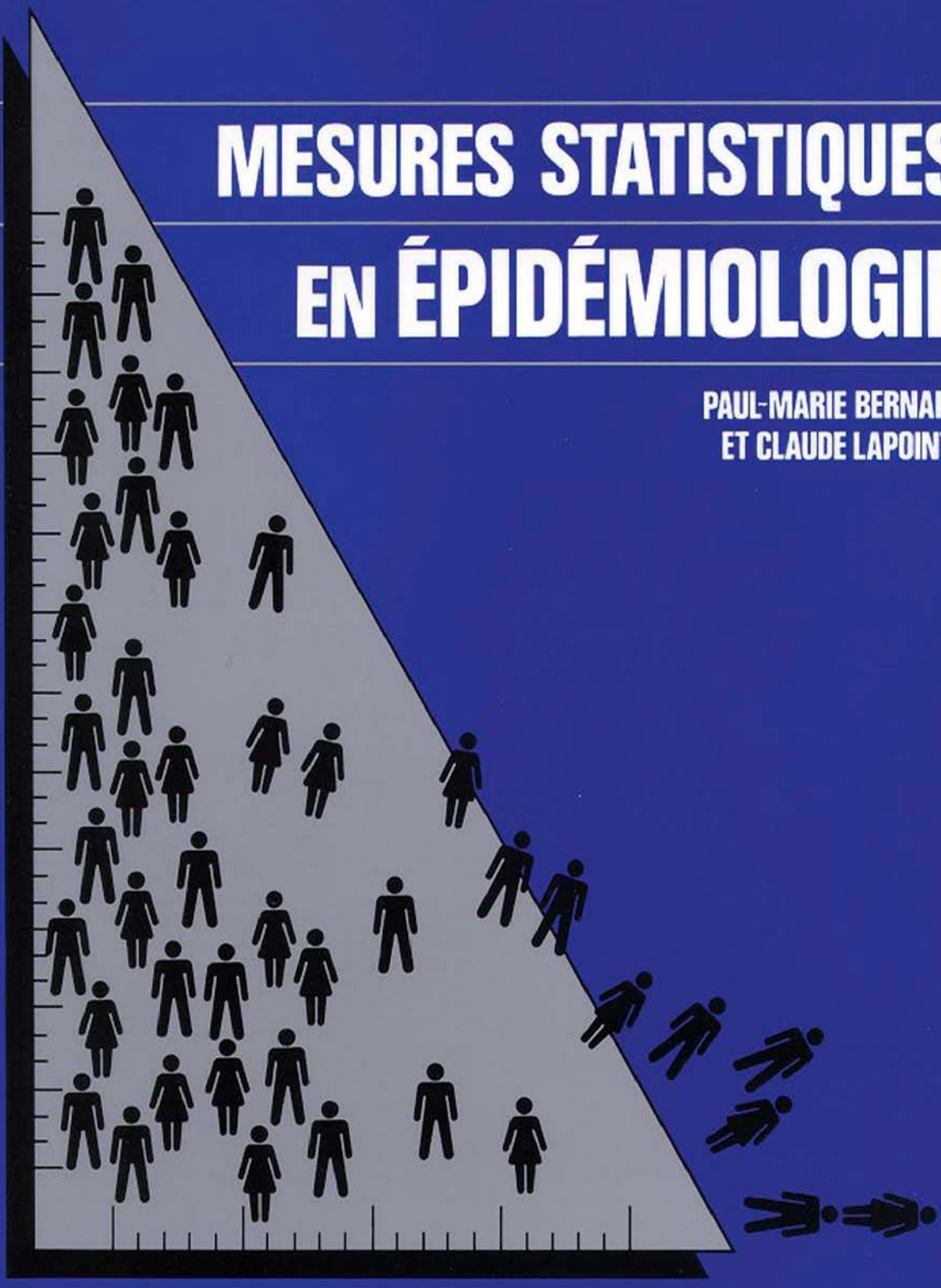
MESURES STATISTIQUES EN ÉPIDÉMIOLOGIE

PAUL-MARIE BERNARD
ET CLAUDE LAPOINTE



Presses de l'Université du Québec

MESURES STATISTIQUES EN ÉPIDÉMIOLOGIE

PRESSES DE L'UNIVERSITÉ DU QUÉBEC
2875, boul. Laurier, Sainte-Foy (Québec) GIV 2M3
Téléphone : (418) 657-4399 • Télécopieur : (418) 657-2096
Courriel : secretariat@puq.uquebec.ca
Catalogue sur Internet : http://www.uquebec.ca/puq

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
LIBRAIRIE DU QUÉBEC À PARIS
30, rue Gay-Lussac, 75005 Paris, France
Téléphone: 33 1 43 54 49 02
Télécopieur: 33 1 43 54 39 15

BELGIQUE
S.A. DIFFUSION—PROMOTION—INFORMATION
Département la Nouvelle Diffusion
24, rue de Bosnie, 1060 Bruxelles, Belgique
Téléphone : 02 538 8846
Télécopieur: 02 538 8842

SUISSE
GM DIFFUSION SA
Rue d'Etraz 2, CH-1027 Lonay, Suisse
Téléphone : 021 803 26 26
Télécopieur: 021 803 26 29

Paul-Marie BERNARD et Claude LAPOINTE

MESURES STATISTIQUES EN ÉPIDÉMIOLOGIE

1998

Presses de l'Université du Québec
2875, boul. Laurier, Sainte-Foy (Québec) G1V 2M3

Données de catalogage avant publication (Canada)

Bernard, Paul-Marie, 1942-

Mesures statistiques en épidémiologie

Comprend des références bibliographiques et un index. ISBN

2-7605-0446-8

1. Épidémiologie – Méthodes statistiques.
1. Lapointe, Claude, 1938- . II. Titre.

RA652.2M3B47 1987 614.4'07'2 C87-011340-2

Les Presses de l'Université du Québec remercient le Conseil des arts du Canada
et le Programme d'aide au développement de l'industrie de l'édition du Patrimoine canadien pour
l'aide accordée à leur programme de publication.

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

Dépôt légal – 3e trimestre 1987
Bibliothèque nationale du Québec / Bibliothèque nationale du Canada
Imprimé au Canada

TABLE DES MATIÈRES

AVANT-PROPOS

Le titre de ce livre évoque le caractère grandement quantitatif de l'épidémiologie. L'épidémiologie est la science qui permet de quantifier l'apparition et la répartition de la maladie dans les populations humaines, de quantifier les relations entre la maladie et les caractéristiques des individus et de leur environnement, de quantifier l'impact de certains facteurs ou interventions sur la santé des individus. La quantification de ces phénomènes passe par l'utilisation de mesures.

Nous présentons les principales mesures devenues essentielles à la pratique actuelle de l'épidémiologie. Nous les qualifions de mesures statistiques puisqu'elles sont définies en référence à des groupes de personnes et calculées à partir d'ensembles de données. Cet ouvrage n'est pas pour autant un livre à proprement parler de statistique. Les méthodes de cette discipline ne sont pas réellement abordées ici, exception faite principalement des deux chapitres sur la valeur-p ou degré de signification et sur l'estimation par intervalle de confiance. Il ne s'agit pas non plus d'un livre d'épidémiologie puisqu'il ne s'intéresse pas aux connaissances acquises par l'épidémiologiste dans le domaine de la santé; il appartient plutôt au domaine de la méthodologie générale en épidémiologie.

Dans la préparation de ce texte, nous avons mis à profit nos années d'expérience dans l'enseignement des méthodes en épidémiologie et biostatistique au deuxième cycle universitaire. A titre de professeurs, nous déplorons de devoir constamment utiliser comme supports pédagogiques des ouvrages de langue anglaise et surtout américains. Ce commentaire, sans rien enlever à la valeur et à la qualité de ces textes, traduit plutôt le souci de mieux respecter la réalité linguistique de notre clientèle étudiante. En plus de répondre à cette préoccupation à forte incidence pédagogique, ce volume comble un vide important dans la littérature de langue française. A notre connaissance, il n'existe pas en français d'ouvrage qui présente de façon systématique et exhaustive les mesures les plus couramment utilisées en épidémiologie intégrant les mesures de base de la statistique descriptive. Notre volume est le résultat d'un effort consenti dans ce sens.

La présentation du contenu de cet ouvrage fait qu'il tient plutôt du manuel que du traité. Chaque mesure est définie, non seulement au plan conceptuel, mais aussi au plan opérationnel. Après la présentation de l'une d'elles, le lecteur devrait être capable non seulement de comprendre ce qu'elle est, ce qu'elle mesure, mais aussi de la calculer sur un ensemble de données. Les exemples concrets, les analogies, les répétitions, comme dans l'enseignement, sont abondants. Et pour donner au texte un caractère universel, nous avons choisi des exemples fictifs mais vraisemblables. Enfin, pour enrayer le caractère parfois ésotérique de certaines relations, nous avons ajouté tantôt des illustrations, tantôt des démonstrations formelles pour le lecteur plus habile ou plus averti. Ces démonstrations accessoires, non

essentielles à la compréhension du texte, sont clairement identifiées dans la présentation par une surface ombragée.

L'approche pédagogique utilisée nous a obligés à une présentation simple mais rigoureuse qui, pour l'essentiel, évite l'utilisation de notions médicales, statistiques ou mathématiques avancées. Dans cette optique, la lecture de ce volume ne suppose aucune formation préalable en médecine ou en statistique. Elle ne requiert que des connaissances mathématiques élémentaires en algèbre et en géométrie. Toute personne ayant une bonne maîtrise des mathématiques du niveau secondaire peut facilement suivre au plan formel les développements ou démonstrations rencontrées au passage.

Ce volume est d'abord destiné à servir de manuel pour un cours de base en épidémiologie des programmes de maîtrise en épidémiologie, en santé publique, en santé communautaire et en santé au travail. Il peut aussi intéresser tous les professionnels de la santé appelés à lire la littérature scientifique, à évaluer un rapport ou à participer à des projets de recherche nécessitant la connaissance des méthodes et des mesures épidémiologiques. L'agent de recherche d'un centre ou service de santé, le clinicien, le médecin en santé au travail sont autant de professionnels qui peuvent tirer profit de cet outil. Enfin, ce manuel s'adresse aux étudiants ou chercheurs de discipline non-médicale appelés à utiliser les mesures de base en épidémiologie; ce peut être des chercheurs en sociologie, en anthropologie, en histoire, en relations industrielles, etc.

Le texte comprend dix-sept chapitres regroupés en cinq parties. La partie I regroupe les deux premiers chapitres qui servent de préambule. On y trouve les notions fondamentales de variable (chapitre 1), de population, de même qu'une présentation des différents types d'études en épidémiologie (chapitre 2). Les parties II, III et IV regroupent les chapitres qui, à proprement parler, présentent les différentes mesures. La partie II porte sur les mesures de base: les mesures descriptives générales d'un ensemble de données (chapitre 3), les mesures de fréquence (chapitre 4) et la mesure particulière qu'est l'espérance de vie (chapitre 5). La partie III porte sur les mesures composées : les mesures d'association comme le risque relatif, le coefficient de corrélation (chapitre 6), les mesures d'impact (chapitre 7) les mesures d'interaction ou de synergie (chapitre 8) et certaines mesures d'accord (chapitre 9). La partie IV porte sur les mesures qui relèvent plus directement des notions de probabilité. Après une présentation sommaire des mesures de probabilité et de leur calcul (chapitre 10), on trouve la probabilité cumulative de survie (chapitre 11), les mesures de validité d'un test diagnostique (chapitre 12) et la valeur-p comme mesure de vraisemblance d'une hypothèse (chapitre 13). La partie V porte sur le problème de la justesse des mesures. D'abord, nous présentons la notion de justesse et ses deux composantes: la validité et la précision (chapitre 14); par la suite, nous approfondissons la notion de validité de certaines mesures d'association (chapitre 15) et discutons du problème de l'ajustement (chapitre 16). Enfin, dans le cadre de la précision des mesures, nous présentons la notion d'intervalle de confiance et décrivons ce type d'intervalle pour les mesures les plus usitées (chapitre 17). Pour alléger la lecture de certains chapitres, nous avons annexé des notions jugées accessoires et d'approche plus difficile pour le lecteur moins averti.

Nous donnons à la fin de chaque chapitre la liste des symboles qui y ont été utilisés et des formules qui y ont été présentées. Nous indiquons aussi quelques suggestions de lecture; elles réfèrent à des parties de livre qui peuvent être l'occasion pour le lecteur soit d'approfondir, soit simplement d'aborder autrement plusieurs des thèmes développés dans le chapitre.

La numérotation des chapitres ne conditionne pas nécessairement l'ordre de leur lecture. On peut choisir différents cheminements de lecture pourvu qu'ils respectent quelques contraintes. Certains chapitres peuvent être utiles, sinon obligatoires, à la compréhension de certains autres. La grille suivante indique, pour chacun des chapitres en ordonnée, celui ou ceux, en abscisse, qui peuvent lui être utiles (U) ou obligatoires (O), en tout ou en partie.

CHAPITRE PRÉALABLEMENT LU

CHAPITRE À LIRE	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1																	
2																	
3	U																
4	U	U															
5	U	U		O													
6	U	O	O	O													
7	U	O		O		O											
8	U	O		O		O	U										
9			O														
10				U													
11	U			U						O							
12	U			U						O							
13										O							
14		U				U											
15	U	O		O		O		U		U		U		O			
16	U	O		O		O	O							O	O		
17			O	O		O			O	U	O			U			

U = UTILE
O = OBLIGATOIRE

De par sa nature, un manuel transmet des connaissances qui sont largement puisées dans le patrimoine scientifique mondial. Nous adressons alors nos premiers remerciements à toutes les personnes, connues et inconnues, qui par leur travail ont enrichi ce patrimoine. Ils ont été inconsciemment de précieux collaborateurs pour nous. Nous remercions aussi tous nos collègues du Département de médecine sociale et préventive de la Faculté de médecine de l'Université Laval. Nous avons apprécié leurs conseils et leurs suggestions. La préparation du manuel a commencé alors que le professeur Fernand Turcotte était directeur du département. Nous lui témoignons toute notre reconnaissance pour nous avoir incités à poursuivre le travail de rédaction. Nous avons été sensibles aux encouragements que nous a apporté l'actuelle directrice du département, madame Thérèse Morais. Nous voulons également souligner l'excellent accueil reçu aux Presses de l'Université du Québec. Nous remercions enfin nos familles respectives pour leur soutien et leur patience pendant tout le temps qu'a duré ce travail.

PARTIE I

Notions préalables

CHAPITRE 1

Variables et échelles de classification

Dans ce chapitre sont d'abord présentées quelques variables importantes, omniprésentes, ou presque, dans plusieurs études épidémiologiques; toutes les variables sont ensuite classifiées formellement, suivant leur nature, en variables quantitatives ou qualitatives, discrètes ou continues. Les valeurs prises par une variable conduisent à l'idée de classe de valeurs, donc à celle d'échelle de classification. Une distinction est finalement faite entre quatre types d'échelle.

Pour étudier la répartition d'une maladie et les différentes circonstances qui entourent son apparition et son développement au sein d'une population, l'épidémiologiste est appelé à regarder un certain nombre de variables descriptives, souvent assez fortement reliées au problème considéré. Le choix de variables se fait en tenant compte de leur pertinence vis-à-vis les objectifs de l'étude envisagée.

VARIABLES EN ÉPIDÉMIOLOGIE

Les variables en épidémiologie peuvent être regroupées suivant les trois aspects qui permettent de caractériser la maladie: les personnes atteintes, le lieu et le moment où elles ont été atteintes. On trouve donc les trois grandes familles de variables en épidémiologie : les variables de personnes, les variables de lieux et les variables (ou, pour mieux dire, la variable) de temps.

Variables de personnes

Les variables de personnes réfèrent aux attributs anatomiques, physiologiques, sociaux ou culturels. Les plus fréquemment utilisées en épidémiologie sont l'âge, le sexe, l'état civil, les habitudes de vie, l'occupation et le niveau socio-économique. On tient compte de certaines variables de personnes en épidémiologie pour les raisons suivantes :

- L'étude de la variation de la fréquence d'une maladie suivant certaines variables de personnes peut permettre de mieux comprendre les facteurs responsables de cette maladie.
- L'association entre certaines de ces variables et la maladie peut voiler le rôle d'autres facteurs.
- L'effet d'autres facteurs peut être modifié par la présence de certaines caractéristiques de personnes.
- Une bonne description de la maladie suivant les caractéristiques de personnes permet généralement de mieux identifier l'intervention préventive ou curative à entreprendre.

Variables de lieux

L'étude de la *répartition géographique* de la fréquence d'une maladie suscite toujours l'intérêt de 1'épidémiologiste. La fréquence d'une maladie peut varier suivant le pays ou la région, le climat ou selon que la population habite dans une zone urbaine ou rurale.

Variables de temps

De façon générale, la fréquence de la maladie varie avec le *temps.* Par exemple, la fréquence du cancer du poumon a fortement augmenté au cours des trente dernières années. La grippe est un phénomène saisonnier. La durée est aussi une caractéristique de la maladie qui permet de marquer sa gravité et son évolution. Le temps, comme variable présente à tout phénomène, est donc un élément nécessaire à la définition des mesures épidémiologiques et une composante de base du concept de cause.

VARIABLES SUR UN PLAN FORMEL

Chaque variable considérée doit être clairement définie; certaines, comme le sexe, le sont d'emblée. Mais aura-t-on retenu le niveau socio-économique comme variable dans une étude qu'il faudra la définir explicite- ment. On écartera ainsi toute ambiguïté dans la compréhension des variables, ce qui est indispensable à la clarté d'une étude, qu'elle soit médicale, épidémiologique ou de santé publique. Mais, formellement, qu'est- ce qu'une variable?

Variable (ce qui peut varier)

On appelle *variable* tout caractère sujet à prendre des états différents suivant les individus, le temps ou le lieu d'observation. Ainsi en est-il, par exemple, du sexe, de l'âge, du groupe sanguin, de la tension artérielle, du nombre de lits par hôpital, de la durée d'hospitalisation.

Tout état possible que peut prendre le caractère étudié est une *valeur* pour une variable. A est une valeur pour la variable groupe sanguin, 420 une valeur pour le nombre de lits dans un hôpital. Une variable qui ne prend que deux valeurs est dite *dichotomique ou binaire.* C'est le cas du sexe: masculin, féminin; ou du fait de fumer: oui, non.

Classification des variables

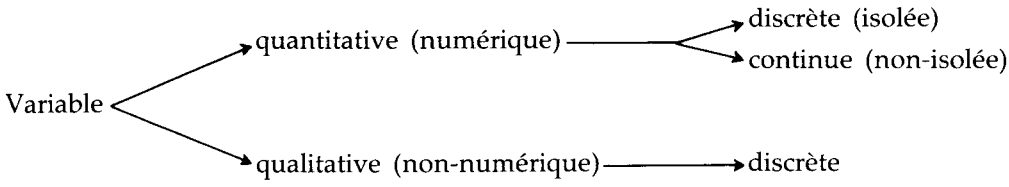
Les variables ne sont pas toutes de même nature. Elles se distinguent d'abord par la nature numérique ou non de leurs valeurs, ensuite par le fait que leurs valeurs sont de nature isolée ou non les unes par rapport aux autres. Ces distinctions que nous faisons ici entraînent la classification des variables en variables quantitatives ou qualitatives, discrètes ou continues, comme l'illustre la figure 1-1.

VARIABLE QUANTITATIVE

Une variable est *quantitative si* les valeurs qu'elle prend sont d'emblée de nature numérique, des quantités. On distingue les variables quantitatives en variables discrètes et en variables continues.

Une variable quantitative est *discrète* lorsque ses valeurs sont des quantités isolées, séparées les unes des autres. Elles sont isolées en ce sens qu'entre deux valeurs quelconques observables de la variable, il existe toujours une valeur non-observable. La variable « nombre d'enfants par famille » prend les valeurs 0, 1, 2, 3, 4, etc. Entre les valeurs observables 3 et 4, il existe au moins une valeur non-observable, comme 3,1. Une famille peut avoir 3 ou 4 enfants, mais non 3,1 enfants. Le nombre d'enfants par famille

Figure 1-1



prend des valeurs isolées. C'est une variable quantitative discrète. Les valeurs d'une telle variable sont connues ou obtenues par dénombrement.

Une variable quantitative est *continue* lorsque ses valeurs sont n'importe quelle quantité dans un certain intervalle. Cela veut dire que toute valeur entre deux valeurs observables quelconques de la variable est théoriquement observable. C'est le cas, par exemple, de la taille des individus. Toute valeur entre les deux valeurs observables 173 et 174 cm, par exemple, est théoriquement observable. La taille exacte d'un individu peut être 173,1 cm, pour un autre 173,14 cm, etc. Toutes les valeurs dans un certain intervalle sont possibles. La taille est une variable quantitative continue. Les valeurs d'une telle variable sont connues ou obtenues par un procédé de mesure au sens strict.

VARIABLE QUALITATIVE

Une variable est *qualitative si* les valeurs qu'elle prend correspondent à des qualités, des attributs. Ainsi en est-il du sexe (masculin, féminin), du groupe sanguin (A, B, AB, O), du stade d'un cancer (I, II, III, IV). La variable qualitative est de nature discrète.

Il est utile de savoir reconnaître si une variable est qualitative ou quantitative, discrète ou continue. Le choix des instruments de description statistique et de mesure d'une variable dépend de la nature de celle-ci. Par exemple, on calculera volontiers la moyenne arithmétique des valeurs d'une variable quantitative discrète ou continue, comme la moyenne d'enfants par famille ou la tension artérielle systolique moyenne. Pour une variable qualitative, on calculera plutôt une proportion, comme la proportion d'individus de groupe sanguin A.

CLASSEMENT DES OBSERVATIONS

Le classement des observations faites sur les individus est la première étape à franchir pour organiser des données statistiques. Pour une variable donnée, les observations référant à une même valeur (ou ensemble de valeurs) sont regroupées dans une même classe (ou catégorie) définie par cette ou ces valeurs. Pour l'âge, par exemple, toutes les observations qui donnent 34 ans comme valeur peuvent être regroupées dans la classe 34 ans. Ou encore, dans la définition de classes plus larges, toutes les observations qui réfèrent à une valeur se situant entre 30 et 34 ans inclusivement peuvent être regroupées dans la classe 30-34 ans. Une *classe* réfère donc à une valeur ou à un regroupement de valeurs contiguës d'une variable.

Échelle de classification

Pour une variable donnée, l'ensemble des classes (ou catégories) définit ce que l'on appelle une *échelle de classification.* Les quatre classes (A, B, AB, O) constituent une échelle de classification pour le groupe sanguin. Les classes 40 kilos et moins, 40-49, 50-59, 60-69, 70-79 et 80 kilos et plus forment une échelle de classification possible pour le poids.

