vraisemblance.

On peut facilement déduire la variance de $\log(\psi)$ par la méthode delta :

$$\begin{aligned} V[\log(\psi)] &= V\left[\log f_{10}\right] + V\left[\log f_{01}\right] + 2\,\mathrm{cov}\left[\log f_{10}, -\log f_{01}\right] \\ &= \frac{1}{{f_{10}}^2} \times \frac{f_{10}(n - f_{10})}{n} + \frac{1}{{f_{01}}^2} \times \frac{f_{01}(n - f_{01})}{n} \\ &\quad + 2 \times \frac{1}{f_{10} f_{01}} \times \frac{f_{10} f_{01}}{n} \\ &= \frac{f_{10} + f_{01}}{f_{10} f_{01}} \end{aligned}$$

La variance étant connue, le calcul du test en découle directement.

Nous pouvons nous en remettre à la procédure PHREG qui fournit non seulement une estimation de ψ et de sa variance, mais aussi la valeur du test de Wald lui-même.

Test du rapport de vraisemblance

Le test du rapport de vraisemblance se définit à partir de la comparaison des deux fonctions de vraisemblance construites sur les données des cellules discordantes : la fonction FV_0, correspondant à l'hypothèse nulle, et FV_1, correspondant au maximum de vraisemblance.

$$FV_0 = C_n^{f_{10}} (1/2)^n$$

$$FV_1 = C_n^{f_{10}} \left(\frac{f_{10}}{n} \right)^{f_{10}} \left(\frac{f_{01}}{n} \right)^{f_{01}}$$

On rappelle alors que $2 \log \frac{FV_1}{FV_0}$ obéit à une loi du khi-carré avec, ici, 1 degré de liberté.

Le test se décrit alors comme :

$$\chi_1^2 = 2 \log \frac{FV_1}{FV_0} = 2 \left[f_{10} \log \left(\frac{f_{10}}{f_{01}} \right) - n \log \left(\frac{n}{2 f_{01}} \right) \right]$$

En termes du rapport de cotes ψ (estimateur du maximum de vraisemblance), ce test peut s'écrire comme suit :

$$\chi_1^2 = 2 \log \frac{FV_1}{FV_0} = 2 \left[f_{10} \log \psi - n \log \left(\frac{\psi + 1}{2} \right) \right] \qquad (14.4)$$

Exemple 14.1

Considérons une étude cas-témoins portant sur l'association entre un facteur *X* et la maladie *Y*. Au total, chacun des 36 cas est apparié à un témoin spécifique pour certaines variables potentiellement confondantes. L'étude génère ainsi 4 paires (cas, témoin) où les deux sujets sont exposés à *X*, 18 paires où seul le cas est exposé, 6 paires où seul le témoin est exposé et enfin 8 paires où ni le cas ni le témoin n'est exposé à *X*. Les données sont présentées au tableau 14.4.

Tableau 14.4

		Nombre de témoins exposés		Total
		X = 1	*X* = 0	
Cas	*X* = 1	4	18	22
	X = 0	6	8	14
	Total	10	26	36

Sur ces données appariées, nous appliquons successivement les trois tests définis précédemment.

Test exact bilatéral

Remarquons d'abord que sous l'hypothèse $\pi_1 = 0{,}5$, $P(A_1 = 18) = 0{,}0080225$. La valeur 6 est une valeur également extrême, située du côté gauche de la distribution de A_1. Sous la même hypothèse, les valeurs 19, 20, 21, 22, 23 et 24

sont plus extrêmes que 18. Il en va aussi pour les valeurs 0, 1, 2, 3, 4 et 5, qui se retrouvent à l'autre extrémité de la distribution de A_1. La valeur-*p* du test binomial bilatéral sera donc égale à :

$$\begin{aligned} p &= 1/2P(A_1 = 18) + 1/2P(A_1 = 6) \\ &\quad + \sum_{u=19}^{24} P(A_1 = u) + \sum_{u=0}^{5} P(A_1 = u) \\ &= 0,004011 + 0,004011 + 0,003305 + 0,003305 \\ &= 0,01463 \end{aligned}$$

Ce résultat étaye l'hypothèse d'une association entre le facteur *X* et la maladie *Y*.

Tests en approximation normale

Test de McNemar

Pour les données du tableau 14.4, $r = 18$ et $n = 24$. Alors,

$$\begin{aligned} \chi_1^2 &= \frac{(2f_{10} - n)^2}{n} \\ &= \frac{(2 \times 18 - 24)^2}{24} \\ &= 6 \end{aligned}$$

La valeur-*p* correspondante est de 0,0143. Ce résultat est très concordant avec le précédent.

Test de Wald

Pour le test de Wald, on a :

- $\log(\psi) = \log\left(\frac{18}{6}\right) = \log(3) = 1,0986$
- $V\left[\log(\psi)\right] = \frac{(18+6)}{18 \times 6} = 0,2222$

Ainsi, $\chi_1^2 = \left(\frac{1,0986}{0,2222}\right)^2 = 5,4313$, pour une valeur-*p* de 0,0198.

Test du rapport de vraisemblance

$$\begin{aligned} \chi_1^2 &= 2\left[f_{10} \log\left(\frac{f_{10}}{f_{01}}\right) - n_1 \log\left(\frac{n}{2f_{01}}\right)\right] \\ &= 2\left[18 \log\left(\frac{18}{6}\right) - 24 \log\left(\frac{24}{2 \times 6}\right)\right] \\ &= 6,2790 \end{aligned}$$

La valeur-*p* correspondante est de 0,0122. Ce résultat est concordant avec les précédents.

Dans le tableau 14.5, nous reprenons les résultats des différents tests conduits sur les données du tableau 14.4.

TABLEAU 14.5

Test statistique	Khi-carré	Valeur-*p* bilatérale	Programme
Binomial	–	0,0146	PR14.1
McNemar	6,0000	0,0143	PR14.2
De Wald	5,4313	0,0198	PR14.3
Du rapport de vraisemblance	6,2790	0,0122	PR14.4

◆

14.1.2 INTERVALLE DE CONFIANCE DU *RC*

Nous proposons trois approches pour le calcul de l'intervalle de confiance du rapport de cotes ψ : le calcul exact, le calcul en approximation normale et le calcul par la méthode du rapport de vraisemblance (RV).

INTERVALLE DE CONFIANCE EXACT

L'intervalle de confiance exact de niveau $100(1-\alpha)\,\%$ se calcule dans le cadre de la loi binomiale. Ainsi, rappelant que F_{10} est une variable binomiale Bin[(π_1,n] où $n = f_{10} + f_{01}$ et $\pi = \psi/(1+\psi)$, les limites de confiance de y sont alors définies comme :

ψ_{inf} est la solution de l'équation $\displaystyle\sum_{x=f_{10}}^{n} C_n^x \pi^x (1-\pi)^{n-x} = \frac{\alpha}{2}$

ψ_{sup} est la solution de l'équation $\displaystyle\sum_{x=0}^{f_{10}} C_n^x \pi^x (1-\pi)^{n-x} = \frac{\alpha}{2}$

Pour le calcul de ces limites, nous proposons d'utiliser l'approche dans la convention mi-*p*.

INTERVALLE DE CONFIANCE EN APPROXIMATION NORMALE

EN CONCORDANCE AVEC LE TEST DE MCNEMAR

La variable F_{10} est une variable binomiale de paramètres $\pi = \dfrac{\psi}{1+\psi}$ et n.

La valeur attendue $E(F_{10})$ et la variance $V(F_{10})$ sont respectivement données par :

$$E(F_{10}) = \frac{n\psi}{\psi + 1} \text{ et } V(F_{10}) = \frac{n\psi}{(\psi + 1)^2}$$

Nous recherchons alors les valeurs ψ_{inf} et ψ_{sup} de ψ telles que :

$P(F_{10} \geq f_{10} \mid \psi_{inf}) = \alpha/2$ et $P(F_{10} \leq f_{10} \mid \psi_{sup}) = \alpha/2$

En d'autres termes, nous voulons déterminer les valeurs ψ_{inf} et ψ_{sup} de ψ telles que :

$$\left| \frac{f_{10} - E(F_{10})}{\sqrt{V(F_{10})}} \right| = z_{\alpha/2}$$

ou encore

$$\frac{\left[f_{10} - E(F_{10})\right]^2}{V(F_{10})} = \chi^2_{1,1-\alpha}$$

En substituant dans cette dernière relation les valeurs de $E(F_{10})$ et $V(F_{10})$ précédemment définies, on obtient une équation quadratique en ψ dont les racines correspondent aux deux limites recherchées :

$$f_{01}^2\psi^2 - [2f_{10}f_{01} + n\chi^2_{1,1-\alpha}]\psi + f_{10}^2 = 0$$

Les racines, et donc les limites de confiance, exprimées en fonction du *RC* observé, se présentent simplement comme :

$$\psi_{inf} = RC + U\left[1 - \sqrt{\frac{2RC}{U} + 1}\right]$$

$$\psi_{sup} = RC + U\left[1 + \sqrt{\frac{2RC}{U} + 1}\right]$$

où $U = \dfrac{\chi^2_{1,1-\alpha}}{2f_{01}}(RC + 1)$.

Par la transformation logarithmique du RC (méthode de Wald)

Les limites de confiance peuvent aussi être obtenues par approximation normale sur la transformation logarithmique du *RC*. Les limites sont d'abord calculées pour $\log(RC)$ (ou $\log(\psi)$). Puis, par transformation inverse, on obtient celles du rapport de cotes ψ :

$$\psi_{inf} = RC \times e^{-z_{\alpha/2}\sqrt{V(\log RC)}}$$

$$\psi_{sup} = RC \times e^{+z_{\alpha/2}\sqrt{V(\log RC)}}$$

où $V(\log RC) = \dfrac{f_{10} + f_{01}}{f_{10} f_{01}}$.

MÉTHODE BASÉE SUR LE RÉSULTAT D'UN TEST

La technique est la même que celle déjà décrite en analyse simple (section 6.3.1 du chapitre 6). Les limites de confiance du rapport de cotes ψ sont alors données par :

$$\psi_{\text{inf}} = RC^{\left(1 - \frac{z_{\alpha/2}}{\chi}\right)}$$

$$\psi_{\text{sup}} = RC^{\left(1 + \frac{z_{\alpha/2}}{\chi}\right)}$$

où $\chi = \sqrt{\chi^2(\text{McNemar})}$.

INTERVALLE DE CONFIANCE PAR LA MÉTHODE DU RAPPORT DE VRAISEMBLANCE

On peut déduire les limites de confiance du rapport de cote ψ par la méthode du rapport de vraisemblance en solutionnant pour ψ l'équation suivante :

$$2[L - L(\psi)] - \chi^2_{1,1-\alpha} = 0$$

Rappelons que L représente le logarithme de la fonction de vraisemblance évaluée en son maximum et que $L(\psi)$ représente le logarithme de la fonction de vraisemblance pour une valeur de ψ quelconque. Pour $\alpha = 0{,}05$, la valeur de $\chi^2_{1,1-\alpha}$ est de 3,84.

La résolution d'une telle équation ne peut se faire que par des méthodes itératives, déjà disponibles dans les procédures informatiques appropriées.

EXEMPLE 14.2

Pour les données du tableau 14.4, nous rappelons que le rapport de cotes est de 3.

Dans le tableau 14.6, nous décrivons les intervalles de confiance obtenus par les différentes méthodes présentées. Pour la méthode basée sur la valeur d'un test, nous utilisons la valeur du test de McNemar.