Une échelle de classification doit permettre de classer toutes les observations, chacune ne pouvant être classée que dans une catégorie. Ainsi, pour qu'un classement des observations soit correct, les classes de l'échelle doivent satisfaire les deux conditions suivantes:

- Elles doivent être *mutuellement exclusives.* Chaque individu ou encore chaque observation de la variable ne peut appartenir qu'à une seule classe. Les classes d'âge 1-5 ans, 5-15, 15-25, et 25 ans et plus ne sont pas mutuellement exclusives, car l'individu de 15 ans appartient à plus d'une classe (ici à deux); par contre, les classes 1-4 ans, 5-14, 15-24, et 25 ans et plus le sont.
- Elles doivent être *collectivement exhaustives.* Chaque individu ou encore chaque observation de la variable doit appartenir à une classe. Les deux classes A et O du groupe sanguin ne sont pas collectivement exhaustives, car un individu peut n'appartenir à aucune de ces deux classes.

Pour certaines variables, comme le sexe, l'investigateur n'a aucune liberté quant au choix de l'échelle de classification, tandis que pour d'autres, plusieurs possibilités s'offrent à lui, d'aucunes étant jugées plus pertinentes. Veut-on uniquement distinguer les fumeurs des non-fumeurs ou cherche-t-on aussi à séparer les petits fumeurs des gros fumeurs? En tout cas, une échelle doit satisfaire les deux propriétés fondamentales qui sont les caractères d'exclusivité et d'exhaustivité de leurs classes.

Le choix, l'adoption ou la construction d'une échelle de classification sont la base de l'organisation des données statistiques.

Types d'échelles

En distinguant les variables quantitatives et qualitatives, discrètes et continues, on peut répartir les échelles de classification suivant essentiellement quatre types.

ÉCHELLE NOMINALE

Dans une échelle *nominale,* les classes ne sont que nommées. Ainsi en est-il pour le sexe (masculin, féminin), le groupe sanguin (A, B, AB, O) ou le diagnostic de sinusite aiguë (sinusite maxillaire aiguë, frontale aiguë, autres sinusites aiguës).

ÉCHELLE ORDINALE

Dans une échelle *ordinale,* il existe une relation d'ordre entre les classes. C'est le cas pour l'échelle relative à l'évolution de l'état de santé d'un patient: amélioration, stabilité, détérioration. C'est le cas aussi pour le degré de satisfaction face aux soins dispensés par un service de santé: peu, moyennement, très satisfait.

ÉCHELLE PAR INTERVALLE

Dans une échelle *par intervalle, il* existe une notion de distance entre les valeurs. Mentionnons, à titre d'exemples, l'échelle moins de 0°C, 0-9, 10-19, 20-29, et 30°C et plus pour la température (climat) ou l'échelle 0-4 ans, 5-14, 15-24, 25-34, et 35 ans et plus pour l'âge. Si des individus appartiennent à la classe d'âge 15-24 ans, d'autres à la classe 25-34 ans, on peut dire que leur différence (distance) d'âge est en moyenne de dix ans.

ÉCHELLE PROPORTIONNELLE

Dans une échelle *proportionnelle,* la notion de rapport entre les grandeurs existe. Pour deux personnes, âgées respectivement de 30 et 10 ans, on peut dire que la première a vingt ans de plus que la seconde, mais on peut dire aussi

qu'elle est trois fois plus âgée. L'échelle pour l'âge est non seulement par intervalle mais aussi proportionnelle. L'échelle pour la température moins de 0°C, 0-9, 10-19, 20-29 et 30°C et plus est une échelle par intervalle mais non proportionnelle. Quand on compare une température de 30°C à celle de 10°C, on constate une différence de 20°C. Mais on ne peut pas dire que 30°C (ou 86°F) indique une température trois fois plus chaude que 10°C (ou 50°F).

En définitive, l'échelle nominale permet de répondre à la question: « Qui est qui? » Qui appartient au groupe sanguin A? L'échelle ordinale permet de répondre à une question plus forte: « Qui est plus? ou Qui est moins? ». Qui, parmi les patients, est le plus satisfait des soins dispensés par une clinique? L'échelle par intervalle permet de répondre à une question encore plus forte, du genre « Combien plus? ou Combien moins? ». De combien d'années un individu est-il plus âgé qu'un autre? Finalement, l'échelle proportionnelle permet de répondre à la question « Combien de fois plus? ou Combien de fois moins? » De combien de fois un individu est-il plus âgé qu'un autre? Il se dégage ainsi une hiérarchie des échelles qui va de la plus simple à la plus complexe:

— nominale: nomination;

— ordinale: nomination, ordre;

— par intervalle : nomination, ordre, distance;

— proportionnelle : nomination, ordre, distance, rapport.

Pour les variables qualitatives, seules les échelles nominale et ordinale peuvent être envisagées; les échelles par intervalle ou proportionnelle ne sont utilisables que pour les variables quantitatives. Toutefois, là où une échelle par intervalle ou proportionnelle peut être utilisée, l'investigateur peut préférer une échelle nominale ou ordinale pour des raisons liées aux objectifs de son étude. Un investigateur pourrait adopter l'échelle ordinale : hypotendu, normotendu, hypertendu, au lieu d'une échelle par intervalle pour la tension artérielle. En pratique, on ne distingue généralement pas les échelles par intervalle et proportionnelle.

RÉSUMÉ

D'un point de vue pratique, les variables considérées en épidémiologie et en santé publique se répartissent en trois grandes familles: les variables de personnes, les variables de lieux et les variables de temps; l'âge, le sexe, les habitudes de vie, l'état civil, l'occupation, le niveau socio-économique, le climat sont autant d'exemples de telles variables. D'un point de vue formel, les variables sont des caractères sujets à prendre des états différents suivant les individus, le temps, le lieu, etc. Les variables sont qualitatives ou quantitatives, discrètes ou continues. Leurs valeurs sont regroupées en classes mutuellement exclusives et collectivement exhaustives pour former une échelle de classification; on compte des échelles nominales, ordinales, par intervalle ou proportionnelles.

LECTURES SUGGÉRÉES

1. JENICEK, M. et CLÉROUx, R. ***Épidémiologie,*** Saint-Hyacinthe, Edisem, 1982, chapitre 5, pp. 93-118.

2. MAC MAHON, B. et PUGH, T.F. ***Epidemiology: Principles and Methods,*** Boston, Little, Brown, 1970, chapitres 7, 8, 9 et 10, pp. 103-206.

CHAPITRE 2

Types d'études en épidémiologie

Les mesures utilisées en épidémiologie sont fortement reliées aux types d'études pratiquées. Ce chapitre propose une classification des études épidémiologiques d'abord en deux grandes catégories: non-expérimentales (ou d'observation) et expérimentales. À leur tour, les études non-expérimentales se divisent en études descriptives et en études à visée étiologique. La subdivision continue pour en arriver finalement, à l'intérieur des études à visée étiologique, aux stratégies bien connues comme les études de cohorte et les études cas-témoins.

Les études en épidémiologie peuvent se diviser en deux grandes catégories: les études non-expérimentales et les études expérimentales. Dans les études *non-expérimentales,* la réalité est observée telle qu'elle se présente spontanément à l'observateur, sans qu'il intervienne sur cette réalité. Pour déterminer une association possible entre, par exemple, le fait de fumer et une maladie, le chercheur ne peut manipuler le facteur cigarette en décidant qui sera fumeur et qui ne le sera pas. Dans les études *expérimentales,* le chercheur manipule le facteur pour ensuite observer l'effet; l'exposition au facteur n'est pas déterminée spontanément mais décidée par l'investigateur. Par exemple, dans une étude pour évaluer un nouveau traitement, un chercheur peut déterminer qui sera sou mis au nouveau traitement et qui sera sou mis au placebo. Généralement, dans une étude expérimentale, le chercheur répartit au hasard les patients ou individus en deux ou plusieurs groupes de comparaison. Lorsqu'il y a manipulation du facteur par l'investigateur sans que les individus soient répartis par tirage au sort, certains auteurs parlent d'études *quasi expérimentales.* Les études non expérimentales sont aussi appelées études *d'observation,* du fait que la réalité est observée comme elle apparaît à l'observateur. Cette appellation, bien que très utilisée, est un peu délicate dans la mesure où des observations sont aussi faites dans les études expérimentales.

Avant de traiter plus spécifiquement des différents types d'études, il est intéressant de distinguer le genre de population auprès de laquelle on est appelé à conduire une étude. Par *population,* on entend l'ensemble des individus visés par l'étude. Du point de vue de l'observateur, une population est statique ou dynamique. Le terme « statique » correspond à l'idée d'un état considéré sans référence à l'évolution dans le temps, tandis que le terme « dynamique » appelle l'idée de mouvement, donc de changement dans le temps.

La *population statique* est composée d'individus observés à un même moment. L'observateur s'intéresse à l'état des individus tels qu'ils existent à ce moment particulier et les observations ne sont faites que pour ce moment. Le terme « moment » réfère soit à une date du calendrier, soit à un âge, soit à un événement particulier.

La *population dynamique* comprend les individus observés au cours d'une période de temps. L'observateur s'intéresse à certains événements qui peuvent affecter les individus durant cette période (la maladie, le décès ...). Le terme « dynamique » traduit deux aspects qui caractérisent une telle population. D'une part, chaque individu subit des changements dans certaines de ses caractéristiques, par exemple, l'âge, le lieu de résidence, la maladie, etc.; d'autre part, la composition de la population elle-même se modifie en cours d'observation: les individus du début de l'étude ne sont pas forcément les mêmes que ceux de la fin de l'étude. Ces changements sont dus aux naissances, à la migration, à la maladie, à la mortalité, etc.

Pour être admis dans la population, un individu doit satisfaire un certain nombre de critères spécifiés, par exemple, pour l'âge, le sexe, le lieu de résidence, etc. Tout individu qui obéit à ces critères pour une période déterminée par l'investigateur est admis dans la population. Une population dynamique peut être soit fermée, soit ouverte.

Une population dynamique est dite *fermée* si les individus admis (entrés en observation) ne peuvent pas la quitter à moins que se produise le décès ou l'événement qui intéresse l'observateur: le décès dans une étude de survie, la maladie dans une étude de morbidité, etc. Dans une population fermée, passée la période d'admission dans l'étude, aucun nouvel individu ne peut s'y ajouter; après cette période, elle ne peut que connaître des pertes au plan des effectifs. Nous convenons que le terme *cohorte* remplace l'expression « population dynamique fermée ».

Les études de survie sont généralement conduites auprès de cohortes, c'est-à-dire de populations fermées. Pour une région et une période déterminées, on recense tous les patients qui ont reçu pour la première fois le traitement (T) contre la maladie (M). Tous ces patients suivis pendant deux ans depuis la date de leur traitement sont observés pour le décès. Ce groupe constitue une population fermée. Une fois admis dans l'étude et pour toute la période d'observation fixée, un patient ne peut quitter l'observation sauf s'il décède. Tout changement, de résidence, d'âge ou d'autres caractéristiques, en théorie, ne modifie nullement son appartenance au groupe observé. (En pratique cependant, il peut arriver qu'à la fin de l'étude l'investigateur soit incapable de retrouver un ou plusieurs sujets admis en observation.)

Une population dynamique est dite *ouverte* si les individus la quittent, non seulement parce que le décès ou l'événement qui intéresse l'observateur a pu se produire, mais aussi parce que l'un ou l'autre des critères d'entrée n'est plus respecté. Dans une population ouverte, des individus peuvent s'ajouter par naissance ou par immigration, par changement d'âge ou autrement; des individus peuvent aussi s'y soustraire par décès, par émigration, par changement d'âge ou autrement. Un individu de 19 ans, par exemple, fera partie de la population ouverte des 20-39 ans quand il aura atteint 20 ans en cours d'observation. Un autre quittera cette même population parce qu'en cours d'observation son âge aura passé de 39 à 40 ans. La population d'une région (ville), pour laquelle on veut connaître la mortalité sur une période d'un an, constitue l'exemple d'une population ouverte. Cette région subit des changements continuels dans sa composition. Il y a des naissances, des décès, des départs, des arrivées, etc.

En résumé, dans une population fermée, les critères d'entrée ne sont applicables qu'au moment de l'entrée. L'individu, une fois admis dans une population fermée, acquiert la permanence pour la période d'observation (à moins de défaillance). Dans une population ouverte, les critères d'entrée sont applicables à tout moment de la période d'observation. L'individu dans une population ouverte n'acquiert pas la permanence; il peut en sortir si l'un des critères d'admission n'est plus satisfait.

ÉTUDES NON-EXPÉRIMENTALES

Dans les études non-expérimentales, rappelons-le, l'investigateur observe la réalité telle qu'elle se présente spontanément à lui. Les études non-expérimentales sont descriptives ou à visée étiologique.

Études descriptives

En épidémiologie, une étude *descriptive* décrit un problème de santé dans une population

ou un groupe d'individus et en établit la fréquence selon certaines variables de personnes, de lieux et de temps.

Selon que la population étudiée est statique ou dynamique, ou encore que l'investigateur s'intéresse à la description d'un état ou du changement, l'étude est transversale ou longitudinale. Une étude descriptive *transversale* permet de décrire un problème de santé à un moment donné; ce genre d'étude génère des données de prévalence. Une étude descriptive *longitudinale* implique l'observation d'une population sur une période et génère des données d'incidence ou de mortalité.

Les études de la fréquence d'une maladie ou du décès, à des dates ou des périodes différentes du calendrier, constituent une classe particulière d'études descriptives. Ce sont les études de *tendance.* Généralement, elles utilisent des données de prévalence, d'incidence ou de mortalité déjà collectées. L'investigateur répète les mêmes observations à des moments ou des périodes différents. L'étude d'une tendance suppose que la collecte des données soit faite au moins à trois moments ou périodes distincts. De telles études peuvent couvrir une période totale aussi longue que vingt ans et même plus. L'analyse et l'interprétation des résultats qui proviennent d'une étude de tendance visent principalement à distinguer trois effets reliés au temps: l'effet de l'âge, l'effet de période et l'effet de cohorte.

- *L'effet de l'âge* existe si la fréquence de la maladie varie avec l'âge indépendamment de la période étudiée et de l'année de naissance des individus.
- *L'effet de période* existe si la fréquence de la maladie varie avec le temps indépendamment de l'âge des individus et de leur année de naissance.
- *L'effet de cohorte* existe si la fréquence de la maladie varie en fonction de l'année de naissance des individus indépendamment de leur âge.

Une tendance observée peut ainsi être le résultat d'un de ces trois effets ou d'une combinaison de ceux-ci.

En santé publique, les études descriptives permettent de mesurer l'importance d'un problème de santé, d'en tracer le profil suivant des variables choisies et conséquemment, d'identifier des groupes à risque. La connaissance de ces groupes permet de mieux orienter les interventions de santé publique. En recherche épidémiologique, ces études permettent d'explorer des associations possibles entre des facteurs et des maladies et conduisent souvent à la formulation d'hypothèses étiologiques.

Études à visée étiologique

Les études à visée étiologique cherchent à déterminer le rôle que peuvent jouer un ou des facteurs dans l'étiologie d'une ou plusieurs maladies. Dans leur forme la plus simple, ces études génèrent des mesures d'association entre un facteur d'exposition et une maladie. Les études à visée étiologique tendent à répondre à des questions comme : Existe-il une relation entre le facteur et la maladie? Quelle est l'intensité de cette relation? etc.

Dans une étude, l'observation du facteur et celle de la maladie peuvent être faites chez les individus au même moment ou à des moments différents. Dans le premier cas, on

dit qu'elle est faite de manière synchrone et on la qualifie donc de *transversale. Si* les observations sont faites de façon asynchrone, c'est-à-dire à des moments différents, l'étude est dite *longitudinale.* Dans une telle étude, les moments d'exposition au facteur et d'apparition de la maladie sont distingués, en présumant que l'exposition précède la maladie.

Nous allons maintenant examiner plus en détail les études étiologiques en commençant par les études longitudinales qui sont de loin les plus utilisées en recherche étiologique. Dans le reste du chapitre, nous nous limiterons, dans le seul souci de simplification, à ne présenter que la situation dichotomique: absence ou présence de l'exposition, absence ou présence de la maladie.

ÉTUDES À VISÉE ÉTIOLOGIQUE LONGITUDINALES

Les études longitudinales suivent une démarche directe ou à rebours. La démarche est *directe si* l'étude débute par la classification des individus suivant la présence ou l'absence du facteur. Ces individus sont alors suivis pendant une période pour déterminer qui sera affecté par la maladie. La démarche est *à rebours (ou* inverse) si l'étude commence par la classification des individus selon qu'ils sont malades ou non. Ces individus sont alors investigués pour déterminer qui a été exposé au facteur.

Une autre distinction des études longitudinales peut être faite quant au rapport entre le début de l'étude et le moment d'exposition au facteur ou de l'apparition de la maladie. De ce point de vue, une étude est prospective, rétrospective ou ambispective (néologisme tiré de la littérature (étatsunienne). L'étude est *prospective si* elle débute à un moment où ni l'exposition au facteur ni l'apparition de la maladie n'ont eu lieu; elle est *rétrospective* si les observations sont faites à un moment où l'exposition au facteur et la maladie ont déjà eu lieu; elle est *ambispective* si elle combine les caractères prospectifs et rétrospectifs. Dans ce cas, l'étude débute à un moment où a déjà eu lieu, en partie, soit l'exposition au facteur, soit l'apparition de la maladie.

Une étude longitudinale directe est généralement appelée *étude de cohorte* parce qu'elle implique l'observation d'une ou plusieurs cohortes (populations fermées). Une étude à rebours est dite *étude cas-témoins.* Les études longitudinales à rebours ne sont pas les seules études cas-témoins. Nous verrons plus loin qu'une étude cas-témoins peut être aussi transversale.

ÉTUDES DE COHORTE

Les *études de cohorte* se définissent essentiellement par la comparaison de deux groupes d'individus, les uns exposés au facteur et les autres non. La comparaison est faite quant à la fréquence de l'événement maladie ou décès survenu en cours d'observation. A partir de ce type d'étude, on peut estimer la fréquence de la maladie chez les sujets exposés et celle chez les sujets non-exposés.

Un exemple célèbre d'étude de cohorte est celle de Doll et Hill sur le tabagisme et le cancer du poumon. La population visée, les médecins britanniques, comportait deux groupes d'individus, les fumeurs et les non-fumeurs, qui ont été suivis pendant plus de vingt ans et observés pour l'événement décès par cancer du poumon.

Suivant le procédé d'identification des groupes de sujets exposés et de sujets non- exposés, les études de cohorte se divisent en études de cohorte de population et en études de cohorte sur échantillons électifs.

Études de cohorte de population. Les études sont conduites auprès de la population en- fière ou d'un échantillon aléatoire. La population (ou l'échantillon obtenu par tirage au sort) se divise spontanément en deux groupes sans l'intervention du chercheur: les sujets exposés et les sujets non-exposés. Spontanément, la population de l'étude de Doll et Hill se divise en deux groupes: les fumeurs et les non-fumeurs.

Dans ce type d'étude, on peut connaître ou estimer la fréquence de la maladie respectivement chez les sujets exposés et chez les sujets non-exposés, ainsi que la proportion de sujets exposés dans la population. La figure 2-1 présente les deux proportions ou fréquences qui font l'objet de comparaison dans ce genre d'étude.

Études de cohorte sur échantillons électifs. Dans ces études, l'investigateur choisit lui-même le groupe de sujets exposés au facteur et le groupe des sujets non-exposés; il détermine également leur taille respective. Ce type d'étude convient particulièrement lorsque l'exposition au facteur est rare. Pour étudier le rôle du chlorure de vinyle dans le cancer hépatique, on choisit un groupe d'individus exposés dans leur milieu de travail à cette substance et un autre groupe d'individus non-exposés, mais le plus comparable possible au premier groupe.

La figure 2-2 présente les deux proportions qui font l'objet de comparaison dans les études de cohorte sur échantillons électifs.

Les résultats d'une étude de cohorte de population ou sur échantillons électifs peuvent être disposés comme dans le tableau 2-1A ou 2-1B selon qu'il s'agit de données de personnes-temps ou de personnes.

Figure 2-1

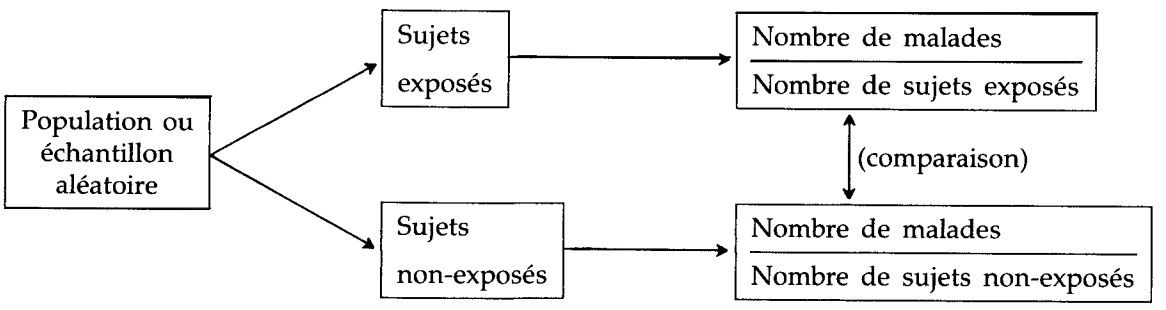


Figure 2-2

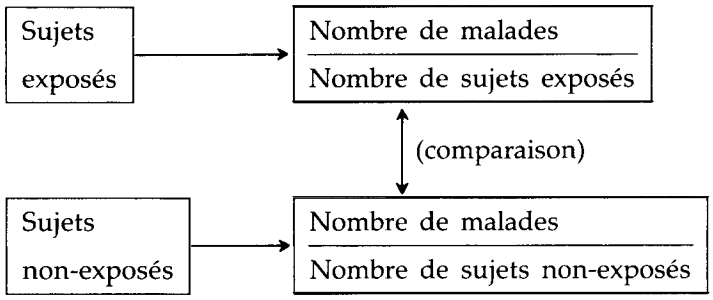


Tableau 2-1A

		E +	E −	Total
M	+	a	b	M_1
Personnes-temps		N_1	N_0	N

Tableau 2-1B

		E +	E −	Total
M	+	a	b	M_1
	−	c	d	M_0
Personnes		N_1	N_0	N

E et M dénotent respectivement le facteur d'exposition et la maladie, et + et — respectivement la présence et l'absence du facteur ou de la maladie.

Une étude de cohorte peut être prospective, rétrospective ou ambispective. Généralement, les études de cohorte rétrospectives peuvent être conduites plus rapidement que les études de cohorte prospectives; les premières sont souvent utilisées en santé au travail pour détecter les problèmes reliés à une occupation ou une exposition spécifique.

ÉTUDES CAS-TÉMOINS

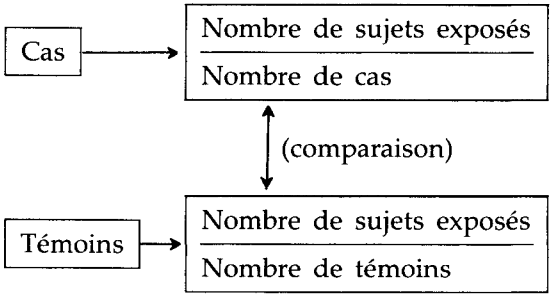
Les *études cas-témoins* se définissent essentiellement par la comparaison de deux groupes, l'un composé d'individus atteints de la maladie considérée (ou décédés des suites de celle-ci), c'est-à-dire les cas, et l'autre composé d'individus n'ayant pas la maladie (ou non décédés des suites de celle-ci), c'est-à-dire les témoins; les deux groupes sont comparés quant à l'importance de l'exposition au facteur. La figure 2-3 présente les deux proportions qui font l'objet de comparaison dans les études cas-témoins.

Un groupe de malades hospitalisés pour cancer broncho-pulmonaire (cas) et un groupe de malades hospitalisés pour d'autres raisons (témoins) sont comparés sur leurs habitudes antérieures de tabagisme (facteur d'exposition). Les résultats d'une telle étude peuvent être disposés dans un tableau 2 *x* 2 comme au tableau 2-2.

Voici quelques points de comparaison entre les études de cohorte et les études cas-témoins.

- Dans les études de cohorte prospectives, l'information obtenue sur l'exposition au facteur et sur d'autres caractéristiques de la personne peut difficilement être faussée par la connaissance du résultat (maladie ou décès). L'information sur l'exposition est obtenue avant que l'on ne sache si la maladie est présente ou absente. Par contre, dans les études cas-témoins, la connaissance de la maladie ou du décès peut fausser l'information sur l'exposition.
- Les études de cohorte se prêtent bien à l'étude d'un facteur susceptible d'être associé à un éventail de maladies. Les études cas-témoins permettent d'examiner plus facilement un éventail de facteurs susceptibles d'être associés à une maladie particulière.
- Les études de cohorte prospectives sont généralement plus coûteuses et plus longues que les études cas-témoins.
- Les études de cohorte sont inefficaces lorsque la maladie est rare; les études cas-témoins sont plus difficiles à réaliser lorsque l'exposition est rare.

Figure 2-3



ÉTUDES À VISÉE ÉTIOLOGIQUE TRANSVERSALES

Rappelons que la présence de l'exposition et celle de la maladie sont déterminées au même moment dans les études à visée étiologique transversales. Ces études sont inaptes à déterminer la chronologie des événements exposition—maladie. Elles se divisent en trois catégories suivant le type d'échantillonnage: 1) échantillons non-électifs, 2) échantillons électifs par rapport à l'exposition; et 3) échantillons électifs par rapport à la maladie.

ÉTUDES SUR ÉCHANTILLONS NON-ÉLECTIFS

Dans une étude à visée étiologique transversale sur *échantillon non-électif,* le sujet est admis dans l'étude sans avoir été au préalable classé quant à l'exposition ou à la maladie. C'est seulement après son admission dans l'étude qu'il est classé suivant ces deux variables. A titre d'exemple, on peut considérer l'examen d'un groupe de travailleurs

Tableau 2-2

	E		
	+	−	Total
Cas	a	b	M_1
Témoins	c	d	M_0

pour déterminer une association possible entre le tabagisme et l'hypertension. L'investigateur mesure la tension artérielle des travailleurs et détermine leurs habitudes actuelles en la matière. Une étude sur échantillon non-électif peut comprendre à la limite tous les sujets d'une population.

ÉTUDES FAITES À PARTIR D'ÉCHANTILLONS ÉLECTIFS PAR RAPPORT À L'EXPOSITION

Dans une étude à visée étiologique transversale faite à partir *d'échantillons électifs par rapport à l'exposition,* l'investigateur choisit un groupe d'individus exposés au facteur et un second groupe de sujets non-exposés. Chaque individu est alors investigué pour déterminer l'existence ou non de la maladie. Il en serait ainsi, par exemple, d'une étude où les travailleurs d'une industrie exposés au bruit seraient examinés pour détecter des problèmes de surdité, puis comparés aux cadres de la même industrie.

ÉTUDES FAITES À PARTIR D'ÉCHANTILLONS ÉLECTIFS PAR RAPPORT À LA MALADIE

Dans une étude à visée étiologique transversale faite à partir *d'échantillons électifs par rapport à la maladie,* l'investigateur choisit un groupe d'individus atteints de la maladie et un groupe d'individus non atteints et les compare quant à l'exposition actuelle au facteur. Pour déterminer l'association entre le diabète juvénile et la présence de certains groupes HLA, un chercheur compare un groupe de jeunes diabétiques à un groupe de jeunes n'ayant pas la maladie. Le dosage des HLA s'effectue chez les deux groupes au moment de l'observation. Une telle étude de type transversal à échantillons électifs par rapport à la maladie se comprend aussi comme une étude cas-témoins. Lorsque le facteur d'exposition est un caractère permanent (par exemple génétique), ce type d'étude peut être confondu avec l'étude longitudinale à rebours.

ÉTUDES EXPÉRIMENTALES

L'étude expérimentale est caractérisée par le fait que le chercheur manipule le facteur étudié. Supposons qu'un investigateur veuille évaluer l'efficacité d'une intervention en santé publique. Le facteur, ici l'intervention, sera manipulé si l'investigateur décide d'appliquer cette intervention à une population de son choix. En la comparant, à partir de certains critères, à une autre population non soumise à l'intervention, il pourra se prononcer sur l'efficacité de l'intervention. Dans cet exemple, l'assignation d'une population au facteur, c'est-à-dire à l'intervention, relève directement du chercheur. On parle alors d'étude expérimentale *non-randomisée.* Dans d'autres études, l'assignation ou l'affectation des individus peut être faite par tirage au sort. La répartition aléatoire des sujets en deux ou plusieurs groupes est connue sous le nom de randomisation. On

parle alors d'étude expérimentale *randomisée.* La figure 2-4 illustre ce type d'étude.

Pour évaluer l'efficacité d'un vaccin corn- me agent préventif d'une maladie, on peut convenir de pratiquer une randomisation sur un échantillon d'individus tiré de la population qu'on veut protéger. A l'un des groupes randomisés, on administre le vaccin, à l'autre le placebo. Ce n'est donc ni le chercheur, ni l'individu qui décide du groupe d'appartenance, mais le hasard. Les résultats d'une telle étude peuvent être disposés dans un tableau 2 x 2 comme le tableau 2-3.

En résumé, s'il y a manipulation du facteur sans randomisation, alors on peut parler d'étude expérimentale non-randomisée ou encore d'étude *quasi expérimentale.* En revanche, l'étude est dite expérimentale randomisée lorsqu'il y a manipulation du facteur avec randomisation. L'étude randomisée permet un meilleur contrôle de l'influence des facteurs autres que celui étudié, puisque la randomisation, en principe, équilibre les groupes quant à ces autres facteurs. La randomisation permet théoriquement de constituer des groupes d'individus aussi semblables que possible sur l'ensemble des caractéristiques comme l'âge, le sexe, le niveau socio-économique, les signes cliniques, etc. à l'exception, bien entendu, du facteur étudié. Si, en principe, la randomisation assure la comparabilité des groupes sans tenir compte de leurs caractéristiques particulières, en pratique le hasard peut entraîner des différences pour certaines caractéristiques, surtout si les groupes comparés comprennent peu de sujets. Toutefois, les études randomisées sont préférables aux études non-randomisées parce qu'elles se soustraient à l'assignation arbitraire ou sélective des sujets. Les procédés de répartition non-aléatoire rendent les groupes plus difficilement comparables.

Le laboratoire, le milieu clinique et la santé communautaire constituent les trois principaux champs d'utilisation des études expérimentales. En milieu clinique, on parle d'essai clinique ou thérapeutique; en santé communautaire, l'étude expérimentale peut être qualifiée d'étude d'intervention et, s'il s'agit d'étude d'intervention à caractère préventif, on peut parler d'essai préventif.

Nous avons proposé une classification des études en épidémiologie. Il faut toutefois remarquer en terminant que toute classification, toute typologie, comprend une part d'arbitraire et qu'elle est rarement définitive. Il peut s'avérer plus ou moins difficile de classer certaines études dans l'une ou l'autre des catégories évoquées ici. La figure 2-5 résume *une* classification possible des études en épidémiologie.

Figure 2-4

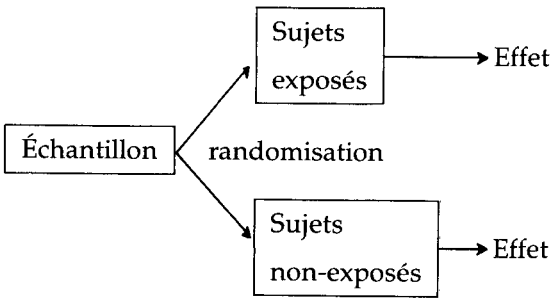
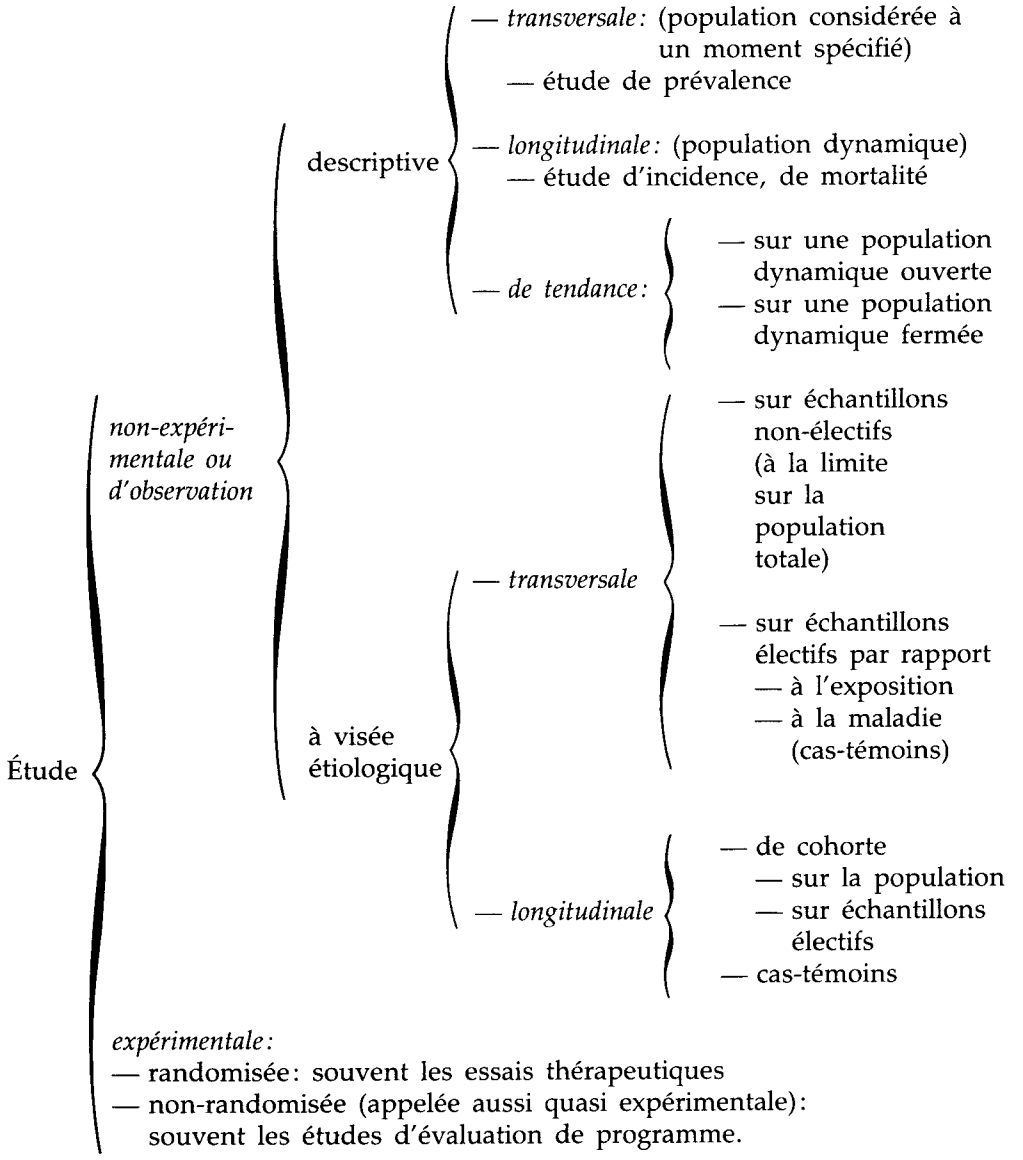


Tableau 2-3

		Vaccin	Placebo
Maladie	Oui	*a*	*b*
	Non	*c*	*d*
	Total	N_1	N_0

Figure 2-5 Classification des études



RÉSUMÉ

En épidémiologie, les études peuvent être expérimentales, mais elles sont généralement non-expérimentales ou d'observation. Les études portent sur des populations statiques ou dynamiques. Les études non-expérimentales se divisent en études descriptives et en études à visée étiologique. L'étude descriptive s'intéresse, en termes de fréquence, aux problèmes de santé dans une population; l'étude à visée étiologique s'intéresse à l'association entre facteurs et maladies. Les études à visée étiologique sont transversales ou longitudinales suivant que le facteur et la maladie sont observés ou non au même moment. Les études à visée étiologique longitudinales se divisent en études de cohorte et études cas-témoins. L'étude de cohorte se définit essentiellement par la comparaison de deux groupes d'individus les uns exposés au facteur, les autres non, la comparaison étant faite quant à la fréquence de l'événement maladie ou décès survenu en cours d'observation. L'étude cas-témoins se définit essentiellement par la comparaison de deux groupes, l'un composé d'individus atteints de la maladie considérée (cas) et l'autre composé d'individus n'ayant pas la maladie (témoins). La comparaison est faite entre les deux groupes sur l'importance de l'exposition. Une étude cas-témoins peut aussi être de type transversal.

LECTURES SUGGÉRÉES

1. GOLDBERG, *M. L'Épidémiologie sans peine,* Paris, Éditions médicales Roland Bettex, 1986, pp. 45-49, 111-123.

2. KLEINBAUM, D.G., KUPPER, L.L. et MORGENSTERN, H. ***Epidemiologic Research,*** Belmont (USA), Lifetime Learning Publications, 1982, chapitre 5, pp. 62-95.

3. ROTHMAN, K.J. ***Modern Epidemiology,*** Boston, Little, Brown, 1986, chapitre 6, pp. 51-76.

4. RUMEAU-ROUQUETTE, C., BREART, G. et PADIEU, R. *Méthodes en épidémiologie,* Paris, Flammarion, 1985, chapitre II, pp. 14-21.

PARTIE II

Mesures de base

CHAPITRE 3

Mesures descriptives générales d'un ensemble de données

L'observation répétée d'une variable conduit à la formation d'un ensemble plus ou moins grand de valeurs, de données, qu'il peut être utile d'organiser ou de résumer. Ce chapitre présente le tableau de fréquences et le graphique comme modes d'organisation de données. Sont définies ensuite des mesures de tendance centrale qui servent à résumer les données, comme les deux moyennes, arithmétique et géométrique, la médiane et le mode. Quatre mesures de dispersion sont étudiées: l'étendue, la variance, le coefficient de variation et l'intervalle semi-interquartile. Enfin, sont présentés les centiles.

Dans une étude portant sur 121 patients de plus de 50 ans, on a déterminé pour chacun d'eux la tension artérielle systolique au mmHg près. Les 121 valeurs observées figurent au tableau 3-1.

Ce tableau présente une masse de données quantitatives qui se distingue plutôt par un manque que par un excès d'organisation. Les 121 observations ont été notées, enregistrées au fur et à mesure qu'elles sont devenues disponibles. Elles se présentent pour ainsi dire dans le désordre, sans aucune forme particulière d'organisation et, en ce sens, forment un ensemble de données *brutes*. Généralement, ce désordre voile les réponses à des questions, même simples, que l'on pourrait se poser. Par exemple, on n'a pas d'emblée une idée de l'ordre de grandeur de la proportion de patients dont la tension systolique dépasse 170 mmHg. On ne connaît pas davantage la manière dont se répartissent les valeurs observées, c'est-à-dire l'importance relative de différents groupes de valeurs. Une masse de données est un grand déploiement d'information qui, paradoxalement, en cache une bonne partie.

Il devient nécessaire d'organiser les données brutes si on veut les rendre plus intelligibles, plus accessibles, plus claires. Acquis à l'idée d'organisation, on doit envisager une forme de présentation des données, c'est-à-dire une manière de faire connaître les traits essentiels, de fournir une description de la structure générale de la série de valeurs observées. Les tableaux de fréquences, éventuellement leurs représentations graphiques, et certaines mesures comme celles de tendance centrale et de dispersion sont des instruments utiles à l'organisation et à la présentation des données.

TABLEAUX DE FRÉQUENCES

La présentation d'une série de valeurs observées d'une variable, comme les 121 valeurs de la tension artérielle, peut se faire à l'aide d'un tableau de fréquences. Il faut d'abord convenir d'une échelle de classification pour la variable tension artérielle. Supposons que l'échelle retenue soit 120-129 mmHg, 130-139, ..., 190-199 mmHg. Il n'y a pas de règle ferme sur le

Tableau 3-1

140	150	138	146	146	198	152	136	143	152	145
130	162	138	152	139	154	146	136	144	133	151
142	151	127	148	154	148	150	158	151	148	142
137	145	149	141	157	158	150	158	154	134	142
122	157	154	168	157	161	150	156	141	149	151
143	133	160	142	151	140	148	143	134	152	148
144	159	160	171	165	155	159	145	159	152	144
145	158	160	175	166	163	147	153	147	167	146
180	166	141	131	151	132	139	153	139	156	143
144	155	147	135	150	164	157	146	155	149	151
149	162	149	153	145	151	156	144	151	169	155

nombre de classes, mais le plus souvent il se situe entre 5 et 15. De façon générale, il faut éviter les extrêmes, c'est-à-dire trop ou trop peu de classes.

Le nombre d'observations (de valeurs observées) qui tombent dans une classe est appelé la *fréquence (ou* l'effectif) de cette classe. La fréquence de la classe de tension artérielle 130-139 mmHg est égale à 16, c'est-à-dire qu'il y a 16 patients dont la tension se situe entre 130 et 139 mmHg, (pour être plus exact entre 129,5 et 139,5). La proportion d'observations qui tombent dans une classe est appelée la *fréquence relative* de cette classe et peut être exprimée en pourcentage. La fréquence relative de la classe 130-139 mmHg pour la tension artérielle est égale à $^{16}/_{121}$, soit 13,2 %.

L'ensemble des classes d'une échelle avec leur fréquence (relative) constitue ce que l'on appelle une *distribution de fréquences (relatives)*. Une distribution de fréquences illustre comment se distribuent, se répartissent les différentes valeurs ou groupes de valeurs observées d'une variable.

Un *tableau de fréquences* est une façon concrète de présenter une distribution de fréquences. Il comprend essentiellement deux colonnes :

1. la colonne des classes;
2. la colonne des fréquences (ou des fréquences relatives).

Se référant toujours à l'exemple sur la tension artérielle et en adoptant l'échelle déjà convenue, on obtient le tableau de fréquences 3-2.

On peut aussi construire un tableau de fréquences pour une variable qualitative. Le tableau 3-3 en est un exemple; il décrit la distribution selon le sexe des 121 patients.

REPRÉSENTATION GRAPHIQUE DES DISTRIBUTIONS DE FRÉQUENCES

La représentation graphique est un complément visuel au tableau de fréquences. Elle permet de saisir rapidement et facilement les grands traits d'une distribution. Les modes de représentation graphique varient selon le type d'échelle, donc, d'une certaine façon, suivant la nature de la variable. Nous nous limiterons ici aux représentations graphiques des distributions

Tableau 3-2

Tension artérielle systolique (mmHg)	Patients	
	Nombre	Pourcentage
120-129	2	1,7
130-139	16	13,2
140-149	41	33,9
150-159	44	36,4
160-169	14	11,6
170-179	2	1,7
180-189	1	0,8
190-199	1	0,8
Total	**121**	**100**

Tableau 3-3

	Patients	
Sexe	Nombre	Pourcentage
Masculin	55	45,5
Féminin	66	54,5
Total	**121**	**100**

de fréquences de variables quantitatives, plus spécifiquement à l'histogramme et au polygone de fréquences.

Histogramme

L'histogramme est un mode de représentation utile pour les distributions de fréquences (relatives) de variables quantitatives continues ou discrètes. Il est construit dans un système d'axes rectangulaires et comprend un ensemble de rectangles adjacents qui répondent aux trois conditions suivantes :

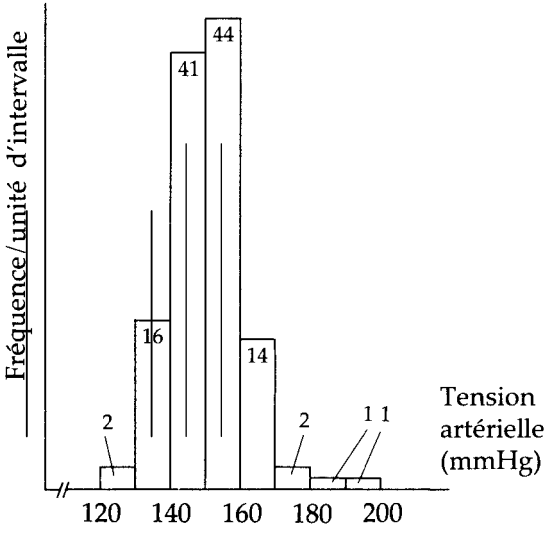
- Les rectangles se suivent selon l'ordre des classes. Pour l'exemple au tableau 3-2, le premier rectangle correspond à la première classe, 120-129; le deuxième rectangle à la deuxième classe, 130-139; etc.
- Chacune des bases des rectangles coïncide avec l'intervalle de la classe correspondante. Dans l'exemple, la base du premier rectangle va de 120 à 129, plus strictement de 119,5 à 129,5; celle du second de 129,5 à 139,5; etc.
- Chacune des aires des rectangles (base x hauteur) mesure la fréquence (relative) de la classe correspondante. En conséquence, l'échelle verticale (la hauteur) doit représenter les fréquences par unité d'intervalle de la variable. Pour l'exemple du tableau 3-2, l'aire du deuxième rectangle de l'histogramme correspondant doit mesurer 16: 10 (base: intervalle) x 1,6 (hauteur: fréquence par unité d'intervalle). Ainsi construit, ce deuxième rectangle a une aire huit fois plus grande que celle du premier rectangle, étant donné que leurs fréquences sont dans le rapport de 16 à 2. Il en va ainsi pour les autres rectangles.