TABLEAU 14.6

Méthode	Intervalle de confiance (95 %)	Programme SAS
Exacte	[1,2261 ; 8,2519]	PR14.5
McNemar	[1,2272 ; 7,3339]	PR14.6
De Wald	[1,1908 ; 7,5577]	PR14.7
Basée sur le résultat d'un test	[1,2455 ; 7,2260]	PR14.8
RV	[1.2578 ; 8.2734]	PR14.9

Nous remarquons que les méthodes exactes et du rapport de vraisemblance sont très concordantes. Elles nous apparaissent les plus correctes. Les méthodes de McNemar et de Wald, sujettes à l'hypothèse de la normalité, ne conduisent à des résultats acceptables qu'avec des échantillons de bonne taille. La méthode basée sur le résultat d'un test conduit à un intervalle sensiblement plus étroit que les autres, traduisant pour la mesure une stabilité plus grande qu'elle n'est en réalité. Bien que cette méthode soit d'application très simple, elle nous apparaît de dernier recours.

◆

14.2 APPARIEMENT 1 À 2

À chaque cas sont appariés deux témoins. Les observations sont faites pour l'exposition à un facteur. Pour K triplets ainsi constitués, on dispose les données dans un tableau 2×3 dont les lignes spécifient l'état du cas quant à l'exposition et les colonnes spécifient l'état des témoins du triplet correspondant (tableau 14.7). Chaque cellule du tableau représente donc le nombre de triplets d'une configuration donnée.

TABLEAU 14.7

		Nombre de témoins exposés		
		2	1	0
Cas	$X = 1$	f_{12}	f_{11}	f_{10}
	$X = 0$	f_{02}	F_{01}	f_{00}

Dans ce tableau, par exemple, f_{12} représente le nombre de triplets où le cas et les deux témoins sont exposés au facteur X. À partir de ces données, on peut estimer le rapport de cotes mesurant l'association entre le facteur d'exposition et la maladie par :

$$RC = \frac{f_{11} + 2f_{10}}{2f_{02} + f_{01}}$$

Ce rapport de cotes correspond au rapport de cotes ajusté par la méthode Mantel-Haenszel.

14.2.1 TESTS STATISTIQUES

Comme c'est le cas pour la section 14.1.1, les tests statistiques qui sont présentés ici sont directement rattachés au *RC* (ou ψ). L'hypothèse testée est celle d'un *RC* égal à 1 ($\psi = 1$). La définition de ces tests n'implique que les cellules discordantes. Nous présentons d'abord le test exact défini dans le cadre de la loi binomiale, puis le test en approximation normale (test de McNemar) et, enfin, le test du rapport de vraisemblance.

TEST EXACT

Le test considéré doit permettre de calculer, sous l'hypothèse nulle, la probabilité que le nombre *A* de cas exposés sur les cellules discordantes soit égal ou supérieur à la valeur observée a ($= f_{10} + f_{11}$). Pour ce test unilatéral à droite, la valeur-*p* correspond à $p = P(A \geq f_{10} + f_{11})$

La variable *A* est la somme des variables F_{10} et F_{11}. Chacune des variables F_{10} et F_{11} est une variable binomiale.

- F_{10} représente le nombre de triplets où seul le cas est exposé. Rapportée à la diagonale (I) (correspondant aux triplets de la forme f_{10} et f_{01}) qui regroupe les n_1 triplets où 1 seul sujet est exposé, F_{10} peut être considérée comme une variable binomiale de paramètres n_1 et π_1. Sous l'hypothèse nulle, le paramètre π_1 correspond à la probabilité qu'un triplet soit de la forme « le cas est exposé » parmi les n_1 triplets de la diagonale I. Ainsi, $\pi_1 = 1/3$.
- F_{11} représente le nombre de triplets où le cas est exposé, tout comme l'un des deux témoins. Rapportée à la diagonale (II) (correspondant aux triplets de la forme f_{11} et f_{02}) qui regroupe les n_2 triplets où 2 sujets sont exposés, F_{11} peut être considérée comme une variable binomiale de paramètres n_2 et π_2. Sous l'hypothèse nulle, le paramètre π_2 correspond à la probabilité qu'un triplet soit de la forme « le cas et un des témoins sont exposés » parmi les n_2 triplets de la diagonale II. Ainsi, $\pi_2 = 2/3$.

Le calcul de la valeur-*p* se fait donc en traitant la variable *A* comme la somme de deux variables binomiales indépendantes. On détermine les valeurs de *A* également et plus extrêmes que la valeur observée a ($= f_{10} + f_{11}$) et dont les probabilités doivent être cumulées pour obtenir la valeur-*p* en maintenant fixes les effectifs sur les diagonales correspondant aux

variables binomiales. Pour la valeur particulière a de A, qui est la somme des nombres f_{10} de la diagonale I et f_{11} de la diagonale II, la probabilité $P(A = a)$ se calcule simplement comme suit :

$$\begin{aligned} P(A=a) &= P(F_{10}=f_{10}) \times P(F_{11}=f_{11}) \\ &= C_{n_1}^{f_{10}} \pi_1^{f_{10}} (1-\pi_1)^{f_{01}} \times C_{n_2}^{f_{11}} \pi_2^{f_{11}} (1-\pi_2)^{f_{02}} \end{aligned}$$

Nous pouvons décrire ce test exact (unilatéral à droite) par

$$p = \sum_{j=a}^{n} \left(\sum_{i=\inf}^{\sup} C_{n_1}^{i} \pi_1^{i} (1-\pi_1)^{n_1-i} \times C_{n_2}^{j-i} \pi_2^{j-i} (1-\pi_2)^{(n_2-j+i)} \right) \qquad (14.5)$$

où $a = f_{10} + f_{11}$, $n = n_1 + n_2$, $\inf = \max(j - n_2, 0)$ et $\sup = \min(j, n_1)$.

L'application du test sur un exemple numérique simple pourra mieux faire comprendre la procédure (voir l'exemple 14.3).

Tests en approximation normale

Test de McNemar

Un test en approximation normale, et donc un test du khi-carré, est facile à construire à partir de la variable A. Rappelons que A est la somme de deux variables binomiales indépendantes F_{10} et F_{11}. Sous l'hypothèse nulle, les paramètres de chacune de ces variables se décrivent comme suit :

Pour la variable F_{10} : $\pi_1 = 1/3$ et $n_1 = f_{10} + f_{01}$.

Pour la variable F_{11} : $\pi_2 = 2/3$ et $n_2 = f_{11} + f_{02}$.

Ainsi, sous l'hypothèse nulle, on peut aisément en déduire la valeur attendue $E(A)$ et la variance $V(A)$ de la variable A :

$$\begin{aligned} E(A) &= (n_1 \times \pi_1) + (n_2 \times \pi_2) \\ &= (n_1 \times 1/3) + (n_2 \times 2/3) \\ V(A) &= [n_1 \times \pi_1 \times (1-\pi_1)] + [n_2 \times \pi_2 \times (1-\pi_2)] \\ &= [n_1 \times 1/3 \times 2/3] + [n_2 \times 2/3 \times 1/3] \end{aligned}$$

Par approximation normale, on obtient donc le test de McNemar :

$$\chi_1^2 = \frac{(a - E(A))^2}{V(A)}$$

En remplaçant $E(A)$ et $V(A)$ par leurs expressions respectives, le test du khi-carré peut se présenter comme :

$$\chi_1^2 = \frac{\left[3(f_{10}+f_{11})-(n_1+2n_2)\right]^2}{2(n_1+n_2)}$$
$$= \frac{\left[(2f_{10}+f_{11})-(f_{01}+2f_{02})\right]^2}{2(n_1+n_2)}$$

Test de Wald

Le test de Wald se présente simplement comme :

$$\chi_1^2 = \frac{\left[\log(RC)\right]^2}{V[\log(RC)]}$$

Lorsque l'appariement est de 2 témoins ou plus par cas, on ne dispose de formule simple ni pour l'estimation du maximum de vraisemblance du rapport de cotes *RC* ni pour la variance V[log(*RC*)]. Nous nous en remettons donc à la procédure PHREG qui fournit non seulement une estimation pour cette variance, mais aussi la valeur du test de Wald et du SCORE.

Test du rapport de vraisemblance

Le test du rapport de vraisemblance pour les analyses appariées 1 à 2 se définit théoriquement de façon analogue à celui pour l'appariement 1 à 1. On compare les deux fonctions de vraisemblance suivantes : FV_0, correspondant à l'hypothèse nulle ($\psi = 1$), et FV_1, correspondant à l'hypothèse de l'estimateur du maximum de vraisemblance de ψ.

$$FV_0 = C_{n_1}^{f_{10}}(1/3)^{f_{10}}(2/3)^{f_{01}} \times C_{n_2}^{f_{11}}(2/3)^{f_{11}}(1/3)^{f_{02}}$$
$$FV_1 = C_{n_1}^{f_{10}}\left(\frac{\psi}{\psi+2}\right)^{f_{10}}\left(\frac{2}{\psi+2}\right)^{f_{01}} \times C_{n_2}^{f_{11}}\left(\frac{2\psi}{2\psi+1}\right)^{f_{11}}\left(\frac{1}{2\psi+1}\right)^{f_{02}} \quad (14.6)$$

La statistique $2\log\frac{FV_1}{FV_0}$ obéit à une loi du khi-carré avec, ici, 1 degré de liberté.

En substituant dans cette statistique les valeurs de FV_1 et FV_0, on obtient l'expression suivante pour le test du rapport de vraisemblance :

$$\chi_1^2 = 2\log\frac{FV_1}{FV_0} = 2\left[\begin{array}{l}(f_{11}+f_{10})\log\psi + n_1\log\left(\dfrac{3}{\psi+(3-1)}\right)\\ +n_2\log\left(\dfrac{3}{2\psi+(3-2)}\right)\end{array}\right]$$

L'estimateur du maximum de vraisemblance du rapport de cotes ψ s'obtient en solutionnant l'équation quadratique de vraisemblance déduite à partir de la fonction de vraisemblance décrite à l'expression (14.6). En effet, on a :

$$\frac{\partial\log[FV_1(\psi)]}{\partial\psi} = \frac{f_{10}}{\psi} + \frac{f_{11}}{\psi} - \frac{n_1}{\psi+2} - \frac{2n_2}{2\psi+1}$$

En annulant cette dérivée, on obtient l'équation quadratique suivante :

$$2(a-n_1-n_2)\psi^2 + (5a-n_1-4n_2)\psi + 2a = 0$$

où $a = f_{10} + f_{11}$.

La racine positive de cette équation quadratique correspond au ψ recherché :

$$\psi = \frac{(5a-n_1-4n_2)+\sqrt{(3a+n_1+4n_2)^2-48an_2}}{4(n_1+n_2-a)}$$

Nous pouvons aussi l'obtenir, plus facilement, par la procédure PHREG de SAS adaptée à ce genre d'analyse.

EXEMPLE 14.3

Considérons une étude cas-témoins portant sur l'association entre le facteur X et la maladie Y. Chaque cas a été apparié à deux témoins. Les résultats sont présentés dans le tableau suivant.

TABLEAU 14.8

I_7		Nombre de témoins exposés à X		
		2	**1**	**0**
Cas	$X = 1$	6	4	3
	$X = 0$	1	2	5

On veut mesurer le *RC* décrivant l'association entre *Y* et *X* et tester cette association.