L'histogramme à la figure 3-1 représente la distribution de fréquences de la tension artérielle des 121 patients, décrite au tableau 3-2. Le point milieu (point médian) du premier intervalle est 125, du second 135, etc.

La distribution des fréquences des tensions artérielles systoliques des 121 patients monte graduellement jusqu'à 155 mmHg, pour ensuite redescendre. La représentation graphique de la distribution des fréquences relatives conduirait au même histogramme.

Notons qu'au tableau 3-2 les classes de tension artérielle ont un même intervalle, soit 10 mmHg. Dans ce cas, les bases des rectangles à la figure 3-1 sont toutes égales. Il s'ensuit que les hauteurs des rectangles sont proportionnelles aux aires, donc aux fréquences. Le deuxième rectangle est huit

Figure 3-1



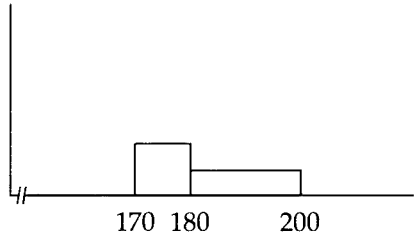
fois plus haut que le premier. Dans ce cas, l'échelle verticale pourrait tout aussi bien représenter les fréquences que les fréquences par unité d'intervalle de la variable, sans que l'allure de l'histogramme en soit modifiée.

Si les intervalles de classes sont inégaux, les bases le sont aussi. Dans ce cas, il n'est plus vrai que les hauteurs des rectangles sont proportionnelles aux aires, donc aux fréquences. L'axe vertical ne peut pas décrire les fréquences de classe. Imaginons que la dernière classe du tableau 3-2 soit 180-199. L'intervalle de cette classe est alors le double de celui de la classe précédente, 170-179, 20 contre 10 mmHg. Mais pour les deux classes, les fréquences sont identiques, soit deux sujets. Les rectangles correspondants doivent avoir la même aire suivant la troisième condition. Pour ce faire, il faut corriger en abaissant la hauteur du dernier rectangle par un facteur de 2 comme à la figure 3-2. Les aires respectives sont obtenues par:

10 (base) x 0,2 (hauteur) pour l'avant dernier rectangle;

20 (base) x 0,1 (hauteur) pour le dernier rectangle.

Figure 3-2

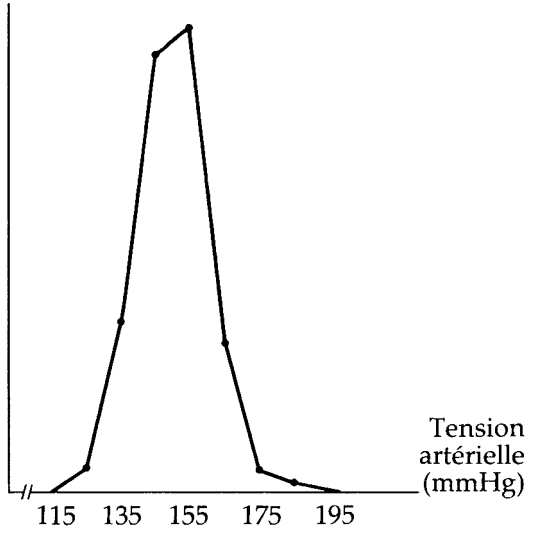


Polygone de fréquences

Le *polygone de fréquences* est un autre mode de représentation graphique pour les distributions de fréquences (relatives) de variables quantitatives continues ou discrètes. Il convient généralement mieux que l'histogramme lorsqu'il s'agit de représenter plusieurs distributions de fréquences sur un même système d'axes. On peut construire le polygone de fréquences soit à partir de l'histogramme, soit encore sans aucune référence à celui-ci. Pour simplifier, disons que le polygone est obtenu en joignant le point milieu du sommet de chaque rectangle d'un histogramme au milieu du sommet des rectangles adjacents. Pour les données du tableau 3-2, on obtient le polygone de fréquences de la figure 3-3.

Ce polygone de fréquences, qui est une sorte de ligne brisée, a été construit en considérant des intervalles de classes égaux à 10 (120-129, 130-139, ...).

Figure 3-3



Si on pouvait augmenter indéfiniment le nombre de sujets ou d'observations tout en réduisant de plus en plus l'intervalle de classe, le polygone deviendrait à la limite parfaitement lisse et tendrait, pour la distribution des tensions, vers une courbe symétrique comme celle de la figure 3-4. D'autres variables peuvent conduire, après lissage, à des courbes asymétriques comme celles des figures 3-5A et 3-5B.

MESURES DE TENDANCE CENTRALE

Si on se reporte au tableau 3-2 sur la tension artérielle systolique, les valeurs proches de 150 mmHg apparaissent plus centrales que les valeurs 130 et 180, lesquelles occupent des positions plutôt périphériques dans la distribution. Des *valeurs centrales* se caractérisent donc par le fait que les valeurs observées de la variable tendent à se rassembler autour d'elles. On peut imaginer une valeur centrale comme une sorte de valeur typique autour de laquelle gravitent les valeurs observées d'une variable. Une valeur centrale est une valeur unique qui résume une série d'observations.

Nous allons maintenant définir des mesures qui nous permettent d'obtenir des valeurs centrales. Les deux mesures les plus répandues sont la moyenne arithmétique et la médiane. Nous définissons également la moyenne géométrique et le mode.

Figure 3-4

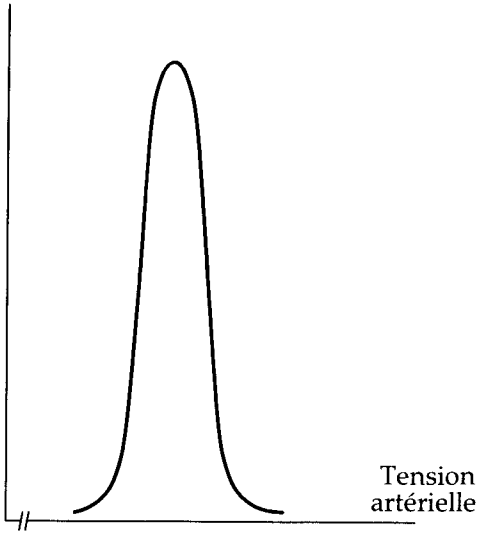


Figure 3-5A

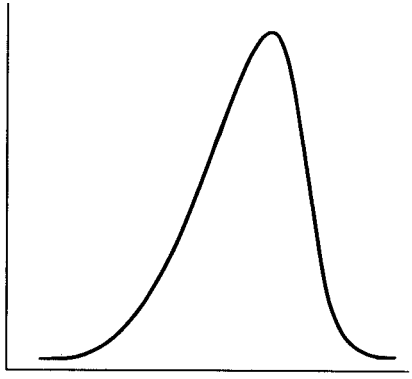
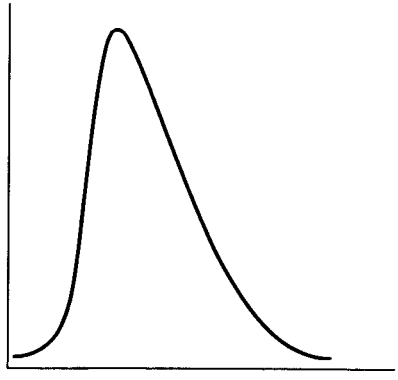


Figure 3-5B



Moyenne arithmétique

La moyenne arithmétique est la plus connue et la plus utilisée de toutes les mesures de tendance centrale. On l'appelle simplement « la moyenne ». Elle correspond à l'expression courante « en moyenne ». Qu'on se rappelle l'exemple scolaire de la note moyenne de la classe.

La *moyenne arithmétique (ma)* est la somme de chacune des valeurs observées d'une variable divisée par le nombre de valeurs observées, c'est-à-dire par la fréquence totale. Si on désigne les n valeurs observées d'une variable par $x_1, x_2, \ldots, x_n$, alors on aura compris que

$$m_a = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

ou, en abrégé,

$$m_a = \frac{\Sigma x_i}{n}$$

La moyenne arithmétique de la tension artérielle systolique des 121 patients est donc:

$$m_a = \frac{18\,143}{121} = 149{,}9 \text{ mmHg}.$$

La moyenne arithmétique peut être calculée aussi bien sur des données quantitatives discrètes que sur des données quantitatives continues. Si le nombre moyen d'enfants par famille de 2,4 heurte le sens commun, il n'en demeure pas moins que cette abstraction est utile.

Moyenne géométrique

Il est assez facile de vérifier que:

$$(5 \times 10)^{1/2} = 7{,}1$$

$$(5 \times 10 \times 14)^{1/3} = 8{,}9$$

$$(5 \times 10 \times 14 \times 17)^{1/4} = 10{,}4$$

etc.

Chaque fois, on observe que la valeur ainsi calculée est relativement centrale par rapport aux valeurs sur lesquelles s'effectue le calcul. Ce procédé de calcul définit une mesure de tendance centrale appelée moyenne géométrique. De façon plus formelle, si on désigne les n valeurs observées d'une variable par $x_1, x_2 \ldots x_n$, la *moyenne géométrique (m_g)* est la racine nième de leur produit :

$$m_g = (x_1 \times x_2 \times \ldots \times x_n)^{1/n}$$

ou

$$m_g = \sqrt[n]{x_1 \times x_2 \times \ldots \times x_n}$$

Elle n'est définie que pour des valeurs observées positives. Pour la tension artérielle systolique des 121 patients, on trouve $m_g = 149{,}6$ mmHg.

Le logarithme de la moyenne géométrique est égal à la moyenne arithmétique des logarithmes de chaque donnée. En effet,

$$\log m_g = \log (x_1 \times x_2 \times \ldots \times x_n)^{1/n}$$

$$= \frac{1}{n} \log (x_1 \times x_2 \times \ldots \times x_n)$$

$$= \frac{1}{n} [\log x_1 + \log x_2 + \ldots + \log x_n]$$

$$= \frac{1}{n} \Sigma \log x_i$$

Médiane

La *médiane (mé)* est une valeur qui divise l'ensemble des valeurs observées, disposées en ordre croissant ou décroissant, en deux parties égales, c'est-à-dire de même fréquence. Théoriquement, 50 % des valeurs sont inférieures à la médiane, 50 % lui sont supérieures.

Pour la série 3 — 5 — 8 — 9 — 12, la médiane est égale à 8, soit deux valeurs à sa gauche et deux à sa droite.

Pour la série 3—5—8—9—12—13, la médiane est située entre 8 et 9. On la prendra égale à 8,5 soit le point milieu entre 8 et 9. On trouve trois valeurs à sa gauche et trois valeurs à sa droite.

La médiane est une valeur (observée ou non) de rang $(n + 1)/2$, si n est le nombre total de valeurs observées. Pour la première série ci-dessus, la médiane occupe le troisième rang, $(5 + 1)/2$. Elle est de rang $(6 + 1)/2 = 3,5$ dans la deuxième série, c'est-à-dire se situe entre la troisième valeur et la quatrième. En ce qui regarde la tension artérielle systolique des 121 patients, l'ordre des valeurs observées est inscrit au tableau 3-4.

La 61e valeur, de rang $\frac{121 + 1}{2}$, correspond à la médiane; c'est la valeur en caractère gras au tableau 3-4.

On devrait théoriquement obtenir le même nombre de valeurs, tant au-dessus qu'en-dessous de 150. Or, on en compte 59 d'un côté et 57 de l'autre. Cette difficulté pratique, souvent inévitable, résulte du phénomène de l'arrondi. Les cinq valeurs 150 sont indiscernables pour une tension connue au mmHg près. Les tensions seraient connues au dixième de mmHg que déjà plu- sieurs de ces valeurs 150 seraient différenciées. On pourrait avoir ainsi :

149,7 ⟶ 60e valeur

150,1 ⟶ 61e valeur

150,2 ⟶ 62e valeur

150,4 ⟶ 63e valeur

150,4 ⟶ 64e valeur

Tableau 3-4

122	136	141	144	146	149	151	152	155	158	164
127	137	142	144	146	149	151	153	156	159	165
130	138	142	144	147	149	151	153	156	159	166
131	138	142	145	147	149	151	153	156	159	166
132	139	142	145	147	150	151	154	157	160	167
133	139	143	145	148	150	151	154	157	160	168
133	139	143	145	148	150	151	154	157	160	169
134	140	143	145	148	150	152	154	157	161	171
134	140	143	146	148	150	152	155	158	162	175
135	141	144	146	148	151	152	155	158	162	180
136	141	144	146	149	151	152	155	158	163	198

La médiane serait alors de 150,1 avec 60 valeurs de part et d'autre. La connaissance de la tension au dixième de mmHg est si peu réaliste qu'il faut comprendre que la médiane est égale à 150 mmHg. En pratique, cela signifie qu'il y a de chaque côté de 150 un nombre à peu près égal d'individus. Il est possible d'obtenir une approximation de la médiane à partir de l'histogramme d'une distribution. La perpendiculaire à l'axe horizontal (l'axe de la variable), qui partage l'aire sous l'histogramme en deux parties égales, détermine la médiane. Il y a de chaque côté de cette perpendiculaire un même nombre de valeurs: l'aire étant proportionnelle à la fréquence. La médiane est la valeur lue à l'intersection de cette perpendiculaire avec l'axe horizontal.

Mode

Dans une série de valeurs observées, le *mode (mo)* est la valeur qui revient le plus souvent. C'est la valeur dominante. Pour la série 3 — 5 — 6 — 6 — 7 — 7 — 7 — 7 — 8 — 8 — 9, le mode est égal à 7.

Le mode de la tension artérielle systolique des 121 patients est égal à 151 *(mo* = 151 mmHg).

Si on se réfère à un tableau de fréquences, la *classe modale* est celle ayant la fréquence la plus élevée. Au tableau 3-2, la classe modale est 150-159 mmHg.

Propriétés et comparaisons des quatre mesures

INFLUENCE DES VALEURS EXTRÊMES

La moyenne arithmétique d'un ensemble de données (série statistique) dépend de toutes les observations qui le composent. Elle est influencée par les valeurs extrêmes qui peuvent s'y trouver qu'elles soient élevées ou basses. La moyenne arithmétique est sensible à la valeur extrême 16 présente dans la série 3 — 4 — 5 — 16. L'influence des valeurs extrêmes est bien sûr atténuée si le nombre d'observations est grand.

À l'instar de la moyenne arithmétique, la moyenne géométrique dépend de toutes les observations et subit l'influence des valeurs extrêmes, mais avec une intensité différente. Une valeur extrême élevée influe moins sur la moyenne géométrique que sur la moyenne arithmétique. C'est l'inverse qui se produit avec une valeur extrême basse. Le tableau 3-5 illustre ces phénomènes.

Dans un domaine comme la pollution, où l'on peut rencontrer des concentrations de particules à des valeurs extrêmement élevées, les hygiénistes industriels utilisent souvent la concentration moyenne géométrique de particules plutôt que la concentration moyenne arithmétique, qui est moins centrale.

La moyenne arithmétique est cependant plus utilisée, plus facile à comprendre et possède la belle propriété que si on ajoute une même valeur à chaque observation d'une série statistique, la moyenne arithmétique change par la même valeur. Quand on ajoute 2 à chacune des quatre valeurs de la

Tableau 3-5

Série	M_a	m_g
3—4—5	4	3,91
3—4—5—16	7	5,57
14 — 15 — 16	15	14,98
3 — 14 — 15 — 16	12	10,02

série statistique 3 — 4 — 5 — 8, la moyenne arithmétique se trouve être augmentée de la même valeur 2. Il en va autrement pour la moyenne géométrique. Le tableau 3-6 illustre ces faits.

La médiane dépend avant tout du rang des observations et non de leurs valeurs. Elle est invariante à une augmentation d'une valeur qui lui est supérieure ou à une diminution d'une valeur qui lui est inférieure. Que l'on remplace 8 par 14 ou 3 par 2 dans la série 3 — 4 — 5 — 8 ne change pas la valeur de la médiane. On peut déduire de cette invariance, par exemple, que le calcul du temps de survie médian d'un groupe de patients est possible dès que la moitié des patients du groupe sont décédés.

La médiane n'est pas influencée par les valeurs extrêmes contrairement à la moyenne arithmétique. De ce fait, on pourra préférer la médiane pour décrire la tendance centrale d'une distribution fortement asymétrique.

Le mode est plutôt influencé par les fréquences des observations que par leur valeur ou leur rang.

STABILITÉ

La moyenne arithmétique est moins sensible que la médiane et le mode aux fluctuations d'échantillonnage. Ainsi, pour des séries différentes d'observations d'une même variable se rapportant à une même population d'individus, il y a, en règle générale, moins de variation entre les moyennes de chacune des séries qu'entre leurs médianes ou leurs modes. Le mode est, de ces trois mesures, la plus sensible aux changements et, un peu comme la mode, il se démode. Le mode est moins stable que la médiane ou la moyenne, surtout si l'on dispose de peu d'observations. La plus grande stabilité de la moyenne arithmétique joue en sa faveur. Au plan des fluctuations, la moyenne arithmétique est plus stable que la médiane ou le mode.

TYPE D'ÉCHELLE ET MESURE

Suivant le type d'échelle de classification, certaines mesures de tendance centrale pourront être calculées alors que d'autres ne le pourront pas. On ne peut calculer les moyennes arithmétique ou géométrique que pour des observations de variables quantitatives. Le calcul des moyennes est possible seulement si l'on dispose d'échelles par intervalle ou proportionnelles. La médiane, qui suppose une mise en ordre des observations d'une variable, exige au moins une échelle ordinale. Sur une échelle nominale, on ne peut pas calculer la médiane. Quant au mode, on peut le déterminer aussi bien pour des variables qualitatives que pour des variables quantitatives. Il s'agit seulement d'identifier la classe ou la valeur la plus

Tableau 3-6

Série	m_a	m_g
3 — 4 — 5 — 8	5	4,7
5 — 6 — 7 — 10	7	6,8

Tableau 3-7

Échelle	Mesure centrale disponible
nominale	mode
ordinale	mode, médiane
par intervalle	mode, médiane, moyennes
proportionnelle	mode, médiane, moyennes

fréquente. Le mode est donc disponible sur les quatre échelles. Pour les quatre types d'échelles, le tableau 3-7 indique les mesures de tendance centrale qui sont naturellement calculables.

Position relative des quatre mesures

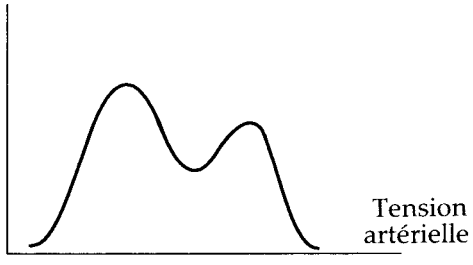
Il faut d'abord souligner qu'une distribution de fréquences peut avoir deux ou plusieurs modes. La présence de plusieurs modes peut suggérer que le groupe d'individus considéré est non-homogène. Par exemple, une distribution bimodale de la tension, comme à la figure 3-6, peut indiquer que nous sommes en présence de deux groupes de personnes. Soulignons au passage que l'interprétation de la moyenne ou de la médiane fait problème dans le cas de distributions bimodales ou multimodales.

Nous limiterons notre discussion sur la position relative des quatre mesures de tendance centrale aux distributions de fréquences unimodales, c'est-à-dire ayant un seul mode.

POSITION RELATIVE DES MOYENNES GÉOMÉTRIQUE ET ARITHMÉTIQUE

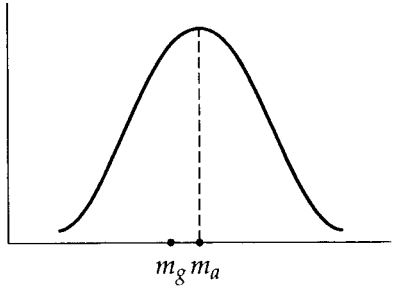
En supposant les valeurs x_i observées positives, on peut démontrer que $m_g \leq m_a$.

Figure 3-6



La valeur d'une moyenne géométrique ne peut pas être plus grande que celle de la moyenne arithmétique sur une même série de valeurs. De plus, elle lui est égale seulement si toutes les valeurs observées x_i sont égales entre elles. Pour une distribution unimodale, peu importe qu'elle soit symétrique, fortement ou légèrement asymétrique, à gauche ou à droite, la position de la valeur moyenne géométrique sur l'axe horizontal est toujours à gauche de celle de la valeur de la moyenne arithmétique. Les deux positions coïncident si les observations sont toutes égales à une même valeur. Nous illustrons leur position relative à la figure 3-7 pour une distribution unimodale symétrique.

Figure 3-7



POSITION RELATIVE DE LA MOYENNE ARITHMÉTIQUE, DE LA MÉDIANE ET DU MODE

Si une distribution unimodale est symétrique, comme à la figure 3-8, les valeurs des trois mesures de tendance centrale sont identiques. La moyenne arithmétique est égale à la médiane et au mode.

Le mode correspond au sommet le plus élevé. La médiane partage les fréquences en deux parties égales, donc la surface aussi, puisqu'il faut se rappeler que l'aire sous une courbe de distribution mesure les fréquences.

Si une distribution est asymétrique à droite, la moyenne arithmétique, qui subit davantage l'influence des valeurs extrêmes que la médiane, est déplacée vers la droite plus fortement que celle-ci. La médiane qui partage la surface en deux parties égales est forcément plus grande que le mode. On décrit à la figure 3-9 la position relative du mode, de la médiane et de la moyenne arithmétique.

Si la distribution est asymétrique à gauche, la situation est inversée comme à la figure 3-10.

Figure 3-8

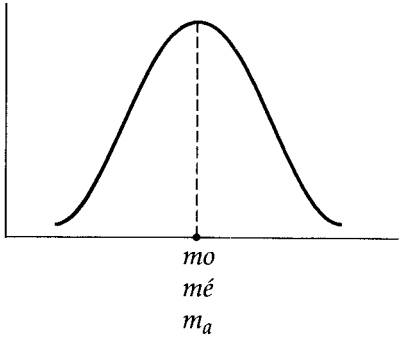
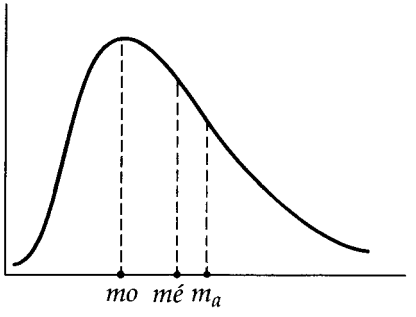


Figure 3-9



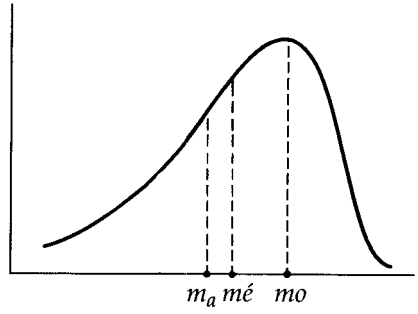
Pour terminer, que la dissymétrie soit à gauche ou à droite, la médiane est plus centrale que la moyenne arithmétique. Plus une dissymétrie est forte, plus la médiane est favorisée, en comparaison de la moyenne arithmétique, pour décrire le centre d'une distribution. Pour une distribution légèrement asymétrique, il existe une règle (empirique) entre le mode, la médiane et la moyenne arithmétique :

3(moyenne – médiane) $\simeq$ moyenne – mode.

MESURES DE DISPERSION

De toute évidence, les mesures de tendance centrale sont, en termes géométriques ou graphiques, des mesures de position. Elles permettent de localiser, à des degrés divers, le centre d'une distribution. Si la moyenne arithmétique des 121 patients avait été de 159,9 mmHg plutôt que 149,9, la position de l'histogramme à la figure 3-1 aurait été différente. Deux courbes de distribution se rapportant à des distributions de fréquences ayant une même

Figure 3-10

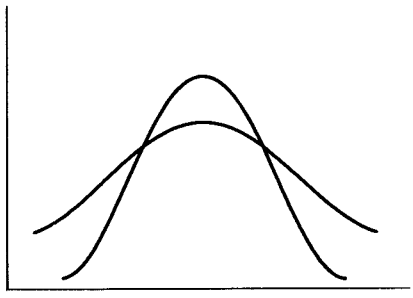


valeur centrale ont une même localisation, comme à la figure 3-11.

Ces deux courbes se distinguent cependant par l'étalement de leurs valeurs. Les mesures de tendance centrale ne suffisent pas à caractériser complètement une distribution de fréquences. Deux groupes de nouveau-nés peuvent avoir le même poids moyen à la naissance avec une dispersion différente de leurs poids. Que la tension artérielle des 121 patients soit de 149,9 mmHg ne nous apprend rien sur la dispersion des 121 valeurs. Ni la moyenne, ni la médiane, ni le mode en soi nous éclairent sur la dispersion d'une distribution de fréquences, sur l'étalement des valeurs dans une série d'observations. C'est pourquoi, il est essentiel de définir des *mesures de dispersion* qui nous renseignent sur la variabilité des observations, un concept fondamental en statistique.

Il existe un grand nombre de mesures de dispersion. Nous en retenons d'abord deux: la plus simple, l'étendue et, la plus connue, la variance. Nous présentons ensuite le coefficient de variation et l'intervalle semi-interquartile.

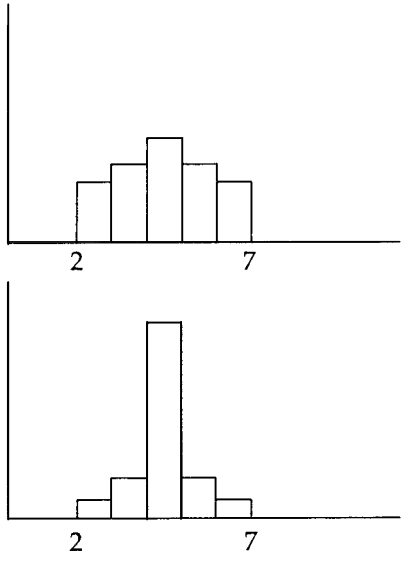
Figure 3-11



Étendue

L'étendue, rappelons-le, est la différence entre la plus grande et la plus petite des valeurs observées. Pour les 121 patients de plus de 50 ans, l'étendue des valeurs de leur tension artérielle est égale à 198 – 122, soit 76 mmHg. Cette mesure de dispersion a le défaut de ne tenir compte que de deux valeurs, la plus petite et la plus grande, donc d'ignorer les autres observations et par conséquent leurs fréquences. Ce défaut est accentué si, de surcroît, la plus petite et la plus grande sont des valeurs extrêmes, voire des valeurs aberrantes. En conséquence, l'étendue néglige une partie importante de l'information. Les deux histogrammes décrits à la figure 3-12 ont trait à des distributions de même étendue (7 – 2 = 5). Pourtant le premier réfère manifestement à une distribution aux valeurs plus dispersées (moins concentrées).

Figure 3-12



Variance (et écart-type)

Une bonne mesure de dispersion doit refléter la manière dont toutes les observations s'écartent d'une valeur centrale. La variance, comme nous allons le voir, mesure la dispersion autour de la moyenne arithmétique. Plus difficile à comprendre que l'étendue, nous allons l'expliquer graduellement.

Considérons la série de valeurs 130 — 131 — 131 — 133 — 140 où la movenne est de:

$$m_a = \frac{130 + 131 + 131 + 133 + 140}{5} = 133.$$

Chacune des valeurs observées s'écarte plus ou moins de la valeur moyenne arithmétique, certaines valeurs lui sont supérieures, d'autres inférieures ou encore égales. La différence (valeur observée – valeur moyenne) est naturellement la plus simple des mesures de l'écart entre une valeur observée et la moyenne. Ainsi, en suivant l'ordre des valeurs de cette série, les écarts, tels que définis, sont respectivement de -3, -2, -2, 0, + 7.

Le signe indique de quel côté par rapport à la moyenne se situe la valeur observée. On aurait, pour la même variable, une série plus dispersée 113 — 127 — 132 — 133 — 160 de même moyenne (m_α = 133) que, dans leur ensemble, les écarts seraient plus importants comme : -20, – 6, – 1, 0, + 27.

Il est raisonnable de penser qu'une synthèse appropriée des cinq différents écarts individuels obtenus puisse conduire à mesurer la dispersion globale de la série. Une première synthèse qui peut venir à l'esprit est la moyenne des cinq écarts. Cette synthèse définit une mesure de dispersion qui ne l'est qu'en apparence, puisque la moyenne des écarts à la moyenne est toujours égale à zéro, quelle que soit la dispersion de la distribution considérée. On le constate aisément pour les deux séries précédentes, où

$$\frac{(-3) + (-2) + (-2) + 0 + 7}{5} = 0$$

$$\frac{(-20) + (-6) + (-1) + 0 + 27}{5} = 0$$

Ce résultat est général et facile à démontrer. En effet, il y a une sorte de compensation entre les écarts positifs et les écarts négatifs et, en définitive, la somme s'annule.

On peut contourner cette difficulté en ignorant le signe des écarts individuels, c'est-à-dire en ne retenant que leur valeur absolue. On obtient ainsi une mesure de dispersion qui est la moyenne des écarts absolus: c'est l'écart *moyen absolu.* Pour les deux mêmes séries précédentes, les écarts moyens absolus sont respectivement:

$$\frac{3 + 2 + 2 + 0 + 7}{5} = 2{,}8$$

et

$$\frac{20 + 6 + 1 + 0 + 27}{5} = 10{,}8.$$

La plus dispersée des deux a le plus grand écart moyen absolu, comme il se doit d'une bonne mesure de dispersion quand elle est appliquée à une même variable, exprimée dans les mêmes unités. L'écart moyen absolu est peu utilisé pour des raisons liées aux valeurs absolues. On contourne la difficulté des valeurs absolues en élevant les écarts au carré ce qui, d'une autre façon, rend positifs les écarts négatifs. Chaque écart x — m_α est élevé au carré. Les écarts au carré sont ensuite additionnés pour donner finalement la mesure de dispersion décrite

par l'expression suivante:

$$\frac{\Sigma (x - m_a)^2}{n},$$

où n est le nombre d'observations. Cette mesure de dispersion, dénotée S^2, est appelée la *variance.*

$$s^2 = \frac{\Sigma (x - m_a)^2}{n}.$$

Toujours pour les deux mêmes séries ci-dessus, les variances sont respectivement:

$$s^2 = \frac{3^2 + 2^2 + 2^2 + 0^2 + 7^2}{5} = 13,2$$

$$s^2 = \frac{20^2 + 6^2 + 1^2 + 0^2 + 27^2}{5} = 233,2.$$

ce qui montre bien, pour une même variable exprimée dans les mêmes unités, que la série la plus dispersée a la plus grande variance. La variance apparaît comme une mesure de dispersion intéressante pour quantifier la variabilité, caractère inhérent à toute variable. Elle est nulle lorsque toutes les valeurs observées sont égales. La variance s'exprime en unités carrées, ce qui peut parfois être ennuyeux. Si la variable x est l'âge, exprimé en années, la variance S^2 est en années carrées. On peut supprimer cet inconvénient en prenant la racine carrée de la variance, qu'on appelle *l'écart-type* et que l'on désigne par s.

$$s = \sqrt{\frac{\Sigma (x - m_a)^2}{n}}.$$

Pour les deux séries déjà considérées,

$$s = \sqrt{13,2} = 3,6$$

$$s = \sqrt{233,2} = 15,3.$$

S'il est difficile de donner un sens concret à la variance, ou à l'écart-type, on peut se rappeler qu'ils permettent de comparer, du point de vue de leur dispersion, plusieurs distributions d'une même variable exprimées dans les mêmes unités. L'écart-type de la tension artérielle des 121 patients est de 10,9 mmHg; s'il avait été de 22 mmHg, les tensions artérielles auraient été environ deux fois plus dispersées autour de la moyenne 149,9 mmHg.

Dans un souci de continuité et de simplicité, on a utilisé pour la variance (ou l'écart-type) le diviseur n. Toutefois, pour des raisons propres à la théorie statistique de l'estimation, le diviseur doit être plutôt $(n - 1)$ lorsqu'il s'agit de calculer, d'estimer la variance d'une série de valeurs observées, lesquelles ne sont qu'une partie de toutes les valeurs possibles. La justification pour ce diviseur est quelque peu difficile à présenter ici. On se contentera d'une explication facile. Les valeurs extrêmes d'une variable, plus rares que les valeurs plus centrales, ont moins de chance de se retrouver parmi les *n* valeurs observées de la variable. Ce fait probable donne naissance à une série de *n* valeurs observées en moyenne moins dispersée que la totalité des valeurs possibles d'une variable. Il découle qu'en principe la variance calculée à partir de *n* valeurs observées est vraisemblablement inférieure à celle que l'on devrait normalement obtenir. On corrige cette sous-estimation en utilisant un diviseur

plus petit; on démontre qu'il est *(n* — 1). La division par *(n* — 1) conserve à la variance (et à l'écart-type) son caractère de mesure de dispersion. L'essentiel est préservé. On aura plutôt les expressions suivantes:

$$s^2 = \frac{\Sigma (x - m_a)^2}{n - 1}$$

$$s = \sqrt{\frac{\Sigma (x - m_a)^2}{n - 1}}$$

Coefficient de variation

L'écart-type n'est pas un nombre pur. Il porte les unités de mesure de la variable. Si la variable est le rythme cardiaque mesuré en pulsations par minute, l'écart-type s'exprime en pulsations par minute. Cela présente de sérieuses difficultés lorsqu'il s'agit de comparer les distributions de variables différentes, du point de vue de leur dispersion. Comment comparer, par exemple, les dispersions respectives de la tension artériel-le en mmHg et le taux de cholestérol total en mg/100 ml de plasma? Il y a aussi des difficultés lorsqu'il s'agit de comparer les écarts-types de deux distributions d'une même variable mais de tendances centrales différentes. Comment comparer les dispersions des tensions artérielles de deux groupes dont l'un a une tension moyenne de 130 mmHg, l'autre de 150 mmHg?

Pour faciliter les comparaisons évoquées, on définit une mesure relative de dispersion qu'on appelle *coefficient de variation* (C.V.). C'est le rapport de l'écart-type sur la moyen-ne arithmétique.

$$C.V. = s/m_a$$

ou, si on le donne en pourcentage,

$$C.V. = 100\ s/m_a.$$

Ce coefficient de variation est sans dimension: les unités de mesure présentes dans S et dans m_a se simplifient pour finalement disparaître. Le coefficient de variation est donc un nombre pur qui facilite la comparaison de dispersions.

Si, pour un groupe d'individus, la tension artérielle systolique moyenne est de 120 mmHg avec un écart-type de 10 mmHg et le taux moyen de cholestérol total de 180 mg/100 ml de plasma avec un écart-type de 30 mg/ 100 ml de plasma, nous aurons :

$$C.V._{\text{tension}} = 100 \times {}^{10}\!/_{120} = 8{,}3\ \%$$

$$C.V._{\text{cholestérol}} = 100 \times {}^{30}\!/_{180} = 16{,}7\ \%.$$

Le taux de cholestérol total a une dispersion environ deux fois plus grande que celle de la tension artérielle systolique. Les individus sont relativement plus homogènes du point de vue de leur tension artérielle que de leur cholestérol total.

Si des jeunes étudiants ont une tension artérielle systolique moyenne de 124 mmHg avec un écart-type de 6 mmHg, et des jeunes ouvriers une tension artérielle moyenne de 128 mmHg avec un écart type de 9, alors

$$C.V._{\text{(tension étudiants)}} = 100 \times {}^{6}\!/_{124} = 4{,}8\ \%$$

$$C.V._{\text{(tension ouvriers)}} = 100 \times {}^{9}\!/_{128} = 7{,}0\ \%.$$

Les ouvriers sont relativement moins homogènes que les étudiants quant à la tension artérielle systolique.

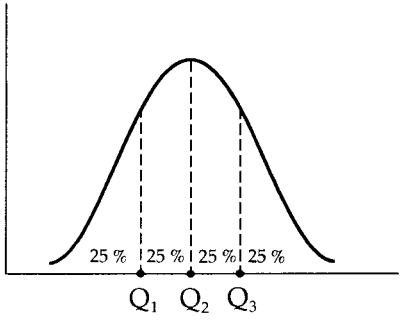
Intervalle semi-interquartile

La variance ou l'écart-type est une mesure de la dispersion des observations autour de leur moyenne arithmétique. La variance est donc aussi influencée par les valeurs extrêmes. De ce fait, elle est moins appropriée à décrire la dispersion des distributions fortement dissymétriques. On pourra lui préférer alors une mesure de dispersion moins affectée par les valeurs extrêmes, en privilégiant par exemple la dispersion autour de la médiane.

La médiane découpe l'intervalle de variation d'une variable en deux parties de même fréquence : 50 % de part et d'autre. On peut bien sûr penser à d'autres découpages, par exemple en quatre morceaux de même effectif : 25 % chacun, comme à la figure 3-13. Les trois points qui séparent ces quatre parties sont nommés dans l'ordre, le *premier quartile* Q_1, le *deuxième quartile* Q_2 et le *troisième quartile* Q_3. Le deuxième quartile n'est autre que la médiane.

Leur calcul, comme pour la médiane, nécessite que les données brutes soient ordonnées. Pour la série ordonnée relative à la tension artérielle systolique des 121 patients (vue précédemment), les trois quartiles sont approximativement 143, 150 et 156 mmHg.

Figure 3-13



La différence $Q_3 - Q_1$ est l'intervalle interquartile. La demi-différence de Q_3 et Q_1, que l'on désigne par Q, est appelé l'intervalle *semi-interquartile.* On écrit:

$$Q = \frac{Q_3 - Q_1}{2}$$

L'intervalle semi-interquartile Q est bien une mesure de dispersion, puisque plus les valeurs observées sont concentrées, plus les quartiles Q_1 et Q_3 sont rapprochés, finalement plus Q est petit. Dans notre exemple sur la tension artérielle, Q est de l'ordre de

$$6{,}5\ \text{mmHg, soit :}\ \frac{156 - 143}{2}.$$

Moins utilisé que la variance (ou l'écart-type), l'intervalle semi-interquartile est intéressant lorsqu'il s'agit de distributions fortement dissymétriques.

AUTRES MESURES DE POSITION: LES CENTILES

L'intervalle de variation d'une variable peut être découpé encore davantage qu'il ne le fut pour la définition de l'intervalle semi-interquartile. En cinq morceaux de 20 % chacun, nous avons quatre points de séparation nommés dans l'ordre premier, deuxième, troisième et quatrième quintile. En dix morceaux de 10 % chacun, nous avons neuf points de séparation. Ce sont les déciles qui vont du premier au neuvième. Enfin, en cent morceaux de 1 % chacun, on compte 99 points de séparation appelés les centiles; ils vont du premier au quatre-vingt-dix-neuvième.

Plus concrètement, considérons un groupe de garçons du même âge dont on a mesuré la taille. Rangeons les garçons en ordre croissant, du plus petit au plus grand. Il peut s'avérer intéressant de connaître la valeur de la taille par rapport à laquelle 5 % des garçons en ont une qui lui est inférieure et 95 % une qui lui est supérieure. Cette valeur est appelée le cinquième centile. Elle partage les garçons en deux groupes : l'un qui contient les 5 % plus petits, l'autre 95 % des plus grands. Si le cinquième centile de la distribution des tailles d'un groupe de garçons est de 130 cm, cela signifie que 5 % des garçons ont une taille inférieure à 130 cm et 95 % une taille supérieure.

Sur un groupe de cent garçons, les cinquième et sixième garçons ont les tailles les plus rapprochées du cinquième centile. De part et d'autre du cinquième garçon, on trouve 4 et 95 garçons, soit respectivement 4 % et 95 %; quant au sixième garçon, on en trouve de chaque côté 5 et 94. Le cinquième centile se situe donc entre la taille du cinquième et celle du sixième garçon. En première approximation, on peut dire, pour plus de simplicité, que le cinquième centile est donné par la taille du cinquième garçon.

Pour déterminer les centiles des valeurs observées d'une variable, il suffit d'abord de les mettre en ordre, ensuite de repérer parmi ces données celle qui vérifie le mieux la définition du centile désiré. Par exemple, pour la tension artérielle dont les valeurs ordonnées ont été présentées plus haut, la valeur qui se rapproche le plus du dixième centile est située entre 136 et 137 mmHg.

RÉSUMÉ

Le tableau de fréquences est un mode de présentation qui regroupe, synthétise, suivant certaines classes, l'ensemble des valeurs observées d'une variable. Pour une variable quantitative, la synthèse opérée par ce regroupement peut être poursuivie en réduisant l'ensemble des valeurs observées à quelques valeurs typiques de tendance centrale et de dispersion. La moyenne arithmétique et la variance (écart-type) figurent parmi les mesures les plus utilisées respectivement de tendance centrale et de

Symboles

m_a: moyenne arithmétique

m_g: moyenne géométrique

$mé$: médiane

mo: mode

Q_1, Q_2, Q_3: premier, deuxième, troisième quartile

s^2, s: variance, écart-type

$C.V.$: coefficient de variation

Q: intervalle semi-interquartile.

Formules

$$m_a = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\Sigma\, x_i}{n}$$

$$m_g = (x_1 \times x_2 \times \ldots \times x_n)^{1/n}$$

$$s^2 = \frac{\Sigma\, (x_i - m_a)^2}{n} \text{ ou } \frac{\Sigma\, (x_i - m_a)^2}{n - 1}$$

$$C.V. = 100\ s/m_a$$

$$Q = \frac{Q_3 - Q_1}{2}$$

LECTURE SUGGÉRÉE

1. SCHERRER B. ***Biostatistique,*** Chicoutimi, Gaétan Morin Éditeur, 1984, deuxième partie, pp.99-190.

ANNEXE DU CHAPITRE 3

Calcul de mesures de tendance centrale et de dispersion sur des données regroupées

Commençons par un avertissement. Le regroupement des données est une opération superflue s'il s'agit de calculer des valeurs de tendance centrale ou de dispersion alors que les données sont peu nombreuses ou traitées par ordinateur. Hormis la volonté de ne pas alourdir inutilement le texte, cela explique que les calculs aient été reportés en annexe. Pour le bénéfice du lecteur *non-informatisé,* nous allons illustrer les calculs sur données regroupées en utilisant celles du tableau 3-2.

Le regroupement des données en classes comporte une perte d'information manifeste. En effet, on sait par exemple qu'il y a 16 patients dont la tension est comprise entre 130 et 139 mmHg, mais on ignore la valeur particulière de tension pour chacun d'eux. Pour remédier à cette perte d'information, on va supposer pour chaque classe, faute de mieux, que les observations s'y répartissent uniformément. Cette supposition, dite de répartition uniforme, est la plus raisonnable. Elle signifie que la différence entre deux valeurs observées adjacentes est partout constante dans une même classe.

La supposition de répartition uniforme implique que le produit de la multiplication de la valeur médiane d'une classe par la fréquence de classe est égale à la somme des valeurs (uniformément réparties) qui tombent dans la classe. Par exemple, la somme 135 des valeurs uniformément réparties (41 — 45 — 49) est égale à 45 x 3, soit le point milieu multiplié par le nombre d'observations.

Calcul de la moyenne arithmétique et de la variance

Sur cette base de la répartition uniforme, nous remplaçons le tableau 3-2 par le tableau A3-1, qui servira au calcul de la moyenne m_a et de la variance S^2 (l'écart-type S).

On obtient:

$$m_a = \frac{\Sigma n_i x_i}{n} = \frac{18\,154,5}{121} = 150,0 \text{ mmHg}$$

$$s^2 = \frac{\Sigma n_i (x_i - m_a)^2}{n - 1} = \frac{14\,590,25}{120} = 121,6 \text{ mmHg}^2$$

$$s = 11,0 \text{ mmHg}$$

Tableau A3-1

x_i (point milieu*)	n_i (fréquence)	$n_i x_i$	$x_i - m_a$	$(x_i - ma)^2$	$n_i (x_i - m_a)^2$
124,5	2	249	-25,5	650,25	1 300,50
134,5	16	2 152	-15,5	240,25	3 844,00
144,5	41	5 924,5	-5,5	30,25	1 240,25
154,5	44	6 798	4,5	20,25	891,00
164,5	14	2 303	14,5	210,25	2 943,50
174,5	2	349	24,5	600,25	1200,50
184,5	1	184,5	34,5	1190,25	1190,25
194,5	1	194,5	44,5	1980,25	1980,25
	121	**18154,5**			**14 590,25**

* La première classe (120-129) commence réellement à 119,5 pour se terminer à 129,5, puisque la tension est connue au mmHg près. Le point milieu est alors 124,5. II en va de même pour les autres classes.

On note une différence, bien que très faible ici, entre la moyenne calculée en utilisant les données brutes (149,9) et celle obtenue à partir du tableau de fréquences (150,0). Cette différence résulte du fait que la supposition de répartition uniforme n'est pas toujours correcte. La moyenne calculée à partir des données regroupées est une approximation de celle évaluée en utilisant les valeurs brutes. L'approximation est souvent acceptable et elle est généralement d'autant meilleure que le nombre de valeurs observées est élevé et que l'intervalle de classe est petit. La même remarque s'applique à l'écart-type.