Mesure du RC

Estimation de Mantel-Haenszel

Le rapport de cotes est estimé comme :

$$RC = \frac{f_{11} + 2f_{10}}{2f_{02} + f_{01}}$$
$$= \frac{4 + (2 \times 3)}{(2 \times 1) + 2}$$
$$= 2,5$$

Estimation du maximum de vraisemblance du RC

$$\psi = \frac{(5a - n_1 - 4n_2) + \sqrt{(3a + n_1 + 4n_2)^2 - 48an_2}}{4(n_1 + n_2 - a)}$$
$$= \frac{(5 \times 7 - 5 - 4 \times 5) + \sqrt{(3 \times 7 + 5 + 4 \times 5)^2 - 48 \times 7 \times 5}}{4(5 + 5 - 7)}$$
$$= 2,5734$$

Test exact

Le test exact doit conduire au calcul de la probabilité $P(A \geq 7)$.

Pour les données du tableau 14.8, la valeur de la variable *A* est donnée par : $A = 4 + 3 = 7$ et la probabilité $P(A = 7)$ est donnée par $P(A = 4 + 3) = P(F_{11} = 4) \times P(F_{10} = 3)$.

Or, F_{11} et F_{10} sont des variables binomiales respectivement de paramètres $n_2 = 4 + 1$ et $\pi_2 = 2/3$, et $n_1 = 3 + 2$, $\pi_1 = 1/3$.

Ainsi :

$$P(F_{11} = 4) = C_5^4 \frac{2}{3}^4 \frac{1}{3}^1 = 0,3292$$

$$P(F_{10} = 3) = C_5^3 \frac{1}{3}^3 \frac{2}{3}^2 = 0,1646$$

et $P(A = 4 + 3) = 0,3292 \times 0,1646 = 0,05419$.

Mais il faut bien compter que le tableau 14.8 n'est pas la seule configuration (I_7) qui, sous la condition des diagonales fixes, conduira à $A = 7$. Il a plusieurs autres configurations qui donneront également la valeur $A = 7$, mais aussi des valeurs plus extrêmes. Disposant de la probabilité $P(A = 7)$ pour la configuration I_7, nous présentons également les probabilités d'obtenir une valeur de *A* également extrême ou plus extrême que 7 suivant les autres configurations (tableau 14.9).

TABLEAU 14.9

6 5 2 0 3 5 $P(A = 7) = 0{,}04335$	II_7	6 5 3 0 2 5 $P(A = 8) = 0{,}02168$	I_8	6 5 4 0 1 5 $P(A = 9) = 0{,}00054$	I_9
6 3 4 2 1 5 $P(A = 7) = 0{,}01355$	III_7	6 4 4 1 1 5 $P(A = 8) = 0{,}00135$	II_8	6 4 5 1 0 5 $P(A = 9) = 0{,}00135$	II_9
6 2 5 3 0 5 $P(A = 7) = 0{,}00068$	IV_7	6 3 5 2 0 5 $P(A = 8) = 0{,}00135$	III_8	6 5 5 0 0 5 $P(A = 10) = 0{,}00054$	I_{10}

La valeur-*p* finale $P(A \geq 7)$ est donnée par la somme des probabilités sur toutes ces configurations également ou plus extrêmes que 7.

$$\begin{aligned} p &= P(A \geq f_{10} + f_{11}) \\ &= P(I_7) + P(II_7) + P(III_7) + P(IV_7) + P(I_8) + \ldots + P(I_{10}) \\ &= 0{,}05419 + 0{,}04335 + 0{,}01355 + 0{,}00068 + \ldots + 0{,}00054 \\ &= 0{,}1557 \end{aligned}$$

Dans la convention mi-*p*, la valeur-$p = 0{,}10$.

Nous renvoyons le lecteur à un petit programme SAS qui fait très bien ces calculs. (**PR14.10**)

EXEMPLE 14.4

Considérons les données suivantes issues d'une étude cas-témoins, où chacun des 54 cas a été apparié à 2 témoins.

TABLEAU 14.10

		Nombre de témoins exposés		
		2	**1**	**0**
Cas	$X = 1$	4	18	16
	$X = 0$	6	8	2

MESURE DU *RC*

ESTIMATION DE MANTEL-HAENSZEL

Le rapport de cotes Mantel-Haenszel est estimé comme :

$$\begin{aligned} RC &= \frac{f_{11} + 2f_{10}}{2f_{02} + f_{01}} \\ &= \frac{18 + (2 \times 16)}{(2 \times 6) + 8} \\ &= 2{,}5 \end{aligned}$$

ESTIMATION DU MAXIMUM DE VRAISEMBLANCE

$$\psi = \frac{(5a - n_1 - 4n_2) + \sqrt{(3a + n_1 + 4n_2)^2 - 48an_2}}{4(n_1 + n_2 - a)}$$
$$= \frac{(5 \times 34 - 24 - 4 \times 24) + \sqrt{(3 \times 34 + 24 + 4 \times 24)^2 - 48 \times 34 \times 24}}{4(24 + 24 - 34)}$$
$$= 2,6889$$

TEST EXACT UNILATÉRAL À DROITE

En utilisant l'expression (14.5) où $a = 18 + 16$, $n_1 = n_2 = 24$ et $n = 48$, on obtient 0,0016279 pour la valeur-p unilatérale à droite suivante dans la convention intégrale. Dans le convention mi-p, on a valeur-p = 0,001091.

TESTS EN APPROXIMATION NORMALE

TEST DE MCNEMAR

$\chi_1^2 = \frac{[3(16+18) - (24 + 2 \times 24)]^2}{2(24+24)} = 9,375$, valeur qui correspond à valeur-p = 0,002.

L'hypothèse nulle n'est pas du tout compatible avec ces données.

TEST DE WALD

$$\chi_1^2 = \frac{[\log(2,689)]^2}{0,33253^2} = 8,848$$

TEST DU RAPPORT DE VRAISEMBLANCE

L'estimateur du maximum de vraisemblance du RC est approximativement de 2,689.

En utilisant cette valeur, on a :

$$\chi_1^2 = 2\log\frac{FV_1}{FV_0} = 2\left[\begin{array}{l}(16+18)\log 2,689 + 24\log\left(\frac{3}{2,689+2}\right) \\ +24\log\left(\frac{3}{2 \times 2,689+1}\right)\end{array}\right]$$
$$= 9,623$$

La valeur de ce test est sensiblement la même que celle obtenue par le test de McNemar.

On résume dans le tableau 14.11 les résultats des différents tests sur ces données en donnant la référence pour les programmes SAS.

TABLEAU 14.11

Test	χ^2_1	p	Programme SAS
Exact (unilatéral à droite)	–	0,0011	PR14.11
De McNemar	9,3750	0,0011*	PR14.12
De Wald	8,8480	0,0014*	PR14.13
RV	9,6227	0,0010*	

* On convertit la valeur-p en valeur unilatérale en faisant mi-p.

Rappelons que la procédure PHREG permet d'obtenir facilement tant l'estimation du maximum de vraisemblance du *RC* que la valeur des tests de Wald, de McNemar (du score), et du rapport de vraisemblance.

◆

14.2.2 INTERVALLE DE CONFIANCE DU *RC*

Pour le calcul de l'intervalle de confiance du rapport de cotes, nous retenons la méthode exacte, deux méthodes en approximation normale : celle basée sur le résultat d'un test et celle de Wald, et la méthode du rapport de vraisemblance.

INTERVALLE DE CONFIANCE EXACT

L'intervalle de confiance exact de niveau $100(1-\alpha)\,\%$ se calcule dans le cadre où la variable A est considérée comme la somme de deux variables binomiales F_{10} et F_{11}.

Les paramètres de ces variables se décrivent respectivement comme suit :

$$\text{Pour } F_{10}:\ \pi_1 = \frac{\psi}{2+\psi} \text{ et } n_1 = f_{10} + f_{01}.$$

$$\text{Pour } F_{11}:\ \pi_2 = \frac{2\psi}{1+2\psi} \text{ et } n_2 = f_{11} + f_{02}.$$

Nous recherchons alors les valeurs ψ_{inf} et ψ_{sup} de ψ telles que, respectivement :

$$P(A \geq f_{10} + f_{11} \mid \psi) = \frac{\alpha}{2} \text{ et } P(A \leq f_{10} + f_{11} \mid \psi) = \frac{\alpha}{2}$$

Ces probabilités sont calculées à partir de l'expression (14.5).

Pour le calcul de ces limites, nous évitons l'approche « mi-p » pour faciliter la programmation en SAS.

Intervalle de confiance en approximation normale

Tant pour la méthode basée sur le résultat d'un test que pour celle de Wald, les procédures de calcul sont analogues à celles déjà présentées à la section 14.1.2. Les limites de confiance obtenues par la méthode de Wald sont directement accessibles dans la procédure PHREG.

Intervalle de confiance par la méthode du rapport de vraisemblance

On utilise ici la fonction de vraisemblance (conditionnelle) :

$$L(\psi) = a\log(\psi) - n_1\log(\psi + 2) - n_2\log(2\psi + 1) + K$$

où K est une valeur indépendante de ψ. Cette fonction est déduite de la relation (14.6).

La fonction $L(\psi)$ atteint son maximum lorsque ψ prend la valeur de l'estimation du maximum de vraisemblance ψ_M :

$$L = a\log(\psi_M) - n_1\log(\psi_M + 2) - n_2\log(2\psi_M + 1) + K$$

Les limites de l'intervalle correspondent alors aux deux valeurs de ψ qui satisfont l'équation logarithmique. Ces valeurs sont obtenues par itération.

Exemple 14.5

Revoyons les données du tableau 14.10. Le rapport de cotes de Mantel-Haenszel estimé sur ce tableau est de 2,5 et la valeur du khi-carré de McNemar est de 9,375. Rappelons aussi que le rapport de cotes estimé par la méthode du maximum de vraisemblance est de 2,689.

Dans le tableau 14.12, nous décrivons les intervalles de confiance obtenus par les différentes méthodes présentées.

Tableau 14.12

Méthode	*RC*	Intervalle de confiance (95 %)	Programme SAS
Exacte	–	[1,356 ; 4,935]	Pr14.14
De Wald	2,689	[1,401 ; 5,160]	Pr14.15
Résultat d'un test	2,500	[1,391 ; 4,494]	Pr14.16
RV*	2,689	[1,429 ; 5,312]	Pr14.17

* La procédure PHREG ne permet pas d'obtenir les limites de confiance par la méthode du rapport de vraisemblance.

On remarque ici une très bonne concordance entre les méthodes de Wald et du RV.

◆

14.3 APPARIEMENT 1 À *R*

À chaque cas sont appariés *r* témoins. Les observations sont faites pour l'exposition à un facteur *X*. Pour tous les $(r + 1)$-uplets constitués, on dispose les données dans un tableau $2 \times (r + 1)$ dont les lignes spécifient l'état du cas quant à l'exposition et les colonnes spécifient l'état des témoins du $(r + 1)$-uplet correspondant. Chaque cellule du tableau représente donc le nombre de $(r + 1)$-uplets d'une configuration donnée.