Les calculs de m_a et de s auraient été impossibles s'il y avait eu des classes ouvertes corn- me 190 et plus.

Calcul de la médiane

Le calcul de la médiane à partir de données regroupées passe par les trois étapes suivantes:

1. la détermination du rang de la médiane;
2. l'identification de la classe qui contient la médiane;
3. le calcul proprement dit de la médiane.

Tableau A3-2

Tension artérielle systolique (mmHg)	Nombre cumulé de patients
120-129	2
130-139	18
140-149	59
150-159	103
160-169	117
170-179	119
180-189	120
190-199	121

Avant de franchir ces trois étapes, nous allons refaire le tableau 3-2 en remplaçant la colonne du nombre de patients par celle du nombre cumulé de patients. On obtient le tableau A3-2 communément appelé tableau des fréquences cumulées.

La tension artérielle médiane est de rang 61, soit $\frac{n+1}{2}$, en l'occurrence $\frac{121+1}{2}$. La 61e valeur se trouve dans la classe 150-159 mmHg, que l'on peut appeler la classe médiane. Les deux premières étapes complétées, nous savons maintenant que la médiane est au moins égale à 149,5 qui est la limite réelle inférieure de la classe médiane. Le 59e patient a une tension au plus égale à 149,5. Quelle est celle du 61e patient? Elle est égale à 149,5 plus une certaine quantité que nous allons chercher à connaître.

Dans la classe 150-159 (149,5 à 159,5), il y a 44 observations pour un intervalle de 10 mmHg. Sur la base de la répartition uniforme, la distance entre chaque observation dans la classe médiane est constante et égale à 10/44 II faut compter deux observations dans la classe médiane pour se rendre à la médiane, puisque nous devons passer de la 59e à la 61e (61-59). La quantité à ajouter à 149,5 pour obtenir une approximation de la médiane est:

$2 \times {}^{10}/_{44}$.
On obtient:

$$mé = 149{,}5 + \frac{2}{44} \times 10$$

$$= 150{,}0 \text{ mmHg}$$

Le calcul ci-dessus peut être formalisé si l'on introduit les notations suivantes:

— 149,5 est la limite inférieure de la classe médiane: $l_{mé}$

— 2 est la différence entre 61 et 59. Dans cette différence, 61 est le rang de la médiane: $\frac{n+1}{2}$, où n est le nombre total d'observations; 59 est la fréquence cumulée de la classe qui précède celle de la médiane: f_C.

— 44 est la fréquence de la classe médiane: $f_{mé}$.

— 10 est l'intervalle de la classe médiane: i.

On obtient donc:

$$mé = l_{mé} + \frac{(\frac{n+1}{2} - f_C)}{f_{mé}} \times i$$

On retrouve aussi la formule:

$$mé = l_{mé} + \frac{(n/2 - f_C)}{f_{mé}} \times i,$$

où $n/2$ a remplacé $(n + 1)/2$. La différence est faible si n est grand. Cette dernière forme est facilement généralisable à d'autres centiles, comme les quartiles par exemple:

$$Q_1 = l_{Q_1} + \frac{(n/4 - f_C)}{f_{Q_1}} \times i$$

$$Q_3 = l_{Q_3} + \frac{(3n/4 - f_C)}{f_{Q_3}} \times i$$

Dans l'exemple du tableau A3-2, on trouve:

$$Q_1 = 139{,}5 + \frac{(121\frac{1}{4} - 18)}{41} \times 10$$

$$= 142{,}5$$

et

$$Q_3 = 149{,}5 + \frac{(3 \times 121\frac{1}{4} - 59)}{44} \times 10$$

$$= 156{,}7$$

L'intervalle semi-interquartile Q, calculé sur les données regroupées, est alors:

$$Q = \frac{156{,}7 - 142{,}5}{2}$$

$$= 7{,}1 \text{ mmHg.}$$

CHAPITRE 4

Mesures de fréquence

Le présent chapitre touche les mesures de fréquence employées en épidémiologie et en santé publique. Une distinction est faite d'abord entre les quatre termes proportion, ratio, indice et taux. Des mesures particulières d'usage courant en rapport avec la maladie sont ensuite discutées, comme les mesures de prévalence et d'incidence. Certaines relations entre ces mesures sont examinées et d'autres, relatives à la mortalité, sont présentées. Parmi ces dernières se trouvent les taux brut et spécifique de décès et la létalité. Enfin, le chapitre aborde la question des opérations simples sur les mesures, telles la somme arithmétique et la somme pondérée.

Nous allons définir et discuter les mesures de fréquence les plus courantes en épidémiologie. Dans le contexte de la santé des populations, la fréquence décrit le nombre d'individus qui, par exemple, sont atteints d'une maladie, exposés à un facteur, soumis à un traitement, décédés d'une certaine cause, etc. Le terme « fréquence » a le même sens que celui défini antérieurement et représente le nombre d'observations appartenant à une classe, par exemple celle des cas d'une certaine maladie.

MESURES DE FRÉQUENCE GÉNÉRALES

Un simple énoncé sur la fréquence de la tuberculose qui touche une population peut prendre la forme suivante : « Il y a eu 158 nouveaux cas de tuberculose ». Un tel énoncé a peu d'utilité à moins qu'il ne spécifie dans quelle population ont été observés les cas de tuberculose et quand ils ont été observés. Un énoncé plus spécifique serait: « En 19X7, il y a eu, dans la région de Sanpulie, 158 nouveaux cas de tuberculose ». Cette dernière formulation n'est pas non plus sans inconvénient lorsqu'il s'agit de comparer la région de Sanpulie à celle d'Épidélie. Il ne suffit pas de dire qu'en 19X7, il y a eu 158 nouveaux cas de tuberculose en Sanpulie et, la même année, 22 nouveaux cas de tuberculose en Épidélie. Pour être adéquate, la comparaison exige que les fréquences 158 et 22 soient rapportées aux tailles respectives des populations des deux régions. La comparaison est alors faite à partir d'une fréquence relative, c'est-à-dire d'un rapport dont nécessairement le numérateur est une fréquence. Dans le reste du texte, la *mesure de fréquence* devra être comprise comme un rapport.

De façon générale, les mesures de fréquence permettent de caractériser au plan quantitatif l'occurrence de la maladie, du décès ou d'autres événements relatifs à la santé des populations. Au plan formel, nous distinguerons quatre catégories de mesures de fréquence: les *proportions,* les *ratios,* les *indices* et les *taux.* Ces mesures s'expriment toutes comme le rapport de deux quantités, mais se distinguent par la quantité qui figure au dénominateur.

Il est important de souligner que les mesures de fréquence sont, par convention, exprimées en unité de taille. Par exemple, un rapport de $^{37}/_{9250}$ s'écrira 4 par 1000 ou 40 par 10 000 plutôt que 0,004. Les valeurs 1000 et 10 000 constituent des unités de taille choisies par l'investigateur. De manière générale, si l'unité de taille est *K,* le rapport *N/D* est exprimé comme $\frac{N}{D} \times K$ par *K* de population. Cette façon de faire évite d'une part la manipulation de fractions décimales et d'autre part concrétise mieux la valeur d'un rapport.

Proportion

En Épidélie, sur 80 000 naissances enregistrées au cours d'une année, 38 616 sont de sexe féminin. Le rapport $^{38\,616}/_{80\,000}$ mesure la fréquence relative des naissances féminines dans la population (statique) des 80 000 naissances. Ce rapport est une proportion. Remarquons que les bébés qui composent le numérateur forment un sous-ensemble de ceux du dénominateur. Le rapport *N/ D* est une *proportion* si les *N* individus du numérateur sont compris dans les *D* individus du dénominateur. Une proportion est toujours comprise entre 0 et 1, à moins qu'elle soit

exprimée en unité *K* de taille de population. Par exemple, si elle est exprimée en pourcentage *(K = 100),* la valeur se trouve entre 0 et 100.

Ajoutons comme autres exemples de proportions :

$$\text{Proportion de décès par accident} = \frac{\text{nombre de décès par accident}}{\text{nombre total de décès}}$$

et

$$\text{Proportion de mortinaissances} = \frac{\text{nombre de morts-nés}}{\text{nombre total de naissances}}.$$

Ratio

Dans la population des 80 000 naissances, pour comparer la fréquence des naissances masculines à celle des naissances féminines, on peut établir le rapport suivant:

$$\frac{\text{41 384 naissances masculines}}{\text{38 616 naissances féminines}}$$

Ce rapport, que nous appellerons ratio, permet de dire qu'A chaque naissance féminine correspond 1,07 naissance masculine, ou mieux encore, que pour chaque 100 naissances féminines, il y a 107 naissances masculines. Remarquons que non seulement le numérateur n'est pas compris dans le dénominateur, mais que tous deux réfèrent à des classes mutuellement exclusives. Nous définirons le *ratio* comme le rapport des fréquences de deux classes d'une même variable. Dans le cas où la variable est dichotomique, le ratio est appelé aussi *cote.*

Indice

Pour calculer la fréquence relative des décès maternels par cause puerpérale, il est naturel de

$$\frac{\text{Nombre de décès maternels par cause puerpérale}}{\text{Nombre total de femmes ayant accouché}}.$$

L'évaluation du dénominateur de cette proportion, c'est-à-dire du nombre de femmes ayant accouché, comporte une difficulté majeure. Le nombre de femmes ayant accouché d'un bébé vivant est généralement bien connu, mais ce n'est pas toujours le cas pour celles ayant accouché d'un enfant mort-né. Dans les régions défavorisées, la déclaration d'un accouchement d'un bébé mort-né est liée à la qualité variable des services. Le nombre recensé de tels accouchements peut être sensiblement inférieur à la fréquence réelle.

Faute de ne pouvoir correctement calculer la proportion des décès maternels, on utilise comme mesure le rapport suivant:

$$\frac{\text{Nombre de décès maternels par cause puerpérale}}{\text{Nombre de naissances vivantes}}.$$

Le dénominateur réfère à l'événement « naissance vivante » généralement bien déclaré et est étroitement lié à l'événement « accouchement ». Cette mesure n'est ni une proportion ni un ratio; en effet, le numérateur n'est pas compris dans le dénominateur et les composantes du rapport réfèrent à deux événements distincts: le décès chez les femmes en couches et les naissances vivantes. Nous qualifierons *d'indice un* tel rapport

de fréquences. C'est en quelque sorte une pseudo proportion, en ce sens que ce rapport sert de substitut à une proportion difficile à calculer.

Taux

Considérons une population dynamique dont les individus sont susceptibles d'être affectés par une certaine caractéristique, disons une maladie *M.* À chaque instant, la population est ainsi subdivisée en deux sous-groupes : celui des malades (M_1) et celui des non-malades (M_0). A chaque moment, des individus peuvent passer d'un groupe à l'autre, soit par l'apparition de la maladie, soit par guérison. On ne s'intéressera ici qu'aux transferts du groupe M_0 vers le groupe M_1, c'est-à-dire à l'apparition de la maladie comme l'illustre la figure 4-1.

Le nombre de transferts du groupe M_0 vers le groupe M_1 par unité de temps définit ce que nous nommons le *débit de transfert (ou* de malade).

Le débit dépend de ce qu'on peut appeler une vitesse de transfert et aussi de la taille du groupe M0. Illustrons cette dépendance par une analogie. Au concept du débit de transfert, on peut faire correspondre celui du débit d'un cours d'eau, c'est-à-dire au nombre de mètres cubes d'eau qui y circulent à la minute. Le débit d'un cours d'eau dépend à la fois de la vitesse d'écoulement et de la taille (largeur et profondeur) du cours d'eau. Nous pouvons écrire plus formellement pour le cours d'eau :

Débit = vitesse d'écoulement X taille du cours d'eau

Par analogie, pour la population M_0 nous pouvons écrire :

Débit de transfert = vitesse de transfert x taille de M_0 *[1]*

Le débit de transfert de M_0 vers M_1, c'est-à-dire le débit de malades, est donc lié à la taille de M0 et à la vitesse de transfert que nous allons maintenant définir plus précisément.

Considérons deux groupes de non-malades qui risquent de développer la maladie M. On observe ces deux groupes pendant une même période pour y déterminer le nombre de transferts de l'état de non-malade à celui de malade. Nous supposons que les tailles de ces deux groupes M_0 sont égales et demeurent constantes durant la période d'observation. On remarque pour la période de temps déterminée qu'il y a respectivement trois et six transferts de M_0 vers M_1 comme le décrivent les figures 4-2A et 4-2B.

Figure 4-2A

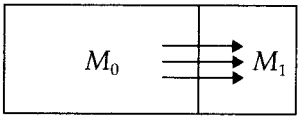


Figure 4-1

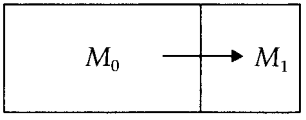
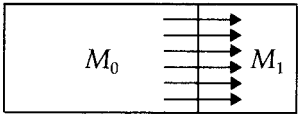


Figure 4-2B



Chaque groupe M_0 est en définitive un réservoir d'un même nombre de sujets qui risquent de développer la maladie. On note par ailleurs une différence entre les deux figures. Le débit de transfert, c'est-à-dire le nombre de transferts de M_0 vers M_1 par unité de temps, est plus élevé à la figure 4-2B qu'à la figure 4-2A. Cette différence entre les débits de transfert ne peut pas être expliquée par une différence des tailles de ces deux groupes M_0 puisque celles-ci sont égales. Comment alors expliquer cette différence si ce n'est que par un transfert plus rapide à la figure 4-2B? Il y a différence dans les *vitesses de transfert,* c'est-à-dire dans la rapidité avec laquelle les réservoirs M_0 s'écoulent vers M_1.

À partir de la relation [1], on peut exprimer la vitesse de transfert en fonction du débit et de la taille de M_0:

$$\text{Vitesse (de transfert)} = \frac{\text{débit (de transfert)}}{M_0 \text{ (taille de)}}$$

$$= \frac{\text{nombre de transferts}/\Delta t}{M_0}$$

$$= \frac{\text{nombre de transferts}}{M_0 \times \Delta t}$$

où Δt représente l'intervalle de temps exprimé en unité de temps (mois, année ...) et M_0 la taille du groupe à risque. (On note que le symbole M_0 désigne, suivant le contexte, ou bien le nom du groupe ou bien le nombre d'individus qui le composent, c'est-à-dire la *taille.* Il en est de même pour M_1.)

Suivant l'usage en épidémiologie, nous emploierons le terme « taux » (de transfert) plutôt que « vitesse » (de transfert).

$$\text{Taux (de transfert)} = \frac{\text{nombre de transferts}}{M_0 \times \Delta t}$$

Toutefois, l'usage ne devrait pas faire oublier qu'il s'agit d'une vitesse de transfert. (Le mot « transfert » est un terme générique. Il peut désigner un décès ou encore un nouveau cas de maladie.)

Dans une population de 300 000 individus risquant de développer une maladie M, sur une période de deux ans, on observe 120 nouveaux cas de transfert (maladie), le taux de transfert correspondant est donné par

$$\frac{120 \text{ cas}}{300\,000 \text{ personnes} \times 2 \text{ ans}}$$

c'est-à-dire 20 cas par 100 000 personnes-années.

Relativement à cette notion de taux, il est opportun d'apporter quelques remarques.

- Dans l'exemple précédent, le débit et le taux ont été calculés sur une période relativement longue, en l'occurrence deux ans: en ce sens, nous obtenons une mesure moyenne. Dans la population, le débit moyen est de 60 nouveaux cas par année et le taux moyen de 20 nouveaux cas par 100 000 personnes par année. Des mesures instantanées, c'est-à-dire calculées sur un très court intervalle de temps (à la limite nul), sont intéressantes au plan théorique mais au plan pratique difficiles, voire impossibles à obtenir. Ainsi, les taux utilisés en épidémiologie sont généralement des mesures moyennes.
- En pratique, si le groupe M_0 compte pour la quasi-totalité de la population, ce qui revient à dire que le groupe M_1 est négligeable par rapport au groupe M_0, alors on peut raisonnablement remplacer la taille M_0 par celle de la population totale. Cette substitution se fait lorsque la population

totale est plus facilement identifiable que le groupe M_0.

- Les épidémiologistes, par tradition, utilisent le terme « taux », soit pour traduire l'idée de transfert, soit pour désigner certaines proportions. De plus, une certaine confusion règne en épidémiologie où ce mot sert à qualifier aussi bien le débit que la vitesse de transfert. Pour plus de clarté et de rigueur, nous réserverons le terme « taux » pour désigner la vitesse de transfert.

CONCEPT DE PERSONNES-TEMPS À RISQUE

Selon la définition d'un taux de transfert, le dénominateur est donné par l'expression $M_0\Delta t$. Le dénominateur d'un taux s'exprime donc en unités complexes : des *personnes-temps à risque,* par exemple des personnes-années à risque. Ces unités, que nous désignerons par le symbole *PT,* comprennent à la fois le nombre de personnes à risque et, pour chacune d'elles, la durée du risque pour la période d'observation. Un individu à risque de la maladie *M,* en observation depuis le 1er juillet et demeuré à risque jusqu'à la fin de l'étude, soit le 31 décembre de la même année, a cumulé six mois ou plus exactement 184 jours de risque. Cette personne compte alors pour 184 personnes-jours à risque. Une autre personne à risque, en observation aussi depuis le 1er juillet mais affectée par la maladie *M* le 1er octobre de la même année, cumule 92 personnes-jours à risque. Notons que la contribution apportée par cette deuxième personne cesse le jour où elle développe la maladie. La contribution des deux personnes totalise 276 personnes-jours à risque.

On peut représenter cette contribution par une surface dans le plan cartésien, comme à la figure 4-3. Le concept de personnes-temps à risque s'interprète bien comme une surface.

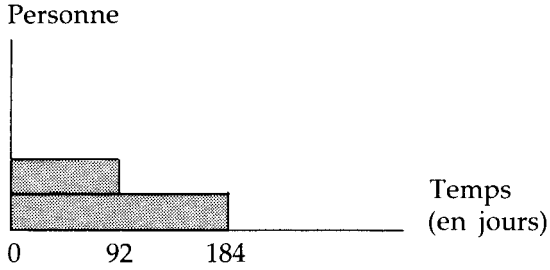
En définitive, pour bien assimiler ce concept étrange de personnes-temps, il suffit de comprendre qu'une personne à risque de maladie M suivie pendant un siècle (Noé par exemple) compte numériquement pour 36 525 personnes-jours à risque ou encore 100 personnes-années à risque.

CALCUL DES PERSONNES-TEMPS À RISQUE DANS UNE POPULATION FERMÉE (OU COHORTE)

CALCUL EXACT

Le calcul du nombre de personnes-temps à risque cumulé par une cohorte est la somme directe du nombre de personnes-temps à risque cumulé par chacun des membres de la cohorte. Par exemple, la cohorte de dix personnes décrite à la figure 4-4 compte exactement pour 152 personnes-semaines à risque, soit dans l'ordre (du haut vers le bas) la somme 20+4+17+20+11+20+6+ 20 + 14 + 20 personnes-semaines.

Figure 4-3

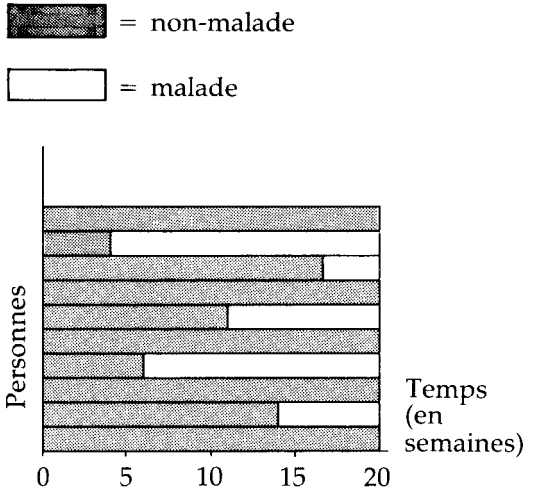


CALCUL APPROXIMATIF

En disposant les dix rectangles de la figure 4-4 suivant la durée du risque, on obtient la figure 4-5.

La surface totale de ces rectangles peut être approximée par celle du trapèze délimité par la ligne discontinue. L'estimation est d'autant meilleure que le débit de nouveaux cas dans la cohorte est constant pour la période. Rappelons que la surface d'un trapèze est obtenue en multipliant la demi-somme des bases par la hauteur. Dans l'exemple, la demi-somme (10 + 5)/2 (personnes) représente une estimation du nombre de personnes à risque au milieu de la période d'observation et la hauteur 20 (semaines) représente la durée de la période. La surface obtenue $\frac{10 + 5}{2} \times 20$ (150 personnes-semaines) est une estimation du nombre exact 152 (personnes-semaines) calculé précédemment.

Figure 4-4



CALCUL DES PERSONNES-TEMPS À RISQUE DANS UNE POPULATION OUVERTE

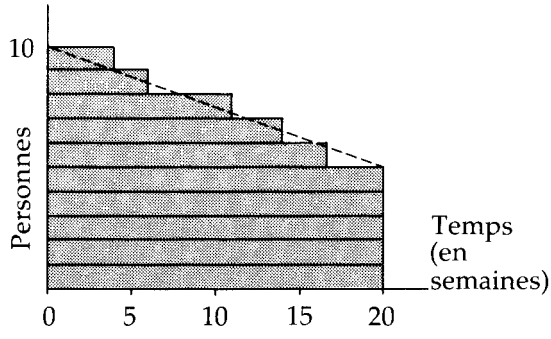
Le calcul du nombre de personnes-temps à risque dans une population ouverte est généralement approximatif puisque, en pratique, il est difficile de déterminer la contribution de chaque individu d'une telle population. Le calcul approximatif se fait comme dans le cas des populations fermées.

Considérons une population ouverte de 500 000 individus à risque d'une maladie, observés pendant une période de deux ans. Supposons que la taille de cette population reste inchangée tout au long de la période d'observation. Le nombre de personnes-années à risque pour cette maladie, cumulé par cette population, est alors de

500 000 personnes x 2 ans = 1000 000 personnes-années.

Considérons encore une population ouverte observée pendant une période de deux ans, mais cette fois dont la taille est instable.

Figure 4-5



De 400 000 qu'elle était au début de la période d'observation, elle passe à 500 000 à la fin de la période. Le nombre cumulé de personnes-années à risque est estimé à

$$\frac{400\,000 + 500\,000}{2} \text{ personnes} \times 2 \text{ ans},$$

soit 900 000 personnes-années.

Rappelons que la valeur numérique de personnes-temps à risque change avec l'unité de temps. Par exemple, 1000 personnes-années à risque a la même valeur, numériquement, que 100 personnes-décades à risque, 12 000 personnes-mois à risque ou encore 52 000 (plus exactement 52 035) personnes-semaines à risque... Bien que les valeurs numériques changent avec le choix de l'unité de temps, la quantité de personnes-temps à risque demeure la même. Il est donc important de bien spécifier l'unité de temps lorsqu'on parle de personnes-temps à risque.