TABLEAU 14.13

		Nombre de témoins exposés					
		r	$r-1$	...	2	1	0
Cas	$X = 1$	f_{1r}	$f_{1(r-1)}$	...	f_{12}	f_{11}	f_{10}
	$X = 0$	f_{0r}	$f_{0(r-1)}$	...	f_{02}	f_{01}	f_{00}

À partir de ces données, le rapport de cotes de Mantel-Haenszel mesurant l'association entre le facteur d'exposition *X* et la maladie *Y* est donné par

$$RC = \frac{\sum_{i=0}^{r-1}(r-i)f_{1i}}{\sum_{i=1}^{m} i f_{0i}} \tag{14.7}$$

L'estimation du *RC* par la méthode du maximum de vraisemblance n'a pas d'expression explicite. Rappelons d'abord que la fonction de vraisemblance a comme expression générale :

$$FV(\psi) = \prod_{i=0}^{r-1} C_{n_{i+1}}^{f_{1i}} \left(\frac{(i+1)\psi}{(i+1)\psi + (r-i)}\right)^{f_{1i}} \left(\frac{r-i}{(i+1)\psi + (r-i)}\right)^{f_{0i}} \tag{14.8}$$

L'équation fondamentale qui en découle,

$\frac{\partial L(\psi)}{\partial \psi} = \sum_{i=0}^{r-1}\left(\frac{f_{1i}}{\psi} - \frac{(i+1)n_{i+1}}{(i+1)\psi + (r-i)}\right) = 0$, ne peut être solutionnée que par des méthodes itératives.

14.3.1 TESTS ET INTERVALLES DE CONFIANCE EXACTS

Au plan théorique, il est facile de généraliser les calculs des tests et des intervalles de confiance exacts. Mais concrètement, ces calculs deviennent rapidement impraticables. Nous nous abstenons donc de les présenter.

Nous retenons uniquement quelques approches approximatives tant pour les tests statistiques que pour les intervalles de confiance.

14.3.2 TESTS EN APPROXIMATION NORMALE ET DU RAPPORT DE VRAISEMBLANCE

Test de McNemar

Le test de McNemar (ou du score) peut facilement être généralisé à une analyse appariée 1 à *r*. Rappelons que la variable *A* est alors la somme de *r* variables binomiales indépendantes F_{10}, F_{11}, …, $F_{1(r-1)}$. Sous l'hypothèse nulle, les paramètres de chacune de ces variables se présentent comme suit :

$$\text{Pour } F_{10}: \quad \pi_1 = \frac{1}{1+r} \text{ et } n_1 = f_{10} + f_{01}.$$

$$\text{Pour } F_{11}: \quad \pi_2 = \frac{2}{1+r} \text{ et } n_2 = f_{11} + f_{02}.$$

…

$$\text{Pour } F_{1(r\text{-}1)}: \quad \pi_2 = \frac{r}{1+r} \text{ et } n_r = f_{1(r\text{-}1)} + f_{0r}.$$

Ainsi, sous l'hypothèse nulle, on peut aisément déduire la valeur attendue $E(A)$ et la variance $V(A)$ de la variable *A* :

$$E(A) = \sum_{i=1}^{r} n_i \pi_i = \sum_{i=1}^{r} \frac{i n_i}{r+1} \quad \text{et} \quad V(A) = \sum_{i=1}^{r} n_i \pi_i (1-\pi_i) = \sum_{i=1}^{r} \frac{i(r+1-i) n_i}{(r+1)^2}$$

Le test en approximation normale apparaît donc comme :

$$z = \frac{a - E(A)}{\sqrt{V(A)}} \text{ ou encore } \chi_1^2 = \frac{(a - E(A))^2}{V(A)}.$$

En substituant les valeurs de $E(A)$ et $V(A)$ dans la dernière expression, on obtient le test de McNemar ou du score pour un appariement 1 à r :

$$\chi_1^2 = \frac{\left\{\sum_{i=0}^{r-1}\left[(r-i)f_{1i} - (i+1)f_{0(i+1)}\right]\right\}^2}{\sum_{i=0}^{r-1}(i+1)(r-i)n_{i+1}}$$

Test de Wald

Le test de Wald est analogue à celui décrit à la section 14.2.1 et se présente simplement comme :

$$\chi_1^2 = \frac{\left[\log(\psi)\right]^2}{V[\log(\psi)]}$$

où ψ est l'estimation du maximum de vraisemblance de *RC*.

Comme précédemment, nous nous en remettons à la procédure PHREG.

Test du rapport de vraisemblance

Le test du rapport de vraisemblance se généralise lui aussi assez facilement à l'appariement 1 à r. Les deux fonctions de vraisemblance à comparer, FV_0 correspondant à l'hypothèse nulle ($\psi = 1$) et FV_1 correspondant à l'hypothèse de l'estimateur du maximum de vraisemblance de ψ, se généralisent assez facilement. Si

$$FV_0 = \prod_{i=0}^{r-1} C_{n_{i+1}}^{f_{1i}} \left(\frac{i+1}{r+1}\right)^{f_{1i}} \left(\frac{r-i}{r+1}\right)^{f_{0(i+1)}}$$

$$FV_1 = \prod_{i=0}^{r-1} C_{n_{i+1}}^{f_{1i}} \left(\frac{(i+1)\psi}{(i+1)\psi + (r-i)}\right)^{f_{1i}} \left(\frac{r-i}{(i+1)\psi + (r-i)}\right)^{f_{0(i+1)}}$$

alors la statistique obéit à une loi du khi-carré avec 1 degré de liberté.

En substituant dans cette statistique les valeurs de FV_1 et FV_0, on obtient l'expression suivante pour le test du rapport de vraisemblance :

$$\chi_1^2 = 2\log\frac{FV_1}{FV_0}$$

$$= 2\sum_{i=0}^{r-1}\left[f_{1i}\log(\psi) + n_{i+1}\log\left(\frac{r+1}{(i+1)\psi + (r-i)}\right)\right]$$

Sans connaître l'estimateur du maximum de vraisemblance de ψ, le test ne peut pas être conduit. On peut recourir encore ici à la procédure PHREG de SAS, qui permet d'obtenir tant l'estimation du maximum de vraisemblance du *RC* que la valeur du test lui-même.

Exemple 14.6

Dans une étude, chacun des 20 cas d'une maladie rare Y a été apparié à quatre témoins. Les cas ont été comparés aux témoins pour le facteur X.

Les données peuvent se présenter dans un tableau comme celui-ci :

Tableau 14.14

		Nombre de témoins exposés				
		4	**3**	**2**	**1**	**0**
Cas	$X = 1$	0	1	5	4	3
	$X = 0$	0	0	0	3	4

Pour les différents tests décrits précédemment, nous présentons au tableau 14.15 les résultats obtenus par la procédure PHREG.

Tableau 14.15

Test	χ^2	p	Programme SAS
Du score	12,488	0,0004	Pr14.18
De Wald	9,819	0,0017	
RV	12,620	0,0004	

◆

14.3.3 Intervalles de confiance du *RC*

Pour le calcul de l'intervalle de confiance du *RC*, nous ne retenons que deux méthodes en approximation normale : 1) la méthode basée sur le résultat d'un test ; 2) la méthode de Wald. Les autres méthodes de calcul, exact ou par le rapport de vraisemblance, ne sont pas disponibles.

Méthode basée sur le résultat d'un test

Cette méthode a déjà été décrite précédemment. Il suffit d'appliquer au *RC* de Mantel-Haenszel le résultat du test du score pour obtenir :

$$\psi_{\text{inf}} = RC^{\left(1-\frac{z_{\alpha/2}}{\chi}\right)}$$

$$\psi_{\text{sup}} = RC^{\left(1+\frac{z_{\alpha/2}}{\chi}\right)}$$

où $\chi = \sqrt{\chi^2(\text{McNemar})}$.

Méthode de Wald

De par cette méthode, les limites de confiance sont simplement décrites comme :

$$\psi_{\text{inf}} = RC \times e^{-z_{\alpha/2}\sqrt{V(\log RC)}}$$

$$\psi_{\text{sup}} = RC \times e^{+z_{\alpha/2}\sqrt{V(\log RC)}}$$

où le *RC* et la variance V[log *RC*] sont les estimations du maximum de vraisemblance.

Exemple 14.7

Revenons aux données du tableau 14.14.

Dans le tableau 14.16, nous présentons les intervalles de confiance du *RC* à 95 %, suivant les deux méthodes retenues.

Tableau 14.16

Méthode	RC	Intervalle de confiance (95 %)	Programme SAS
Résultat d'un test	11,667	[2,987 ; 45,574]	Pr14.19
De Wald	8,168	[2,196 ; 30,383]	Pr14.20

On remarque ici une différence appréciable entre le RC de Mantel-Haenszel (*RC* = 11,667) et celui du maximum de vraisemblance (*RC* = 8,168). En conséquence, il en va de même pour les intervalles de confiance. La méthode basée sur le résultat d'un test est ici fortement influencée par la valeur du *RC*.

◆

CHAPITRE

15

ANALYSE APPARIÉE POUR LES ÉTUDES DE COHORTES

Dans ce chapitre, nous proposons des techniques d'analyse appariée dans le contexte d'études de cohortes, où l'on s'intéresse principalement à la mesure du risque relatif (*RR*) et secondairement à la différence des risques (*DR*). Les techniques développées pour le *RR* sont analogues à celles définies pour le rapport de cotes *RC* et présentées au chapitre 14. Les méthodes statistiques pour le *RR* s'avèrent cependant plus simples que celles pour le *RC*. En effet, dans le contexte des analyses appariées pour le *RR*, les données d'appariement uniforme peuvent être décrites simplement dans un tableau 2×2, quel que soit le niveau d'appariement. Notons qu'au plan formel, le *RR* et le *DR* se confondent respectivement avec un rapport et une différence de proportions.

Considérons une étude de cohorte qui porte sur l'association entre le facteur X dichotomique et une maladie Y. Chacun des K sujets exposés au facteur X ($X = 1$) a été apparié à r sujets non exposés ($X = 0$), pour un ou plusieurs facteurs à contrôler. Cet appariement définit K strates homogènes sur lesquelles on retrouve un exposé et un certain nombre de non-exposés. Par observation, chaque sujet est alors classé malade ($Y = 1$) ou non malade ($Y = 0$).

Sur les données de cette étude, on décide de pratiquer une analyse appariée pour vérifier l'association entre X et Y tout en contrôlant le ou les facteurs d'appariement. Ainsi, l'objectif de cette analyse pourrait être de mesurer par le RR l'association entre X et Y, d'établir un intervalle de confiance de cette mesure ou de tester cette association entre X et Y.

Notation : dans ce chapitre, RR représente en général le risque relatif mesuré sur les données observées et φ, le risque relatif théorique, paramétrique, dont la valeur est en général inconnue.

Pour l'analyse du RR, les K strates peuvent être fondues en un seul tableau de données comme le tableau 15.1.

TABLEAU 15.1

	Facteur d'exposition X		Total
	$X = 1$	$X = 0$	
$Y = 1$	a	b	m_1
$Y = 0$			
Total	K	rK	n

Les données d'un tel tableau peuvent être analysées suivant les outils déjà présentés au chapitre 6 pour les proportions dans un tableau 2×2. De telles analyses se fondent principalement sur la loi hypergéométrique.

Cependant, nous allons définir ici une autre approche d'analyse conditionnelle, analogue à celle présentée pour les analyses appariées dans les études de cas-témoins. Cette approche est intéressante pour trois raisons : 1) elle concorde avec celle des études cas-témoins ; 2) elle est centrée sur la mesure du RR ; 3) elle est directement accessible dans la procédure PHREG au moyen de la fonction de vraisemblance partielle de Breslow. Cependant, elle a le désavantage d'une perte de puissance liée au fait que les analyses portent uniquement sur les sujets ayant eu la maladie ($Y = 1$). Aussi, en guise de complément, pour le contexte d'appariement 1 à 1, nous suggérerons une approche plus efficiente basée sur l'estimation de la variance de $\log(RR)$.

L'approche conditionnelle étant peu ou pas connue, sa présentation requiert une introduction aux concepts de base de risque conditionnel et de vraisemblance partielle.

15.1 RISQUE CONDITIONNEL ET FONCTION DE VRAISEMBLANCE PARTIELLE

Considérons une strate particulière *i*, où 1 sujet exposé au facteur *X* est apparié à *r* non-exposés. Sur cette strate, on suppose pour le moment que le risque de *Y* est fonction du facteur *X* et qu'il n'y a qu'un seul sujet qui soit affecté par *Y*.

Alors, pour décrire ce risque $R_i(X)$, on utilise la relation $R_i(X) = R_{0i} \times \varphi^X$, où

- R_{0i} est un risque de base non précisé ;
- *X* décrit l'état d'exposition (*X* = 1 ou 0) du sujet affecté par *Y* ;
- φ correspond au risque relatif postulé constant à travers les strates.

Ainsi, pour la strate *i* où le sujet *u* est affecté par *Y*, on écrira $R_i(x_u) = R_{0i} \times \varphi^{x_u}$.

Afin d'éliminer le paramètre de nuisance R_{0i}, on définit sur les données le risque conditionnel $R_C(X)$. Pour la strate *i*, on écrira :

$$R_{C(i)}(x_u) = \frac{R_i(x_u)}{\sum_{j=1}^{r+1} R_i(x_{ij})} = \frac{\varphi^{x_u}}{\sum_{j=1}^{r+1} \varphi^{x_{ij}}}$$

où x_{ij} désigne l'état d'exposition du j^e sujet de la strate *i*.

Précisons immédiatement que le risque conditionnel n'est pas véritablement un risque. Il mesure plutôt la probabilité que le sujet atteint par *Y* soit un sujet ayant x_u comme valeur d'exposition (x_u =1 ou 0 dans le présent contexte). En conséquence, ce risque conditionnel ne peut être défini que sur les strates où il existe au moins 1 sujet atteint par *Y*.

Cette définition du risque conditionnel est associable à la vraisemblance partielle de Breslow dans la modélisation du hasard (risque) proportionnel en analyse de survie[1]. Dans la situation où il y a multiplicité des événements *Y* sur une même strate, l'approche de la vraisemblance partielle se présente comme suit :

1. Kalbfleisch, J.D. et R.L. Prentice, *The Statistical Analysis of Failure Data*, Toronto, Wiley, 1980.

pour d_i sujets ayant eu l'événement Y sur la strate i, le risque conditionnel R_{Ci} se définit comme

$$R_{Ci} = \prod_{u \in D_i} \left[\frac{\varphi^{x_u}}{\sum_{j=1}^{r+1} \varphi^{x_{ij}}} \right] = \frac{\varphi^{s_i}}{\left[\sum_{j=1}^{r+1} \varphi^{x_{ij}} \right]^{d_i}}$$ où u parcourt l'ensemble D_i

des sujets affectés par Y et $s_i = \sum_{u \in D_i} x_u$.

La fonction de vraisemblance partielle, regroupant l'information de l'ensemble des strates, prend la forme suivante :

$$FV(\varphi) = \prod_{i=1}^{k} R_{Ci} = \prod_{i=1}^{k} \frac{\varphi^{s_i}}{\left[\sum_{j=1}^{r+1} \varphi^{x_{ij}} \right]^{d_i}}$$

Cette fonction permettra d'estimer le paramètre φ.

15.2 ESTIMATION DU *RR*

Sur les données du tableau 15.1, le *RR* est simplement mesuré par :

$$RR = \frac{rK \times a}{K \times b} = \frac{ra}{b}$$

On remarque que, tenant compte du niveau m d'appariement, seules les cellules a et b comportent de l'information pertinente à la mesure du *RR*. Les analyses portant sur le *RR* peuvent alors être réduites aux seules strates sur lesquelles a été recensé au moins un événement ($Y = 1$). On peut montrer que ce *RR* est à la fois l'estimateur de Mantel-Haenszel et du maximum de vraisemblance.

15.2.1 CONTEXTE D'UN APPARIEMENT 1 À 1

Considérons le tableau 15.2 qui décrit schématiquement les données pour une analyse appariée en étude de cohorte. Chaque cellule du tableau décrit le nombre de paires d'une configuration donnée.

Tableau 15.2

		Sujet non exposé	
		$Y = 1$	$Y = 0$
Sujet exposé	$Y = 1$	f_{11}	f_{10}
	$Y = 0$	f_{01}	f_{00}

Estimation de Mantel-Haenszel

Le risque relatif est estimé par le rapport

$$RR = \frac{f_{11} + f_{10}}{f_{11} + f_{01}} = \frac{a}{b} \tag{15.1}$$

où a et b désignent les nombres de cas respectivement chez les sujets exposés et chez les sujets non exposés. Ce risque relatif est celui de Mantel-Haenszel. La démonstration est analogue à celle déjà faite pour le *RC* (section 14.1 du chapitre 14).

Rappelons (section 10.2.3 du chapitre 10) l'ajustement de Mantel-Haenszel pour le RR :

$$RR_{MH} = \frac{\sum_i \frac{a_{1i} n_{0i}}{n_i}}{\sum_i \frac{a_{0i} n_{1i}}{n_i}}$$

Les paires se regroupent suivant quatre types de strates que l'on peut décrire à l'aide du tableau 15.3 suivant :

Tableaux 15.3

Type I : f_{11}				**Type II : f_{10}**			
	$X = 1$	$X = 0$	Total		$X = 1$	$X = 0$	Total
$Y = 1$	1	1	2	$Y = 1$	1	0	1
$Y = 0$	0	0	0	$Y = 0$	0	1	1
Total	1	1	2	Total	1	1	2
Type III : f_{01}				**Type IV : f_{00}**			
	$X = 1$	$X = 0$	Total		$X = 1$	$X = 0$	Total
$Y = 1$	0	1	1	$Y = 1$	0	0	0
$Y = 0$	1	0	1	$Y = 0$	1	1	2
Total	1	1	2	Total	1	1	2

Pour chacun de ces tableaux, $n_{0i} = n_{1i} = 1$ et $n_i = 2$ (le total).

En appliquant la formule d'ajustement de Mantel-Haneszel, on obtient précisément :

$$RR_{MH} = \frac{a}{b}.$$

Estimateur du maximum de vraisemblance

Le risque relatif $RR = \frac{a}{b}$ est aussi l'estimateur du maximum de vraisemblance dans le cadre de la vraisemblance partielle.

Le risque conditionnel est défini pour tous les types de paires sauf pour le type IV. Les paires de type IV ne comportent aucune information pertinente à l'estimation du risque conditionnel et, conséquemment, à celle du *RR*.

Pour chacun des trois types de paires utilisées, nous estimons le risque conditionnel.

- pour une paire de type I, le risque conditionnel est de

  $$RC_{C1} = \frac{\varphi}{(1+\varphi)^2}$$

- pour une paire de type II, le risque conditionnel est de

  $$RC_{C2} = \frac{\varphi}{(1+\varphi)}$$

- pour une paire de type III, le risque conditionnel est de

  $$RC_{C3} = \frac{1}{(1+\varphi)}$$

En tenant compte de la contribution de chacune des cellules du tableau, on peut donc construire la fonction de vraisemblance partielle :

$$FV(\varphi) = \left(\frac{\varphi}{(1+\varphi)^2}\right)^{f_{11}} \left(\frac{\varphi}{(1+\varphi)}\right)^{f_{10}} \left(\frac{1}{(1+\varphi)}\right)^{f_{01}} = \frac{\varphi^{(f_{11}+f_{10})}}{(1+\varphi)^{2f_{11}+f_{10}+f_{01}}} \quad (15.2)$$

En annulant la dérivée de $\log[FV(\varphi)]$ et en solutionnant pour φ, on retrouve $\varphi = \frac{f_{11}+f_{10}}{f_{11}+f_{01}} = \frac{a}{b}$, qui est l'estimateur présumé.

15.2.2 CONTEXTE D'APPARIEMENT 1 À 2

Considérons le tableau suivant décrivant n triplets.

TABLEAU 15.4

		Nombre de cas chez les sujets non exposés		
		2	**1**	**0**
Sujet exposé	Y = 1	f_{12}	f_{11}	f_{10}
	Y = 0	f_{02}	f_{01}	f_{00}

Les triplets se regroupent suivant six types de strates que l'on peut décrire à l'aide du tableau 15.5.

TABLEAU 15.5

Type I : f_{12}				**Type II : f_{11}**			
	$X = 1$	$X = 0$	**Total**		$X = 1$	$X = 0$	**Total**
$Y = 1$	1	2	3	$Y = 1$	1	1	2
$Y = 0$	0	0	0	$Y = 0$	0	1	1
Total	1	2	3	Total	1	2	3
Type III : f_{10}				**Type IV : f_{02}**			
	$X = 1$	$X = 0$	**Total**		$X = 1$	$X = 0$	**Total**
$Y = 1$	1	0	1	$Y = 1$	0	2	2
$Y = 0$	0	2	2	$Y = 0$	1	0	1
Total	1	2	3	Total	1	2	3
Type V : f_{01}				**Type VI : f_{00}**			
	$X = 1$	$X = 0$	**Total**		$X = 1$	$X = 0$	**Total**
$Y = 1$	0	1	1	$Y = 1$	0	0	0
$Y = 0$	1	1	2	$Y = 0$	1	2	3
Total	1	2	3	Total	1	2	3

Le risque conditionnel est défini pour tous les types de triplets sauf pour le type VI. Les triplets de type VI ne comportent aucune information pertinente à l'estimation du risque conditionnel et, conséquemment, à celle du *RR*.

Estimation de Mantel-Haenszel du *RR*

En appliquant la formule d'ajustement de Mantel-Haneszel, on obtient précisément :

$$RR_{MH} = \frac{2a}{b}$$

Estimation du maximum de vraisemblance du *RR*

Pour chacun des cinq types de paires utilisés, nous estimons le risque conditionnel :

- pour une paire de type I, le risque conditionnel est de

$$RC_{C1} = \frac{\varphi}{(2+\varphi)^3}$$

- pour une paire de type II, le risque conditionnel est de

$$RC_{C2} = \frac{\varphi}{(2+\varphi)^2}$$

- pour une paire de type III, le risque conditionnel est de

$$RC_{C3} = \frac{\varphi}{(2+\varphi)}$$

- pour une paire de type IV, le risque conditionnel est de

$$RC_{C4} = \frac{1}{(2+\varphi)^2}$$

- pour une paire de type V, le risque conditionnel est de

$$RC_{C5} = \frac{1}{(2+\varphi)}$$

La fonction de vraisemblance prend alors la forme :

$$\begin{aligned} FV(\varphi) &= \left(\frac{\varphi}{(2+\varphi)^3}\right)^{f_{12}} \left(\frac{\varphi}{(2+\varphi)^2}\right)^{f_{11}} \left(\frac{\varphi}{(2+\varphi)}\right)^{f_{10}} \\ &\quad \left(\frac{1}{(2+\varphi)^2}\right)^{f_{02}} \left(\frac{1}{(2+\varphi)}\right)^{f_{01}} \\ &= \frac{\varphi^{(f_{12}+f_{11}+f_{10})}}{(2+\varphi)^{3f_{12}+2(f_{11}+f_{02})+(f_{10}+f_{01})}} \end{aligned} \qquad (15.3)$$

En annulant la dérivée de log[$FV(\varphi)$] et en solutionnant pour φ, on retrouve $\varphi = \dfrac{2(f_{12} + f_{11} + f_{10})}{2(f_{12} + f_{02}) + (f_{10} + f_{01})} = \dfrac{2a}{b}$, qui est l'estimateur présumé.