MESURES DE FRÉQUENCE PARTICULIÈRES

Nous allons maintenant aborder des mesures de fréquence particulières à la description de la maladie et du décès dans une population.

Mesures de fréquence de la maladie

Nous distinguons les mesures de prévalence et celles d'incidence. Les premières réfèrent à la description de la fréquence des cas de maladie à un moment donné. Elles s'adressent donc à la description d'un état « être malade » à un moment donné. Les mesures d'incidence se rapportent à la description de la fréquence des nouveaux cas de maladie qui se sont déclarés au cours d'une période de temps déterminée. Elles s'adressent donc à la description d'un événement « devenir malade » dans une période déterminée.

MESURES DE PRÉVALENCE

Considérons une population (statique) dont les individus sont ou affectés par une maladie spécifiée, ou non-affectés mais à risque de l'être. Le nombre P de cas affectés par la maladie à un moment déterminé est appelé le *nombre de cas prévalents ou* la *prévalence* de la maladie. On définit la *prévalence relative (Pr)* d'une maladie dans une population comme le rapport entre la prévalence et le nombre de personnes dans cette population au moment considéré. Si, dans cette population, S désigne le nombre d'individus non-affectés mais qui risquent de l'être, alors:

$$Pr = \frac{P}{S + P}$$

Il faut souligner que la prévalence relative est la proportion des cas prévalents de la maladie dans la population à un moment donné. Comme proportion, cette mesure est un nombre pur, n'a pas d'unité. Par ailleurs, c'est une mesure instantanée en ce sens qu'elle donne cette proportion à un instant donné. Dans le langage courant, le terme prévalence » est utilisé pour désigner aussi bien la prévalence que la prévalence relative. Sans aller à l'encontre de cet usage, nous avons choisi de distinguer dans ce texte les deux termes.

On a recensé au 1er juin d'une année X, 100 cas de diabète dans une population de 5000 individus. La prévalence relative du

diabète dans cette population à la date considérée est alors de:

$$\frac{100 \text{ cas}}{5000 \text{ individus}} = 0{,}02 \text{ (ou 2 \%)}.$$

Au cours d'une enquête portant sur 50 282 naissances, on a diagnostiqué 404 cas de malformation cardio-vasculaire. La prévalence relative à la naissance de malformations cardio-vasculaires dans ce groupe est de :

$$\frac{404 \text{ cas}}{50\,282 \text{ naissances}} = 0{,}008 \text{ (ou 8 par 1000)}.$$

MESURES D'INCIDENCE

Au cours d'une période allant de *t0 à* t1, un individu peut, en regard d'une certaine maladie, passer de l'état de non-malade à celui de malade. Si un individu non-malade le devient en cours de période, il est appelé « cas incident de la maladie » pour cette période. Le nombre I de cas incidents observés dans une population dynamique au cours d'une période est appelé *l'incidence* de la maladie dans cette population. Nous définirons deux mesures relatives à l'incidence, l'une ayant trait à la vitesse de transfert ou de propagation de la maladie, soit le taux d'incidence, l'autre se rapportant au risque d'être affecté par la maladie, soit l'incidence cumulative.

TAUX (OU DENSITÉ) D'INCIDENCE

Par définition, le *taux d'incidence (Ti)* est le rapport de l'incidence I sur le nombre de personnes-temps à risque *(PT)* cumulé par la population en observation.

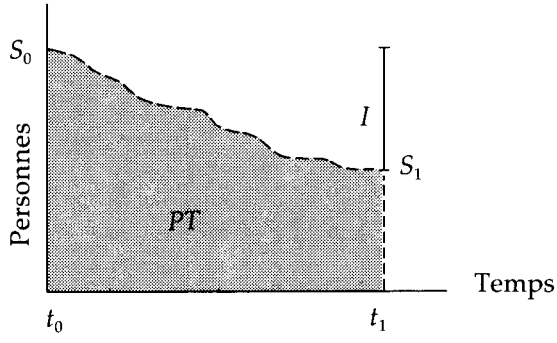
$$Ti = \frac{I}{PT}$$

Le taux d'incidence est une mesure de la vitesse de transfert de l'état de non-malade à celui de malade ou encore de la vitesse de propagation d'une maladie dans une population à risque. Le taux d'incidence est aussi connu sous le nom de *densité d'incidence.*

Considérons la figure 4-6 qui décrit l'évolution de la maladie dans une population fermée. Au temps t_0, la population comprend S_0 individus à risque, alors qu'au moment t_0, qui correspond à la fin de la période, la population à risque n'est plus que de S_l individus. Si l'on admet que les sorties de la population ne sont dues qu'aux cas incidents, l'incidence sur la période est de $S_0 - S_1$, soit la mesure du segment I. La surface tramée *PT* représente le nombre de personnes-temps à risque cumulé. Le taux

d'incidence est le rapport $\frac{I}{PT}$.

Figure 4-6



À la rentrée des classes, sur 1000 écoliers, 100 sont déclarés atteints d'une infection les marquant pour quelques années. Un an plus tard, les 900 écoliers non-affectés sont réexaminés et on décèle cette fois 50 nouveaux cas d'infection. On peut donc estimer à 50 l'incidence pour la période d'un an. Le nombre d'écoliers-années à risque est estimé par le produit du nombre moyen d'écoliers à risque, soit:

$$\frac{900 + 850}{2}$$

par la durée de la période, soit un an. On obtient alors 875 écoliers-années à risque.

$$\frac{900 + 850}{2} \times 1 = 875 \text{ (écoliers-années à risque).}$$

Ce nombre d'écoliers-années obtenu par calcul approximatif correspond à la surface du trapèze de la figure 4-7.

Enfin le taux d'incidence est estimé à:

$$Ti = \frac{50 \text{ nouveaux cas}}{875 \text{ écoliers-années à risque}}$$

$$= 57{,}1 \text{ cas par } 1000 \text{ écoliers-années à risque.}$$

Considérons maintenant la figure 4-8 qui, pour la période de t_0 à t_1, décrit l'évolution de la maladie dans une population ouverte. Chaque point représente un cas incident et la surface *PT* les personnes-temps à risque cumulées. Le taux d'incidence est alors estimé par le rapport du nombre de points à la surface *PT*.

En Sanpulie, on a noté 276 nouveaux cas de tuberculose au cours d'une période de deux ans pour une population estimée à 5200 000 personnes. Le taux d'incidence de la tuberculose en Sanpulie pour la période considérée est donc de

$$Ti = \frac{276 \text{ nouveaux cas}}{5200\,000 \text{ personnes} \times 2 \text{ ans}}$$

$$= 2{,}65 \text{ cas par } 100\,000 \text{ personnes-années.}$$

Figure 4-7

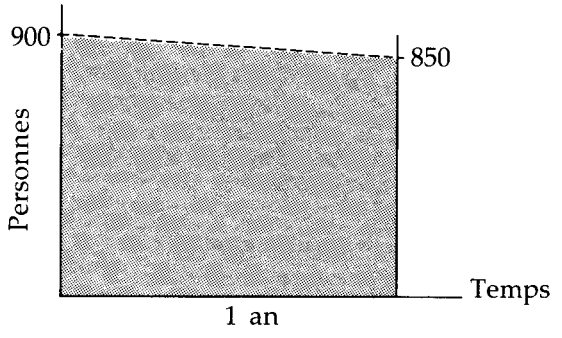
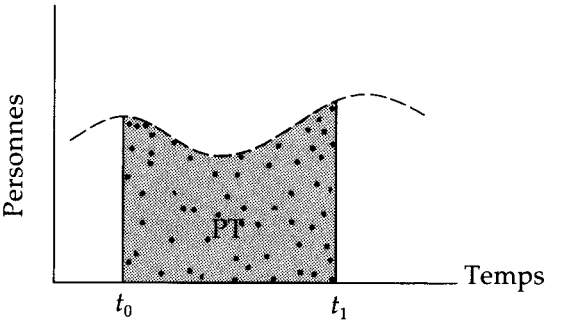


Figure 4-8



INCIDENCE CUMULATIVE

Retournons à la figure 4-6 qui décrit l'évolution de la maladie dans une population fermée pour la période t_0 *à* t_1. *L'incidence cumulative (Ic)* est définie pour cette période par le rapport de l'indicence I sur le nombre S_0 d'individus à risque au début de la période. On peut écrire

L'incidence cumulative est une proportion; son calcul suppose que tous les individus de la cohorte à risque ont été observés pour la période déterminée, c'est-à-dire qu'il n'y a eu de la cohorte à risque aucun retrait autre que ceux attribuables à la maladie considérée. Pour cette raison, dans le reste du texte, nous comprendrons toujours que l'incidence cumulative est une « mesure (probabilité) conditionnelle », dans le sens qu'elle est calculable seulement si cette condition est respectée. En pratique, le calcul direct de l'incidence cumulative satisfait rarement cette condition; il y a toujours de la cohorte des retraits inévitables comme les « perdus-de-vue », les décès, ... Cependant, si l'amoindrissement de la cohorte est faible, l'estimation de l'incidence cumulative peut être tout à fait convenable. Dorénavant, nous comprendrons toujours que l'incidence cumulative est une probabilité conditionnelle (la notion de probabilité est présentée au chapitre 10).

L'incidence cumulative est un nombre pur exprimé le plus souvent en pourcentage. La durée de la période doit toujours être explicitée, par exemple, l'incidence cumulative à un an, deux ans, etc. Sans cette spécification, l'incidence cumulative n'est pas interprétable.

Considérons à nouveau l'exemple des 1000 écoliers. L'incidence cumulative à un an est de 50 nouveaux cas parmi les 900 écoliers à risque du début de la période. On peut donc dire que l'incidence cumulative à un an est de $^{50}/_{900}$ ou 5,6 %.

$$Ic\ (1\ \text{an}) = \frac{50}{900} = 0{,}056\ (\text{ou}\ 5{,}6\ \%)$$

Dans une population de 1500 écoliers chez qui aucune infection n'a été décelée à la rentrée des classes, on détecte, deux ans plus tard, six cas d'infection. Alors l'incidence cumulative à deux ans pour cette infection est de six cas pour 1500 écoliers, soit 0,4 %.

$$Ic\ (2\ \text{ans}) = \frac{6}{1500} = 0{,}004\ (\text{ou}\ 0{,}4\ \%).$$

Dans le langage courant, certains utilisent le terme « incidence » pour désigner le taux d'incidence, d'autres pour désigner l'incidence par unité de temps, d'autres comme synonyme d'incidence cumulative. Quant à nous, nous réservons ce terme pour désigner le nombre de cas incidents. L'incidence cumulative est souvent remplacée par *taux d'attaque,* particulièrement en infectiologie.

RELATIONS ENTRE CERTAINES MESURES

Nous allons examiner maintenant quelques relations entre certaines mesures de fréquence de la maladie, entre la prévalence (P) et l'incidence (I), entre la prévalence relative *(Pr)* et le taux d'incidence (Ti), enfin entre le taux d'incidence *(Ti)* et l'incidence cumulative (*I*c).

RELATION ENTRE LA PRÉVALENCE ET L'INCIDENCE

De façon générale, la prévalence varie comme le produit de l'incidence et de la durée moyenne de la maladie. On écrit:

$$P \sim \frac{I}{\Delta t} \times \bar{d}$$

où P représente la prévalence, $I/\Delta t$ le débit des nouveaux cas et d la durée moyenne de la maladie. Les figures 4-9A et 4-9B illustrent cette dépendance. Dans ces diagrammes, l'axe vertical représente la population à risque et l'axe horizontal, le temps. Chaque segment horizontal représente un cas de maladie dont la durée est mesurée par la longueur du segment compris entre les lignes obliques D et F.

Pour une durée constante (figure 4-9A), une variation du débit ($I/\Delta t$) entraîne une variation dans le même sens pour la prévalence. Pour la période observée, le débit augmente et, de façon équivalente, la prévalence. Par exemple, au moment t_1, elle est de 3 et au moment t_2, elle est de 6. Pour un débit constant (figure 4-9B), une variation de la durée $\bar{d}$ entraîne une variation dans le même sens pour la prévalence (P). Par exemple, au moment t_1 ce nombre est de 4 et au moment t_2, il est de 6.

Pour une maladie fatale donnée, l'amélioration du traitement permettant de reculer l'échéance du décès sans toutefois guérir peut se traduire dans un effet « paradoxal » d'un accroissement de la prévalence de la maladie en raison d'un accroissement de la durée moyenne de la maladie. Ce fut le cas de la tuberculose au début des traitements par les antibiotiques. Une diminution dans la prévalence des malades hospitalisés dans les établissements psychiatriques peut s'expliquer par une réduction de la durée moyenne d'hospitalisation de ces malades.

Si la maladie est en situation d'équilibre dans une population stable, la relation devient :

$$P = \frac{I}{\Delta t} \times \bar{d} \qquad [2]$$

Figure 4-9A

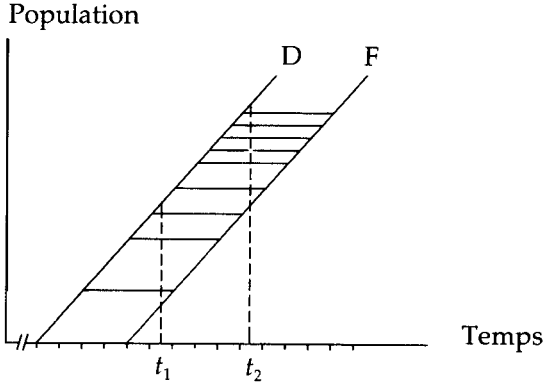
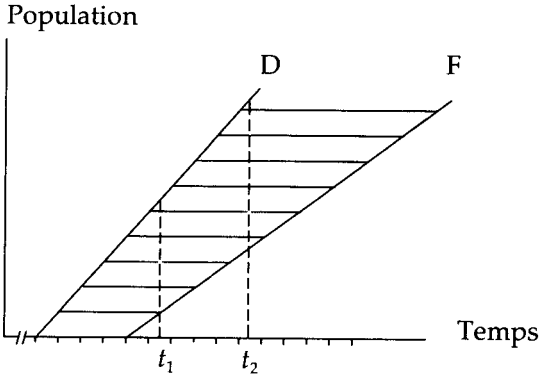


Figure 4-9B



Par maladie en *situation d'équilibre,* on entend que le débit de malades et la distribution de la durée de la maladie sont constants pour la période d'observation. Par *population stable,* on entend une population dont la taille reste constante et sur laquelle la répartition de la maladie suivant ses différents facteurs de risque (âge, sexe, ...) reste la même.

La figure 4-10 nous permet de vérifier la relation [2]. Dans ce diagramme, la maladie est en état d'équilibre: le débit est constant (1 cas par unité de temps), la durée moyenne est constante ($\bar{d} = 4$ unités de temps). Les lignes discontinues permettent de reconnaître les prévalences aux temps spécifiés t_1 et t_2.

On trouve donc:

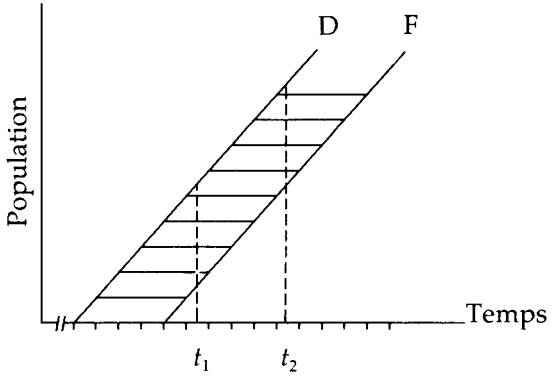
P_1 (prévalence au temps t_1) = 4 cas

P_2 (prévalence au temps t_2) = 4 cas.

Il est facile de vérifier que

$$P_1 = P_2 = \frac{I}{\Delta t} \times \bar{d}.$$

Figure 4-10



RELATION ENTRE LA PRÉVALENCE RELATIVE ET LE TAUX D'INCIDENCE

Pour une maladie en état d'équilibre dans une population stable, on a la relation :

$$P = \frac{I}{\Delta t} \times \bar{d}$$

À partir de cette équation, on peut facilement obtenir par manipulations algébriques une relation entre la prévalence relative *(Pr),* le taux d'incidence *(Ti)* et la durée moyenne $(\bar{d})$ de la maladie. En effet, à partir de la relation citée, on a

$$\frac{P}{S + P} = \frac{I}{\Delta t} \times \frac{\bar{d}}{S + P}$$

$$\frac{P}{S + P} = \frac{I}{\Delta t} \times \frac{\bar{d}}{S + P} \times \frac{S}{S}$$

$$\frac{P}{S + P} = \frac{I}{S \times \Delta t} \times \frac{\bar{d}}{1} \times \frac{S}{S + P}$$

ce qui peut s'écrire

$$Pr = Ti \times \bar{d}\ (1 - Pr)$$

La relation que nous cherchons est donnée par:

$$\frac{Pr}{1 - Pr} = Ti \times \bar{d}$$

ou, dans une forme équivalente, par:

$$Pr = \frac{Ti \times \bar{d}}{1 + Ti \times \bar{d}}$$

Dans le cas où la prévalence relative est faible ($1 - Pr \sim 1$), on obtient:

$$Pr \simeq Ti \times \bar{d}.$$

Si, dans une population stable, le taux d'incidence d'une maladie (en situation d'équilibre) est de deux cas pour 1000 personnes-années à risque ($Ti = 0{,}002$ par année) et la durée moyenne de cette maladie est cinq ans ($\bar{d} = 5$ ans), alors la prévalence relative de cette maladie est, à un moment quelconque, de 0,0099. Puisque *Pr* est faible, on peut en bonne approximation écrire:

$$Pr = \frac{0{,}002}{\text{an}} \times 5 \text{ ans} = 0{,}01$$

RELATION ENTRE LE TAUX D'INCIDENCE ET L'INCIDENCE CUMULATIVE

Dans une population dynamique fermée observée du temps t_0 au temps t_1, aux conditions d'un taux d'incidence *(Ti)* constant pour la période et d'aucun retrait autre que ceux attribuables à la maladie considérée, on peut démontrer que

$$Ic = 1 - e^{-Ti\,(t_1 - t_0)}$$

où e est le nombre népérien (e = 2,71828...). Si le produit $Ti\ (t_1 - t_0)$ est faible, l'expression peut être simplifiée à

$$Ic \simeq Ti\,(t_1 - t_0)$$

Cette relation permet de passer du taux d'incidence à l'incidence cumulative et inversement.

Si on reprend l'exemple des 1000 écoliers où l'incidence cumulative à un an d'infection est de 0,056 et le taux d'incidence de 0,0571 cas par personne-année, on vérifie facilement que

$$Ic\,(= 0{,}056) = 1 - e^{-0{,}0571 \times 1}$$

Cette relation est d'autant plus intéressante qu'elle nous permet d'estimer dans une population ouverte l'incidence cumulative. Rappelons qu'il est difficile de calculer directement l'incidence cumulative dans une population ouverte.

Mesures de fréquence des décès

La description de la fréquence des décès passe par les mesures d'incidence qui sont des mesures de fréquence d'événements. On comprendra que la prévalence, qui est une mesure de fréquence d'état, y est sans intérêt. Le décès, comme événement, intéresse la santé. À la limite, c'est l'événement qu'on veut à court terme éviter, sinon retarder. Le décès est l'événement final. Une fois produit, l'événement décès définit pour l'individu un état irréversible, permanent qui, au plan scientifique, ne connaît pas de terme et a une durée indéfinie. L'état « être mort » soustrait à jamais l'individu de tout risque de toute maladie, y compris du risque de décéder bien entendu. L'état « être mort » ne comporte aucune information qui puisse intéresser la description d'un problème de santé. Comme nous venons de le souligner, il en va autrement pour l'événement décès.

Nous avons différencié, pour la maladie, le taux d'incidence de la probabilité (ou incidence cumulative) : nous ferons de même pour le décès en distinguant taux de décès et probabilité de décès. Les mesures d'incidence du décès ne peuvent être calculées que sur des populations dynamiques.

TAUX DE DÉCÈS (OU DE MORTALITÉ)

Le taux de décès (T_D) ou de mortalité d'une population pour une période déterminée se définit comme un taux d'incidence, soit ici le rapport de l'incidence des cas de décès (I_D) sur le nombre cumulé de personnes-temps à risque *(PT)* de décéder, dans la population pour la période déterminée.

$$T_D = \frac{I_D}{PT}.$$

Généralement, l'unité de temps considérée est un an. Le taux est alors exprimé en personnes-années à risque.

Les taux de décès peuvent être utilisés pour décrire la mortalité dans une population générale sans autre considération particulière : ce sont des taux bruts. Pour une description plus complète, il convient de décrire la mortalité suivant certaines catégories de variables ou de causes: ce sont les taux spécifiques.

TAUX BRUT DE DÉCÈS

Pour une population donnée, on définit le *taux brut* de décès comme le rapport entre le nombre de décès survenus au cours d'une période donnée et le nombre de personnes-temps cumulées pour cette période, sans référence particulière à un sous-groupe de la population considérée ou à une cause spécifique.

$$\text{Taux brut de décès} = \frac{\text{nombre de décès pendant la période}}{\text{personnes-temps}}$$

Si, en Sanpulie au cours d'une année, il y a eu 32 855 décès dans une population de 5133 580, alors le taux brut de mortalité est de 6,4 décès par 1000 personnes-années.

$$T_D \text{ (brut)} = \frac{32\,855 \text{ décès}}{5133\,580 \text{ personnes-années}}$$

$$= 6{,}4 \text{ décès par } 1000 \text{ personnes-années.}$$

Le taux brut est influencé par la structure de la population par âge ou suivant d'autres variables. Le taux brut regroupe également la totalité des causes de décès et ainsi n'apporte pas d'information sur l'importance relative de celles-ci.

TAUX SPÉCIFIQUE DE DÉCÈS

Les taux de décès peuvent être spécifiques en regard d'une cause de décès ou encore pour certains sous-groupes caractérisés par l'âge, le sexe, l'occupation ou autre.

Si l'on s'intéresse à une cause particulière de décès dans une population le taux spécifique de décès par cette cause (ou *taux de décès par une cause)* se définit comme suit:

$$\text{Taux de décès par une cause} = \frac{\text{nombre de décès dus à cette cause}}{\text{personnes-temps}}.$$

En Sanpulie au cours d'une année, il y a eu 159 décès dus à la tuberculose dans une population de 5133 580; le taux spécifique de décès par tuberculose est alors de 3,1 décès par 100 000 personnes-années.