Pour un appariement quelconque 1 à r, les deux estimateurs conduisent à $RR = \dfrac{ra}{b}$.

15.3 TESTS STATISTIQUES

Les tests statistiques qui sont présentés ici sont ceux rattachés au *RR*. L'hypothèse testée est celle d'un *RR* = 1 (hypothèse nulle). La définition des tests n'implique que les cellules *a* et *b* regroupant toutes les strates sur lesquelles a été recensé au moins un événement ($Y = 1$). Nous présentons d'abord le test exact défini dans le cadre de la loi binomiale, puis deux tests en approximation normale (le test du score et le test de Wald) et, enfin, le test du rapport de vraisemblance.

15.3.1 TEST EXACT BINOMIAL

Considérons la variable *A*, correspondant au nombre de cas exposés au facteur *X*. La valeur observée de *A* est *a*. Rapportée à l'ensemble des m_1 cas, cette variable *A* peut être considérée comme une variable binomiale de paramètre $\pi = \dfrac{\varphi}{r + \varphi}$ et $n = m_1$, où φ désigne un *RR* quelconque.

Sous l'hypothèse nulle, $\varphi = 1$ et $\pi = \dfrac{1}{1+r}$. Concrètement, sous l'hypothèse nulle et dans le cadre d'un appariement 1 à *r*, il est raisonnable de penser qu'un cas sur $(1 + r)$ appartiendra au groupe des exposés. Sous cette hypothèse, la valeur-*p* unilatérale à droite est donc tout simplement calculée comme :

$$p = 1/2\ C_n^a \left(\frac{1}{1+r}\right)^a \left(\frac{r}{1+r}\right)^{n-a} + \sum_{x=a+1}^{n} C_n^x \left(\frac{1}{1+r}\right)^x \left(\frac{r}{1+r}\right)^{n-x}$$

On peut définir de façon analogue le test unilatéral à gauche.

Le test exact bilatéral se définit par convention comme :

$$p = \left(1/2\right)^{1-\delta} C_n^a \left(\frac{1}{1+r}\right)^a \left(\frac{r}{1+r}\right)^{n-a} + \sum\nolimits_u C_n^u \left(\frac{1}{1+r}\right)^u \left(\frac{r}{1+r}\right)^{n-u}$$

S'il existe une valeur *u* également extrême à *a* (dans ce cas, $\delta = 1$; autrement, $\delta = 0$), sa probabilité est traitée de même façon que celle de *a* avant d'être cumulée dans la valeur-*p*. Aussi, la sommation est faite sur toutes les valeurs *u* plus extrêmes que *a*.

On peut définir un test binomial pour une hypothèse quelconque sur le *RR*. Sous l'hypothèse $RR = \varphi$, la variable *A* obéit à une variable binomiale de paramètres $\pi = \dfrac{\varphi}{\varphi + r}$ et *n*. Le test se présente alors comme :

$$p = \left(1/2\right)^{1-\delta} C_n^a \left(\frac{\varphi}{\varphi + r}\right)^a \left(\frac{r}{\varphi + r}\right)^{n-a} + \sum\nolimits_u C_n^u \left(\frac{\varphi}{\varphi + r}\right)^u \left(\frac{r}{\varphi + r}\right)^{n-u}$$

15.3.2 TESTS EN APPROXIMATION NORMALE

Test du score

L'approximation normale sera tout simplement appliquée à la variable binomiale *A* considérée sous l'hypothèse nulle, $\pi = 1/2$. Rappelons que la variable *A* représente le nombre de cas exposés parmi les *n* cas observés : $A = F_{11} + F_{10}$ et $n = 2f_{11} + f_{10} + f_{01}$. La valeur observée de *A* est $a = f_{11} + f_{10}$. Le test du score se définit alors comme :

$$\chi_1^2 = \frac{\left[a - E(A)\right]^2}{V(A)} = \frac{\left[ra - b\right]^2}{ra + b} = \frac{b\left[RR - 1\right]^2}{RR + 1} \qquad (15.4)$$

Il est intéressant de comparer cette expression (15.4) à l'expression (14.3) du chapitre 14, définissant le test de McNemar pour le rapport de cotes.

Test de Wald

Comme pour l'analyse appariée portant sur le *RC*, le test de Wald se présente simplement comme :

$$\chi_1^2 = \frac{\left[\log(RR)\right]^2}{V[\log(RR)]}$$

où φ est estimé par la méthode du maximum de vraisemblance.

On peut facilement déduire la variance de log(*RR*) par la méthode delta. Cette variance prend la forme suivante :

$$V[\log(RR)] = \frac{a+b}{ab}$$

Le calcul du test en découle donc directement.

Nous pouvons nous en remettre à la procédure PHREG qui fournit non seulement une estimation de φ et de sa variance, mais aussi la valeur du test de Wald lui-même.

15.3.3 TEST DU RAPPORT DE VRAISEMBLANCE

La fonction de vraisemblance pour l'appariement 1 à *r* a la forme suivante :

$$FV(\varphi) = \frac{\varphi^a}{(r+\varphi)^{(a+b)}}$$

Le test du rapport de vraisemblance se construit en comparant la fonction de vraisemblance FV_1 évaluée sous l'hypothèse la plus vraisemblable $\varphi = \frac{ra}{b}$, et FV_0 évaluée sous l'hypothèse nulle (φ = 1). Le test est alors de la forme

$$\chi_1^2 = 2\log\left(\frac{FV_1}{FV_0}\right)$$
$$= 2\left\{a\log(a) + b\log\left(\frac{b}{r}\right) - (a+b)\log\left(\frac{a+b}{r+1}\right)\right\}$$

EXEMPLE 15.1

APPARIEMENT 1 À 1

Considérons les données du tableau 15.6. On suppose alors qu'elles sont issues d'un appariement pratiqué dans une étude de cohorte où chaque sujet exposé à *X* (*X* = 1) a été assorti d'un sujet non exposé (*X* = 0).

Tableau 15.6

		Non-exposé		Total
		$Y = 1$	$Y = 0$	
Exposé	$Y = 1$	4	18	22
	$Y = 0$	6	8	14
	Total	10	26	36

Ces données peuvent aussi être présentées dans un tableau croisé X par Y après la fonte des strates en un seul ensemble (tableau 15.7) :

Tableau 15.7

	Exposé		Total
	$X = 1$	$X = 0$	
$Y = 1$	22	10	32
$Y = 0$	14	26	40
Total	36	36	72

Mesure du RR : $RR = \dfrac{22}{10} = 2,2$

Test exact

L'hypothèse est $\pi = 0,50$. De plus, $n = 32$ et $a = 22$ (valeur observée). La valeur-p unilatérale à droite est donnée par :

$$\begin{aligned} p &= P(A \geq 22 \mid \pi = 0,50) \\ &= 1/2\, C_{32}^{22}(0,50)^{22}(1-0,50)^{10} + \sum_{i=23}^{32} C_{32}^{i}(0,51)^{i}(1-0,51)^{32-i} \\ &= 0,01754 \end{aligned}$$

La valeur-p du test bilatéral est de 0,035082. On conclut donc que ces données sont non compatibles avec l'hypothèse nulle.

Approximation normale

Test du score

$$\chi_1^2 = \frac{10\left[2,2-1\right]^2}{2,2+1} = 4,5$$

La valeur-p correspondante est de 0,0339.

Test de Wald

$$\chi_1^2 = \frac{\left[\log(2,2)\right]^2}{\dfrac{22+10}{22\times 10}} = 4,2739$$

La valeur-p correspondante est de 0,0387.

Test du rapport de vraisemblance

$$\chi_1^2 = 2\left[22\log(22) + 10\log(10) - 32\log\left(\frac{32}{2}\right)\right] = 4,6119$$

La valeur-*p* correspondante est de 0,0318.

Dans le tableau 15.8, on rappelle les résultats des différents tests.

Tableau 15.8

Test statistique	Khi-carré	Valeur-*p* bilatérale	Programme
Binomial	–	0,0351	PR15.1
Du score	4,5000	0,0339	PR15.2
De Wald	4,2739	0,0387	
Du RV	4,6119	0,0318	

◆

Exemple 15.2

Appariement 1 à 2

Considérons les données du tableau 15.9. On suppose alors qu'elles sont issues d'un appariement pratiqué dans une étude de cohorte où chaque sujet exposé à *X* a été apparié à 2 sujets non exposés à *X*.

Tableau 15.9

		Nombre de non-exposés malades		
		2	1	0
Exposé	$Y = 1$	4	18	16
	$Y = 0$	6	8	2

Fondues dans un seul tableau croisé *X* par *Y*, les strates génèrent le tableau□15.10.

Tableau 15.10

	Exposé		Total
	$X = 1$	$X = 0$	
$Y = 1$	38	46	84
$Y = 0$	16	62	68
	54	2×54	162

Mesure du RR : $RR = \dfrac{2 \times 38}{46} = 1,6522$

Test exact

L'hypothèse nulle est $\pi = ⅓$. De plus, $n = 84$ et $a = 38$ (valeur observée). La valeur-p unilatérale à droite est donnée par :

$$\begin{aligned} p &= P(A \geq 38 \mid \pi = 1/3) \\ &= 1/2\, C_{84}^{38}(0,33)^{38}(1-0,33)^{46} + \sum_{i=39}^{84} C_{84}^{i}(0,33)^{i}(1-0,33)^{84-i} \\ &= 0,0120 \end{aligned}$$

La valeur-p du test bilatéral est de 0,0238. On conclut donc que ces données sont non compatibles avec l'hypothèse nulle.

Test du score

$$\chi_1^2 = \frac{46\left[1,6522 - 1\right]^2}{1,6522 + 2} = 5,3571$$

La valeur-p correspondante est de 0,0206.

Test de Wald

$$\chi_1^2 = \frac{\left[\log(1,6522)\right]^2}{\dfrac{38+46}{38 \times 46}} = 5,2460$$

La valeur-p correspondante est de 0,0220.

Test du rapport de vraisemblance

$$\chi_1^2 = 2\left[38\log(38) + 46\log\left(\frac{46}{2}\right) - 84\log\left(\frac{84}{2+1}\right)\right] = 5,1117$$

La valeur-p correspondante est de 0,0238.

Dans le tableau 15.11, on rappelle les résultats des différents tests.

Tableau 15.11

Test statistique	Khi-carré	Valeur-p bilatérale	Programme
Binomial	–	0,0238	PR15.3
Du score	5,3571	0,0206	PR15.4
De Wald	5,2470	0,0220	
Du RV	5,1117	0,0238	

◆

Exemple 15.3

Appariement 1 à 4

Considérons les données du tableau 15.12. On suppose alors qu'elles sont issues d'un appariement pratiqué dans une étude de cohorte où chacun des 20 sujets exposés à X a été apparié à 4 sujets non exposés.

Tableau 15.12

	Nombre de non-exposés malades				
	4	**3**	**2**	**1**	**0**
Exposé $Y = 1$	0	1	5	4	3
$Y = 0$	0	0	0	3	4

Fondues dans un seul tableau croisé X par Y, les strates génèrent le tableau 15.13.

Tableau 15.13

	Exposé		Total
	$X = 1$	$X = 0$	
$Y = 1$	13	20	33
$Y = 0$	7	60	67
Total	20	4×20	100

Mesure du RR : $RR = \dfrac{4 \times 13}{20} = 2{,}6$

Test exact

L'hypothèse nulle est $\pi = ⅓$. De plus, $n = 33$ et $a = 13$ (valeur observée). La valeur-p unilatérale à droite est donnée par :

$$\begin{aligned} p &= P(A \geq 13 \mid \pi = 1/5) \\ &= 1/2\, C_{33}^{13}(0,20)^{13}(1-0,20)^{20} + \sum_{i=14}^{33} C_{33}^{i}(0,20)^{i}(1-0,20)^{33-i} \\ &= 0{,}0055 \end{aligned}$$

La valeur-p du test bilatéral est de 0,0113. On conclut que les données sont non compatibles avec l'hypothèse nulle.

Test du score

$$\chi_1^2 = \frac{20\left[2{,}6-1\right]^2}{2{,}6+4} = 7{,}7576$$

La valeur-p correspondante est de 0,0053.

TEST DE WALD

$$\chi_1^2 = \frac{[\log(2,6)]^2}{\frac{13+20}{13\times 20}} = 7,1934$$

La valeur-*p* correspondante est de 0,0073.

TEST DU RAPPORT DE VRAISEMBLANCE

$$\chi_1^2 = 2\left[13\log(13) + 20\log\left(\frac{20}{4}\right) - 33\log\left(\frac{33}{4+1}\right)\right] = 6,5196$$

La valeur-*p* correspondante est de 0,0107.

Dans le tableau 15.14, on rappelle les résultats des différents tests.

TABLEAU 15.14

Test statistique	Khi-carré	Valeur-*p* bilatérale	Programme
Binomial	–	0,0113	PR15.5
Du score	7,7576	0,0053	PR15.6
De Wald	7,1934	0,0073	
Du RV	6,5196	0,0107	

◆

15.4 INTERVALLES DE CONFIANCE DU *RR* EN ANALYSE APPARIÉE

Pour le calcul de l'intervalle de confiance du *RR* en analyse appariée, nous retenons la méthode exacte, deux méthodes en approximation normale, soit celle basée sur le résultat d'un test et celle de Wald, et la méthode du rapport de vraisemblance.

INTERVALLE DE CONFIANCE EXACT

L'intervalle de confiance exact de niveau $100(1 - \alpha)$ % se calcule dans le cadre où la variable *A* est considérée comme une variable binomiale de paramètre $\pi = \frac{\varphi}{\varphi + r}$ et $n = m_1$. Les limites de confiance de φ sont alors définies comme suit :

φ_{inf} est la solution de l'équation binomiale : $P(A \geq a \mid \varphi) = \frac{\alpha}{2}$

φ_{sup} est la solution de l'équation binomiale : $P(A \leq a \mid \varphi) = \dfrac{\alpha}{2}$

Pour le calcul de ces limites, nous utilisons l'approche « mi-*p* ».

Intervalle de confiance en approximation normale

Tant pour la méthode basée sur le résultat d'un test que pour celle de Wald, les procédures de calcul sont analogues à celles déjà présentées à la section 14.3.2 du chapitre 14. Les limites de confiance obtenues par la méthode de Wald sont directement accessibles dans la procédure PHREG, avec l'option TIES=BRESLOW (option par défaut).

Intervalle de confiance par la méthode du rapport de vraisemblance

Pour déterminer l'intervalle de confiance du *RR* par la méthode du rapport de vraisemblance, il suffit de déterminer les deux valeurs de φ qui satisfont l'équation logarithmique :

$$2\left[L - L(\varphi)\right] - \chi^2_{1,1-\alpha} = 0$$

Ces valeurs, correspondant aux limites de confiance recherchées, sont déterminées par itération. La fonction de vraisemblance partielle utilisée ici est de la forme :

$$L(\varphi) = a\log(\varphi) - (a+b)\log(\varphi + r) + K$$

où K est une valeur indépendante de φ.

Cette fonction de vraisemblance atteint son maximum lorsque φ est l'estimation du maximum de vraisemblance :

$$L = a\log(\varphi_M) - (a+b)\log(\varphi_M + r) + K$$

où φ_M est l'estimateur du maximum de vraisemblance.

Exemple 15.4

Revenons aux données du tableau 15.13. Le risque relatif est estimé à 2,6 et la valeur du khi-carré du score est de 7,7576.

Dans le tableau 15.15, nous décrivons les intervalles de confiance obtenus par les différentes méthodes présentées.

TABLEAU 15.15

Méthode	*RR*	Intervalle de confiance (95 %)	Programme SAS
Exacte	–	[1,260 ; 5,222]	PR15.7
De Wald	2,600	[1,293 ; 5,227]	PR15.8
Résultat d'un test	2,600	[1,327 ; 5,093]	PR15.9
RV*	2,600	[1,262 ; 5,174]	PR15.10

* La procédure PHREG ne permet pas d'obtenir les limites de confiance par la méthode du rapport de vraisemblance.

On remarque ici une très bonne concordance entre la méthode exacte et celle du RV.

15.5 APPROCHE NON CONDITIONNELLE POUR LE *RR* ET LE *DR*

L'approche non conditionnelle est basée sur la distribution multinomiale que constituent les fréquences (F_{1r}, …, F_{10}, F_{0r}, …, F_{00}). Par exemple, les données du tableau 15.2 pour un appariement 1 à 1 donnent la distribution multinomiale (F_{11}, F_{10}, F_{01}, F_{00}), où $F_{11} + F_{10} + F_{01} + F_{00} = n$. Cette approche s'avère plus puissante que l'approche conditionnelle définie précédemment. Nous ne la décrivons ici que dans le cadre de l'approximation normale.

15.5.1 APPARIEMENT 1 À 1

Les données sont disposées suivant le schéma du tableau 15.2, que nous rappelons ici :

TABLEAU 15.16

		Sujet non exposé	
		$Y = 1$	$Y = 0$
Sujet exposé	$Y = 1$	f_{11}	f_{10}
	$Y = 0$	f_{01}	f_{00}

Sur ce tableau de n paires, les variables (ou cellules) F_{11}, F_{10}, F_{01} et F_{00} constituent une distribution multinomiale. Les risques R_1 chez les sujets exposés et R_0 chez les sujets non exposés sont respectivement donnés par :

$$R_1 = \frac{f_{11} + f_{10}}{n} \text{ et } R_0 = \frac{f_{11} + f_{01}}{n}.$$

Le rapport *RR* des risques correspond à $RR = \dfrac{f_{11} + f_{10}}{f_{11} + f_{01}}$ et la différence *DR* des risques à $DR = R_1 - R_0 = \dfrac{f_{10} - f_{01}}{n}$.

Le rapport *RR* des risques

Pour définir aussi bien le test statistique (de Wald) que l'intervalle de confiance du *RR*, il suffit d'établir la variance de log(*RR*). En appliquant la méthode delta, on la déduit facilement :

$$V\left[\log(RR)\right] = V\left[\log(F_{11} + F_{10})\right] + V\left[\log(F_{11} + F_{01})\right] + 2\,\mathrm{cov}\left[\log(F_{11} + F_{10}), -\log(F_{11} + F_{01})\right] \tag{15.5}$$

Détaillons ici chacun des éléments de l'expression de droite.

$$\begin{aligned} V\left[\log(F_{11} + F_{10})\right] &= \left(\frac{1}{f_{11} + f_{10}}\right)^2 V(F_{11} + F_{10}) \\ &= \left(\frac{1}{f_{11} + f_{10}}\right)^2 \begin{bmatrix} V(F_{11}) + V(F_{10}) \\ + 2\,\mathrm{cov}(F_{11}, F_{10}) \end{bmatrix} \\ &= \left(\frac{1}{f_{11} + f_{10}}\right)^2 \begin{bmatrix} \dfrac{f_{11}(n - f_{11})}{n} + \\ \dfrac{f_{10}(n - f_{10})}{n} - 2 \times \dfrac{f_{11} f_{10}}{n} \end{bmatrix} \\ &= \frac{(f_{01} + f_{00})}{n(f_{11} + f_{10})} \end{aligned} \tag{15.6}$$

Par symétrie on déduit que :

$$\begin{aligned} V\left[\log(F_{11} + F_{01})\right] &= \left(\frac{1}{f_{11} + f_{01}}\right)^2 V(F_{11} + F_{01}) \\ &= \frac{(f_{10} + f_{00})}{n(f_{11} + f_{01})} \end{aligned} \tag{15.7}$$

Enfin,

$$2\,\mathrm{cov}\begin{bmatrix}\log(F_{11}+F_{10}),\\ -\log(F_{11}+F_{01})\end{bmatrix} = -2\left(\frac{1}{f_{11}+f_{10}}\right)\left(\frac{1}{f_{11}+f_{10}}\right)$$
$$\mathrm{cov}\left[F_{11}+F_{10}, F_{11}+F_{01}\right]$$
$$= -\frac{2(f_{11}f_{00}-f_{10}f_{01})}{n(f_{11}+f_{10})(f_{11}+f_{01})} \qquad (15.8)$$

En substituant les expressions (15.6), (15.7) et (15.8) dans l'expression (15.5), on obtient l'expression de la variance de $\log(RR)$:

$$V\left[\log(RR)\right] = \frac{\left(f_{10}+f_{01}\right)}{\left(f_{11}+f_{10}\right)\left(f_{11}+f_{01}\right)} \qquad (15.9)$$

Test statistique de Wald non conditionnel sur le RR en analyse appariée

Le test statistique de Wald découle assez directement de l'expression de la variance de $\log(RR)$: $\chi_1^2 = \dfrac{\left[\log(RR)\right]^2}{V\left[\log(RR)\right]}$.

Intervalle de confiance non conditionnel du RR en analyse appariée

L'intervalle de confiance en approximation normale découle aussi directement de l'expression (15.9). On a : $IC : RR \times e^{\pm z_{\alpha/2}\sqrt{V[\log(RR)]}}$, expression maintes fois rencontrée.

Ainsi : $\varphi_{\text{inf}} = RR \times e^{-z_{\alpha/2}\sqrt{V(\log RR)}}$

$\varphi_{\text{sup}} = RR \times e^{+z_{\alpha/2}\sqrt{V(\log RR)}}$

Exemple 15.5

Appliquons le test statistique de Wald aux données du tableau 15.6.

On a :

$$\chi_1^2 = \frac{\left[\log(RR)\right]^2}{V\left[\log(RR)\right]}$$
$$= \frac{\left[\log 2,2\right]^2}{\dfrac{(18+6)}{(4+18)(4+6)}}$$
$$= 5,70$$

En comparant ce résultat à ceux obtenus à l'exemple 15.1, on remarque que ce test-ci est sensiblement plus puissant que ceux définis à la section 15.3.2.

L'intervalle de confiance à 95 % du *RR* est donné par :

$$\begin{aligned}\varphi_{\text{inf}} &= 2{,}2 \times e^{-1{,}96\sqrt{\frac{24}{22\times 10}}} \\ &= 1{,}15\end{aligned}$$

$$\begin{aligned}\varphi_{\text{sup}} &= 2{,}2 \times e^{+1{,}96\sqrt{\frac{24}{22\times 10}}} \\ &= 4{,}20\end{aligned}$$

(**PR15.11**)

◆

La différence DR des risques

En utilisant la méthode delta, on peut facilement estimer la variance du *DR* :

$$\begin{aligned}V\left(DR\right) &= V\left(\frac{F_{10}-F_{01}}{n}\right) \\ &= \frac{1}{n^2}V\left(F_{10}-F_{01}\right) \\ &= V(F_{10})+V(F_{01})+2\operatorname{cov}\left(F_{10},-F_{01}\right)\end{aligned} \tag{15.10}$$

Détaillons ici chacun des éléments de l'expression de droite.

$$V(F_{10}) = \frac{f_{10}(n-f_{10})}{n} \tag{15.11}$$

$$V(F_{01}) = \frac{f_{01}(n-f_{01})}{n} \tag{15.12}$$

et

$$\begin{aligned}2\operatorname{cov}\left(F_{10},-F_{01}\right) &= -2\operatorname{cov}\left(F_{10},F_{01}\right) \\ &= 2\frac{f_{10}f_{01}}{n}\end{aligned} \tag{15.13}$$

En substituant les expressions (15.11), (15.12) et (15.13) dans l'expression (15.10), on obtient l'expression de la variance de *DR* :

$$V(DR) = \frac{1}{n^3}\times\left[n(f_{10}+f_{01})-(f_{10}-f_{01})^2\right] \tag{15.14}$$

Test statistique de Wald non conditionnel sur le DR en analyse appariée

Le test statistique de Wald découle assez directement de l'expression de la variance de $\log(RR)$: $\chi_1^2 = \dfrac{DR^2}{V(DR)}$.

Intervalle de confiance non conditionnel du DR en analyse appariée

L'intervalle de confiance en approximation normale découle aussi directement de l'expression (15.14). On a : $IC : DR \pm z_{\alpha/2}\sqrt{V(DR)}$, expression maintes fois rencontrée.

$$\text{Ainsi : } \Delta_{\text{inf}} = DR - z_{\alpha/2}\sqrt{V(DR)}$$
$$\Delta_{\text{sup}} = DR + z_{\alpha/2}\sqrt{V(DR)}$$

Exemple 15.6

Appliquons le test statistique de Wald aux données du tableau 15.6.

On a :

$$\chi_1^2 = \frac{(DR)^2}{V(DR)}$$
$$= \frac{[12/36]^2}{\frac{1}{36^3}\times\left[36(18+6)-(18-6)^2\right]}$$
$$= \frac{[12/36]^2}{0{,}01543}$$
$$= 7{,}20$$

La valeur de ce test est sensiblement supérieure à celle du test appliqué sur les mêmes données, à l'exemple 15.5.

Les limites de confiance de l'intervalle de confiance à 95 % du *DR* sont données par :

$$\Delta_{\text{inf}} = \frac{12}{36} - 1{,}96\sqrt{0{,}01543}$$
$$= 0{,}09$$
$$\Delta_{\text{sup}} = \frac{12}{36} + 1{,}96\sqrt{0{,}01543}$$
$$= 0{,}58$$

(**Pr15.12**)

◆

15.5.2 APPARIEMENT 1 À *R*

On suppose que les *n* sujets exposés au facteur *X* sont appariés chacun à *r* sujets non exposés. Les données peuvent être disposées suivant le schéma du tableau 15.1 ou suivant le schéma d'analyse stratifiée du tableau 15.17.

TABLEAU 15.17

		Nombre de cas chez les sujets non exposés				
		r	(*r* – 1)	...	1	0
Sujet exposé	Y = 1	f_{1r}	$f_{1(-1)}$	...	f_{11}	f_{10}
	Y = 0	f_{0r}	$f_{0(r-1)}$	...	f_{01}	f_{00}

On rappelle alors que *RR* peut être décrit comme $RR = \frac{ra}{b}$ suivant les notations du tableau 15.1 ou encore comme $RR = \frac{r\sum_{i=0}^{r} f_{1i}}{\sum_{i=0}^{r} i(f_{1i} + f_{0i})}$ suivant celles du tableau 15.17.

De même, suivant les mêmes notations, la différence *DR* des risques prend respectivement les formes $DR = \frac{ra - b}{rK}$ et $DR = \frac{\sum_{i=1}^{r} \left[if_{1(r-i)} - (r-i+1) f_{0(r-i+1)} \right]}{r \times n}$.

Sans connaître de façon explicite les variances de *RR* et de *DR*, on peut toutefois calculer l'intervalle de confiance de ces mesures en utilisant la procédure GENMOD de SAS. Dans l'exemple suivant, nous présentons les résultats d'analyse obtenues par GENMOD sur les données du tableau 15.12.

EXEMPLE 15.7

Nous rappelons les données du tableau 15.12, qui décrit le cas d'un appariement 1 à 4.

TABLEAU 15.18

		Nombre de non-exposés malades				
		4	**3**	**2**	**1**	**0**
Exposé	$Y=1$	0	1	5	4	3
	$Y=0$	0	0	0	3	4

Le rapport *RR* des risques est calculé comme :

$$RR = \frac{4\times(3+4+5+1+0)}{0\times(3+4)+1\times(4+3)+2\times(5+0)+3\times(1+0)+4\times(0+0)} = 2,6$$

La différence *DR* des risques est calculée comme :

$$DR = \frac{(1\times 1-4\times 0)+(2\times 5-3\times 0)+(3\times 4-2\times 0)+(4\times 3-1\times 3)}{4\times 20} = 0,4$$

Par la procédure GENMOD, on obtient $RR = e^{0,9389} = 2,56$ avec l'intervalle de confiance à 95 % défini par les limites $RR_{inf} = e^{0,5514} = 1,74$ et $RR_{inf} = e^{1,3264} = 3,77$.

La différence des risques correspond à $DR = 0,3734$ avec un intervalle de confiance à 95 % dont les limites correspondent à $DR_{inf} = 0,1962$ et $DR_{sup} = 0,5506$.

PR15.13

Pour chacune des mesures, on remarque une légère différence entre la valeur estimée sur les données et celle fournie par la procédure. La procédure permet de tenir compte de la corrélation entre les observations appariées, mais au prix d'une légère modification dans l'estimation de la mesure.

◆

RÉFÉRENCES BIBLIOGRAPHIQUES (Principaux ouvrages à consulter)

Agresti, A., *Categorical data analysis*, New York, John Wiley & Sons, 1990.

Ahlbom, Anders, *Biostatistics for Epidemiologists,* Ann Arbor, Lewis Publishers, 1993.

Ancelle, T., *Statistique Épidémiologie,* Paris, Éditions Maloine, 2002.

Bernard, P.M. et C. Lapointe, *Mesures statistiques en épidémiologie,* Sainte-Foy, Presses de l'Université du Québec, 1991.

Bouyer, J., D. Hémon, S. Cordier *et al., Épidémiologie. Principes et méthodes quantitatives,* Paris, Éditions INSERM, 1995.

Breslow, N.E. et N.E. Day, *Statistical Methods in Cancer Research. Volume I – The Analysis of Case-Control Studies,* Lyon, IARC Scientific Publications No 32, 1980.

Breslow, N.E. et N.E. Day, *Statistical Methods in Cancer Research. Volume II – The Design and Analysis of Cohort Studies*, Lyon, IARC Scientific Publications No 82, 1987.

Cox, D.R. et E.J. Snell, *Analysis of Binary Data,* New York, Chapman and Hall, 1989.

Fleiss, J.L., *Statistical Methods for Rates and Proportions,* New York, John Wiley & Sons, 1981.

Galot, G., *Cours de calcul des probabilités,* Paris, Dunod, 1967.

Jenicek, M. et R. Cléroux, *Épidémiologie. Principes Techniques Applications,* Montréal, Edisem, 1982.

Kleinbaum, D., L.L. Kupper et H. Morgenstern, *Epidemiologic Research.* Belmont(USA), Lifetime Learning Publications, 1982.

Martel, J.M. et R. Nadeau, *Probabilités en gestion et en économie,* Chicoutimi, Gaétan Morin, 1980.

Rumeau-Rouquette, C., G. Bréart et R. Padieu, *Méthodes en épidémiologie,* Paris, Flamarion Médecine Sciences, 1970.

Rothman, K.J., *Modern Epidemiology,* Boston, Little Brown and Company, 1986.

Rothman, K.J. et S. Greenland, *Modern Epidemiology,* Philadelphia, Lippincott_Raven, 1998.

Scherrer, B., *Biostatistique,* Chicoutimi, Gaétan Morin, 1984.

Schlesselman, J.J., *Case-control Studies,* Oxford, Oxford University Press, 1982.

Schwartz, D., *Méthodes statistiques à l'usage des médecins et des biologistes,* Paris, Flammarion Médecine-Sciences, 1963.

Snedecor, G.W. et W.G. Cochran, *Méthodes statistiques*. Traduction de l'ouvrage publié en langue anglaise sous le titre *Statistical methods*, 6e édition, réalisée par H. Boelle et E. Camhaji, Paris, Association de coordination technique agricole, 1957.

Tricot, C. et J.M. Picard, *Ensemble et statistique,* Montréal, McGraw-Hill, 1969.

INDEX

C

D

E

F

H

I

J

K

L

M

P

R

V

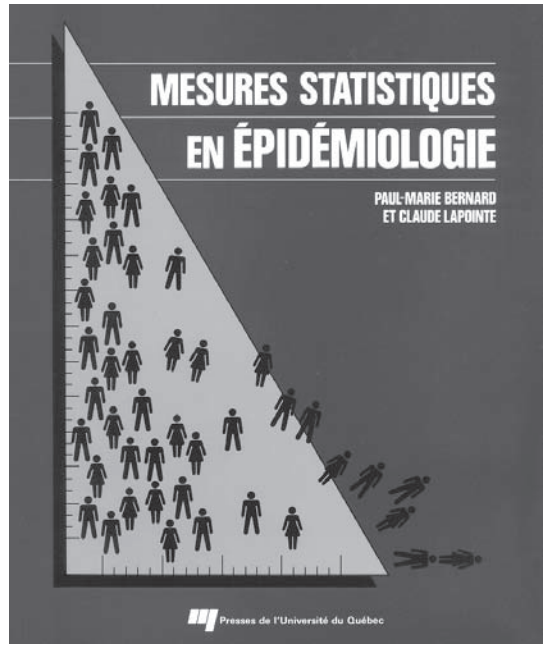
W

MESURES STATISTIQUES EN ÉPIDÉMIOLOGIE

Paul-Marie Bernard et Claude Lapointe

328 pages
ISBN 2-7605-0446-8

32$



À découvrir

Variables et échelles de classification • Types d'études en épidémiologie • Mesures descriptives générales d'un ensemble de données • Mesures de fréquence • Espérance de vie • Mesures d'association • Mesures d'impact • Mesures d'interaction • Mesures d'accord • Mesure de probabilité • Mesure de la probabilité de survie • Mesures de validité des tests diagnostiques (ou de dépistage) • Valeur-p ou degré de signification • Notion de justesse • Biais dans les mesures d'association RR et RC • Ajustement des mesures • Intervalle de confiance.

Prix sujet à changement sans préavis

www.puq.ca 418 · 657-4399

Pour démarrer le cédérom
Analyse des tableaux de contingence en épidémiologie,
insérez-le dans votre lecteur.

Si le cédérom ne démarre pas de lui-même :

- *Pour un ordinateur de type PC*
  Ouvrir « Poste de travail », double-cliquez sur votre lecteur cédérom, puis double-cliquez sur le fichier index.html.
- *Pour un ordinateur de type MAC*
  Double-cliquez sur l'icône du cédérom, puis double-cliquez sur le fichier index.html.

Configuration minimale

Windows®95 et supérieur / Pentium 200 mhz, 64 Mo Ram
Macintosh OS 9.0 et supérieur / Power PC 200 mhz, 64 Mo Ram.

Veuillez activer toutes les fontes de votre ordinateur,
en particulier les polices *Symbol* et *MT Extra*.
Le cédérom est optimisé pour Internet Explorer 5.0 ou supérieur pour PC.