Si l'on s'intéresse à un sous-groupe particulier de la population caractérisé par l'âge,

le sexe ou toute autre variable, le taux spécifique à ce sous-groupe est mesuré par:

$$\frac{\text{Nombre de décès dans ce sous-groupe}}{\text{Personnes-temps cumulées dans ce sous-groupe}}.$$

En Épidélie au cours d'une année, il y a eu 5601 décès dans le groupe d'âge 50-59 ans, celui-ci comprenant 572 900 individus. Le taux de décès pour ce groupe d'âge est de 9,8 décès par 1000 personnes-années.

En Épidélie au cours d'une année, il y a eu 13 721 décès dans la population masculine de 1960 140 individus. Le taux de décès dans la population masculine est de 7,0 décès par 1000 personnes-années.

Nous donnons dans l'annexe de ce chapitre la description de quelques mesures de mortalité particulières se rapportant à la période foeto-infantile.

PROBABILITÉ DE DÉCÈS (OU DE MORTALITÉ)

Une probabilité de décès utile à considérer est la létalité, c'est-à-dire la probabilité de décéder pour une personne atteinte d'une maladie donnée. Une autre probabilité intéressante est celle spécifique à un groupe d'âge. Quelle est, par exemple, la probabilité de décéder avant 65 ans pour un individu qui a atteint l'âge de 60 ans?

LÉTALITÉ

Une personne atteinte d'une maladie peut décéder de cette maladie ou mourir d'une autre cause. Un individu atteint du cancer de la vessie peut décéder des suites de ce cancer, mais peut aussi décéder dans un accident de la route. Ces faits nous amènent à distinguer deux sortes de létalité: 1) la létalité par la cause; et 2) la létalité toute cause.

Pour une cohorte de nouveaux cas d'une maladie observés au cours d'une période donnée, on définit la *létalité toute cause* (ou simplement la *létalité)* comme la proportion de décès toute cause survenus dans cette cohorte durant la période considérée.

$$\text{Létalité (toute cause)} = \frac{\text{nombre de décès toute cause}}{\text{nombre de nouveaux cas dans la cohorte}}.$$

Avec la létalité, on doit toujours indiquer la durée de la période d'observation: létalité à six mois, à un an, à deux ans, etc. Notons au passage que la létalité est une incidence cumulative de décès.

Dans une cohorte de 150 nouveaux cas d'une maladie, on a dénombré, après un an d'observation, douze décès toute cause. La létalité toute cause à un an est donc de $^{12}/_{150}$, soit 8 %.

Selon l'usage, on parle de taux de létalité. Cependant, on aura reconnu que la létalité, comme l'incidence cumulative, n'est pas un taux, mais une proportion (une probabilité). Généralement, on s'intéresse davantage à la probabilité complémentaire de la létalité, c'est-à-dire la probabilité de survie. Nous abordons et développons cette notion au chapitre 11.

La *létalité par la cause* d'une maladie est limitée aux seuls décès par cette maladie.

$$\text{Létalité par la cause} = \frac{\text{nombre de décès par la maladie}}{\text{nombre de nouveaux cas dans la cohorte}}.$$

La létalité par la cause comporte une difficulté majeure dans son estimation. Reportons-nous à l'exemple précédent où trois décès sont survenus par autre cause. Il faut bien souligner qu'au moment où un individu décède, il cesse alors d'être à risque de décéder pour la maladie considérée. Dans cet exemple, les trois individus décédés par autre cause ne peuvent donc être inclus ni totalement dans le dénominateur, ni totalement dans le numérateur, puisqu'ils n'ont pas couvert la période entière de risque et qu'ils y ont été exclus par autre cause que le décès par la maladie. La létalité par la cause à un an est-elle alors de $^{9}/_{150}$ ou de $^{9}/_{47}$? La létalité par la cause devrait être comprise comme la probabilité conditionnelle de décéder de la maladie, la condition étant qu'il n'y ait pas d'autre cause. En pratique, cette mesure est peu utilisée et remplacée par la mesure de survie relative. Cette notion sera abordée au chapitre 11.

PROBABILITÉ DE DÉCÈS DANS UN GROUPE D'ÂGE

Supposons que le taux de décès chez les hommes du groupe d'âge 60-64 ans soit de 25 par 1000 personnes-années. Quelle est la probabilité qu'un homme qui vient d'avoir 60 ans décède avant son 65e anniversaire? Cette probabilité est facile à calculer si on adapte au problème du décès la formule suivante que nous avons vue un peu plus haut:

$$Ic = 1 - e^{-Ti\,(t_1 - t_0)}.$$

Adaptée au problème de décès, cette relation devient :

$$q_x = 1 - e^{-T_{Dx} \cdot h_x}$$

où T_{Dx} dénote le taux de décès du groupe d'âge x, h_x l'intervalle d'âge, par exemple 5 ans pour le groupe d'âge 60-64 ans, *et* q_x la probabilité recherchée. T_{Dx} doit être constant sur le groupe d'âge *x*.

Si on pose $T_{D_{60\text{-}64\ ans}} = 0{,}025$,
$h_{60\text{-}64\ ans} = 5$ ans,

on a:

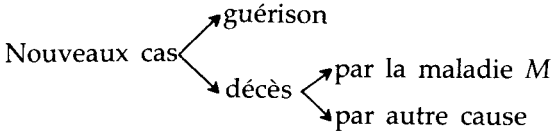
$$q_{60\text{-}64\ ans} = 1 - e^{-0{,}025 \times 5\ ans}$$

$$= 0{,}1175 \text{ ou } 11{,}75\,\%.$$

RELATION ENTRE LE TAUX DE DÉCÈS, LE TAUX D'INCIDENCE ET LA LÉTALITÉ PAR LA CAUSE

Considérons une maladie M en état d'équilibre dans une population stable. Ce pourrait être, par exemple, en bonne approximation, l'infarctus du myocarde considéré dans une vaste population (celle d'un pays) sur une période de cinq ans. Identifions une cohorte de nouveaux cas de cette maladie et observons chaque cas jusqu'à la disparition de la maladie, celle-ci pouvant cesser de deux façons exclusives : la guérison ou le décès. Le décès lui-même est dû à la maladie *M* ou à une autre cause, comme l'illustre la figure 4-11. On y identifie donc trois types de sortie: 1) la guérison; 2) le décès par la cause *M;* et 3) le décès par autre cause.

Figure 4-11



Dans cette cohorte, la proportion de décès par la cause *M* est en quelque sorte une létalité par la cause *(Lc)* que nous qualifierons de finale, en ce sens que chaque nouveau cas est suivi jusqu' à la disparition de la maladie.

La stabilité de la population garantit la constance de cette proportion (L*c*). Si on applique cette proportion (L*c*) au nombre total de sorties S qui se sont produites dans cette population pour une période déterminée, on obtient le nombre de décès par la maladie *M* (D_M) pour cette même période. On peut écrire :

$$D_M = S \times L_C.$$

De plus, la situation d'équilibre de la maladie assure justement l'équilibre entre les sorties S et les entrées I (incidence, nouveaux cas). On peut donc remplacer dans l'équation précédente S par I, pour obtenir:

$$D_M = I \times L_C.$$

La relation tient également pour les taux.

$$T_{D_M} = Ti \times L_C.$$

Cette relation illustre bien le fait que la mortalité est dépendante tant de l'incidence que de la létalité. Tout changement dans la mortalité par une maladie peut s'expliquer par des changements dans l'une ou l'autre des deux composantes *I* et *Lc.* Si, par des mesures préventives, on diminue l'incidence d'une maladie, alors la mortalité sera pour autant diminuée. Si, par les interventions curatives, on réduit la létalité, on trouvera aussi une diminution correspondante dans la mortalité. La prévention et le traitement sont les deux modes complémentaires d'intervention pour réduire la mortalité d'une maladie.

OPÉRATIONS SUR LES MESURES

Peut-on additionner des mesures spécifiques de fréquence pour obtenir une mesure globale? Nous tenterons de répondre à cette question dans les sections suivantes.

Somme arithmétique de mesures

Au cours d'une année, on observe dans un groupe de 20 000 enfants âgés de un à cinq ans, 18 décès par accident de véhicule motorisé et 162 décès par autre type d'accident (accident digestif ou respiratoire, chute ...). On a alors respectivement les taux de décès de 9 par accident de véhicule et de 81 par autre type d'accident pour 10 000 personnes-années. Dans la même population, on observe au total 180 décès par accident, d'où un taux de 90 décès pour 10 000 personnes-années. Il en résulte la règle simple que le taux de décès par accident est la somme (9 + 81 par 10 000) des décès par accident de véhicule et par autre type d'accident. De façon générale, si le numérateur *N* est décomposable en parties mutuellement exclusives N_1 et N_2, on peut écrire la relation suivante pourvu que le dénominateur D ne change pas:

$$\frac{N}{D} = \frac{N_1}{D} + \frac{N_2}{D}.$$

Dans l'exemple, le dénominateur est toujours le même, soit 20 000 personnes-années. Les numérateurs N_1 et N_2 constituent une partition des décès par accident.

Si la décomposition $N_1, N_2, \ldots, N_k$, est une partition de N, alors on a:

$$\frac{N}{D} = \frac{N_1}{D} + \ldots + \frac{N_k}{D}.$$

Suivant cette règle, la somme des taux spécifiques de décès par chacune des causes dans une population est égale au taux brut de décès. Cette règle n'est pas universelle comme nous allons maintenant le constater.

Somme pondérée de mesures

Au cours d'une année, on a observé 500 décès dans une population de 1000 000 de personnes. Le tableau 4-1 illustre la distribution d'âge pour cette population et la répartition des 500 décès par groupe d'âge. Les taux de décès sont exprimés par 100 000 personnes-années.

Chaque mesure spécifique réfère à la mortalité dans un sous-groupe particulier et exclusif de la population. Ces sous-groupes sont, dans notre exemple, 35-54, 55-64 et 65-74 ans; ils sont tous différents et composent la population entière :

600 000 + 300 000 + 100 000 = 1000 000.

Cette situation est différente de celle de l'exemple précédent, où les mesures spécifiques, comme la mesure brute, référaient toutes à la même population de 20 000 personnes. Il est donc erroné de considérer ici le taux brut de décès (50) comme la somme arithmétique des taux de décès spécifiques aux groupes d'âge (10, 60 et 260) :

$$50 \neq 10 + 60 + 260$$

Le taux brut de décès *(Te)* est plutôt une somme pondérée des taux spécifiques de décès *(T_D):*

$$T_D = \Sigma\, p_x T_{Dx}.$$

Cette relation est facilement vérifiable:

50 = (0,60 x 10) + (0,30 x 60) + (0,10 x 260)

où les proportions 0,60, 0,30 et 0,10 sont les coefficients de pondération qui, ensemble, forment un système de poids. De manière générale, tout ensemble (p_x) de coefficients avec les deux propriétés $0 \leq p_x \leq 1$ et $\Sigma p_x = 1$ constitue un système de poids.

Le taux brut apparaît bien comme une somme pondérée des taux spécifiques. De

Tableau 4-1

Groupe d'âge *(x)*	Personnes-années	Proportion *(P_x)*	Décès *(D_x)*	Taux de décès *(T_{Dx})*
35-54	600 000	0,60	60	10
55-64	300 000	0,30	180	60
65-74	100 000	0,10	260	260
Total	**1 000 000**	**1,00**	**500**	**50**

façon générale, si les mesures spécifiques réfèrent à des sous-groupes qui forment une partition de la population, alors la mesure brute est la somme pondérée des mesures spécifiques.

RÉSUMÉ

Les mesures de fréquence sont des proportions, des ratios, des indices ou des taux. Ces mesures s'expriment toutes comme le rapport de deux quantités. Il y a les mesures de fréquence de la maladie et celles du décès. Au nombre des premières, on compte la prévalence relative, le taux (densité) d'incidence et l'incidence cumulative. La prévalence relative d'une maladie dans une population est le rapport entre le nombre de cas prévalents (prévalence) et le nombre de personnes dans cette population au moment considéré. Le taux d'incidence est le rapport entre le nombre de cas incidents (incidence) et le nombre de personnes-temps à risque cumulé par la population en observation. L'incidence cumulative pour une période déterminée est définie comme le rapport entre l'incidence et le nombre d'individus à risque au début de cette période. Les mesures de fréquence de décès sont des mesures d'incidence. Ce sont les taux bruts et les taux spécifiques de mortalité, les probabilités de décès et la létalité, laquelle est une proportion.

Symboles

M, M_1, M_0: maladie, malades, non-malades

S: nombre d'individus non-affectés mais qui risquent de l'être

S_0, S_1: nombre d'individus à risque au temps t_0, au temps t_1

PT: personnes-temps à risque

$\triangle t$: intervalle de temps

P, Pr: prévalence, prévalence relative

I, Ti, Ic: incidence, taux (densité) d'incidence, incidence cumulative

$\overline{d}$: durée moyenne de la maladie

I_D, T_D, T_{D_x}: incidence des cas de décès, taux de décès (de mortalité), taux de décès spécifique au groupe d'âge x.

p_x: proportion d'individus dans le groupe d'âge x

h_x: intervalle du groupe d'âge x

q_x: probabilité de décès dans le groupe d'âge x

L_C: létalité liée à la cause

D_M: décès par la maladie (M)

Formules

$$Pr = \frac{P}{S + P}$$

$$P = \frac{I}{\Delta t} \times \bar{d}$$ (pour une maladie en situation d'équilibre dans une population stable)

$$\frac{Pr}{1 - Pr} = Ti \times \bar{d}$$ (idem au précédent)

$$Pr \simeq Ti \times \bar{d}$$ (idem et Pr faible)

$$Ti = \frac{I}{PT}$$

$$Ic = \frac{I}{S_0}$$

$$Ic = 1 - e^{-Ti(t_1 - t_0)}.$$ (pour Ti constant sur la période t_0 à t_1)

$$Ic \simeq Ti(t_1 - t_0)$$ (idem au précédent et $Ti(t_1 - t_0)$ faible)

$$T_D = \frac{I_D}{PT}$$

$$q_x = 1 - e^{-T_{D_x} h_x}$$ (pour T_{D_x} constant sur le groupe d'âge x)

$$D_M = I \times L_C$$ (pour une maladie en situation d'équilibre dans une population stable)

$$T_{D_M} = Ti \times L_C$$ (idem au précédent)

$$T_D = \Sigma\, p_x T_{D_x}$$

LECTURES SUGGÉRÉES

1. JENICEK, M. et CLÉROUX, *R.* ***Épidémiologie,*** Saint-Hyacinthe, Edisem, 1982, chapitre 3, pp.43-78.
2. KLEINBAUM, D.G., KUPPER, L.L. et MORGENSTERN, *H.* ***Epidemiologic Research,*** Belmont (USA), Lifetime Learning Publications, 1982, chapitre 6, pp. 96-139.
3. ROTHMAN, *K.J.* ***Modern Epidemiology,*** Boston, Little, Brown, 1986, chapitre 3, pp. 23-34.
4. RUMEAU-ROUQUETTE, C., BRÉART, G. et PADIEU, R. ***Méthodes en épidémiologie,*** Paris, Flammarion, 1985, chapitre XXII, pp. 237-253.

ANNEXE DU CHAPITRE 4

Mesures de la mortalité pour la période foeto-infantile

Pour décrire la mortalité par rapport à l'événement naissance, plusieurs mesures sont couramment utilisées en épidémiologie et en santé publique. Chacune d'elles distingue la mortalité suivant des périodes d'âge plus ou moins rapprochées de la naissance elle-même. Il y a d'abord la grande période foeto-infantile qui va approximativement de 20 semaines de gestation à 364 jours d'âge inclus. Ce critère de 20 semaines est, d'un certain point de vue, discutable. L'OMS suggère d'utiliser le poids de préférence au nombre de semaines de gestation pour la définition de la mortalité foetale. Quoi qu'il en soit, la période foeto-infantile peut être divisée en sous-périodes:

— période foetale ..... de 20 semaines de gestation à la naissance; (poids d'au moins 500 g)

— période infantile ..... de la naissance à 364 jours d'âge inclus;

— période néonatale ..... de 0 à 27 jours d'âge inclus;

— période néonatale précoce ..... de 0 à 6 jours d'âge inclus;

— période néonatale tardive ..... de 7 à 27 jours d'âge inclus;

— période post-néonatale ..... de 28 à 364 jours d'âge inclus;

— période périnatale ..... période foetale et période néonatale précoce.

Les mesures de la mortalité correspondant à ces périodes sont définies ci-après. Remarquons qu'elles sont des indices bien que suivant l'usage on les appelle taux.

Pour les indices de mortalité périnatale et de mortinaissances, la plupart des pays occidentaux utilisent un poids d'au moins 500g dans leurs statistiques nationales.

$$\text{Indice de mortinaissances (OMS statistiques nationales)} = \frac{\text{nombre de morts foetales (foetus pesant 500 g et plus)}}{\text{nombre de morts foetales et de naissances vivantes (500 g et plus)}}$$

$$\text{Indice de mortalité infantile} = \frac{\text{nombre de décès d'enfants avant l'âge d'un an pendant l'année}}{\text{nombre de naissances vivantes pendant la même année}}$$

$$\text{Indice de mortalité néonatale} = \frac{\text{nombre de décès d'enfants avant l'âge de 28 jours pendant l'année}}{\text{nombre de naissances vivantes pendant la même année}}$$

$$\text{Indice de mortalité néonatale précoce} = \frac{\text{nombre de décès d'enfants avant l'âge de 7 jours}}{\text{nombre de naissances vivantes pendant la même année}}$$

$$\text{Indice de mortalité néonatale tardive} = \frac{\text{nombre de décès d'enfants âgés de 7 à 27 jours inclus}}{\text{nombre de naissances vivantes pendant la même année}}$$

$$\text{Indice de mortalité post-néonatale} = \frac{\text{nombre de décès d'enfants âgés de 28 jours et plus, mais de moins d'un an pendant l'année}}{\text{nombre de naissances vivantes pendant la même année}}$$

$$\text{Indice de mortalité périnatale (définition de l'OMS pour les statistiques internationales)} = \frac{\text{nombre de morts foetales et de morts néonatales précoces de produits de conception pesant 1000 g et plus à la naissance}}{\text{nombre de morts foetales et nombre de naissances vivantes de produits de conception pesant 1000 g et plus}}$$

$$\text{Indice de mortalité périnatale (définition de l'OMS pour les statistiques nationales)} = \frac{\text{nombre de morts foetales et de morts néonatales précoces de produits de conception pesant 500 g et plus à la naissance}}{\text{nombre de morts foetales et nombre de naissances vivantes de produits de conception pesant 500 g et plus}}$$

Le poids d'au moins 1000 g est recommandé par l'OMS pour les comparaisons internationales. Les poids de 500 à 1000 g correspondraient respectivement, en moyenne, à des périodes de gestation de 20 et 28 semaines.

CHAPITRE 5

Espérance de vie

Nous allons examiner l'espérance de vie dans deux situations: celle où la génération des individus est éteinte au moment du calcul et celle où la génération ne l'est pas encore. Ce qui nous amène à présenter le calcul de l'espérance de vie selon une cohorte réelle d'individus et celui selon une cohorte fictive. Parallèlement, nous définirons l'espérance de vie à un âge quelconque. Enfin, nous considérerons le type d'influence que peuvent avoir les taux spécifiques de décès sur l'espérance de vie.

ESPÉRANCE DE VIE D'UNE COHORTE RÉELLE

Considérons une génération d'individus nés la même année et supposons que l'on puisse la suivre dans le temps jusqu'à son extinction.

Espérance de vie à la naissance

On observe une génération de 200 individus nés la même année, où chacun est suivi jusqu'à ce qu'il décède. Supposons que 80 d'entre eux décèdent la première année, 30 la deuxième, 20 la troisième, 30 la quatrième et 40 la cinquième. On peut alors facilement calculer approximativement l'âge moyen au décès de ces 200 individus en faisant la moyenne sur des données regroupées en classe. En fait, on suppose que les 80 individus qui décèdent la première année ont en moyenne vécu 0,5 année; les 30 qui décèdent la deuxième année ont en moyenne vécu 1,5 année; ainsi de suite. On obtient alors,

$$\text{Âge moyen au décès} = \frac{(80 \times 0{,}5) + (30 \times 1{,}5) + (20 \times 2{,}5) + (30 \times 3{,}5) + (40 \times 4{,}5)}{200}$$

$$= \frac{420}{200} = 2{,}1 \text{ années.}$$

En moyenne les membres de cette cohorte ont vécu 2,1 années. On peut se permettre d'interpréter cet âge moyen au décès comme une espérance de vie à la naissance.

Pour une génération qui serait suivie jusqu'à son extinction, l'âge moyen au décès des individus indique en quelque sorte ce qu'aurait été leur *espérance de vie à la naissance* (e_0). Formellement, si T_0 désigne le total des années vécues par tous les membres de la génération depuis leur naissance et S_0 le nombre d'individus de cette génération à la naissance, alors

$$e_0 = \frac{T_0}{S_0}.$$

Espérance de vie à l'âge x

Dans l'exemple précédent, 120 des 200 individus de la génération ont survécu à la première année. Le nombre moyen d'années qu'il leur reste à vivre est donc de :

$$e_1 = \frac{30 \times (1{,}5 - 1) + 20 \times (2{,}5 - 1) + 30 \times (3{,}5 - 1) + 40 \times (4{,}5 - 1)}{120}$$

$$= 2{,}2 \text{ années.}$$

Ce qui signifie que les membres de la génération qui ont survécu à la première année ont encore vécu en moyenne 2,2 années.

Dans la même génération, 90 membres ont survécu aux deux premières années. Le nombre moyen d'années qui leur reste à vivre après ces deux premières années est de:

$$e_2 = \frac{20 \times (2{,}5 - 2) + 30 \times (3{,}5 - 2) + 40 \times (4{,}5 - 2)}{90}$$

$$= 1{,}7 \text{ année.}$$

Les membres de la génération qui ont survécu aux deux premières années ont donc encore vécu en moyenne 1,7 année.

Les valeurs e_1 et e_2 sont dites respectivement l'espérance de vie à un an et à deux ans. Ainsi peut-on calculer pour les survivants à un âge quelconque x le nombre moyen d'années qu'il leur reste à vivre. C'est l'espérance de vie à l'âge *x:* e_x.

Calcul par la méthode des années de vie contribuées

Une autre façon de déterminer l'espérance de vie consiste à cumuler pour chaque période les années de vie contribuées par les survivants à la période et par les décédés dans la période. On pose l'hypothèse que les décédés dans la période ont vécu chacun en moyenne une demi-période. La somme totale de ces années contribuées divisée par le nombre total d'individus dans la génération donne bien la moyenne des années de vie, c'est-à-dire l'espérance de vie. Pour la génération rapportée dans l'exemple précédent, on construit le tableau 5-1.

Ainsi,

$$e_0 = \frac{320 + 100}{200} = 2{,}1 \text{ années.}$$

On pourrait par cette méthode, bien sûr, calculer l'espérance de vie à un âge x. Par exemple, l'espérance de vie à 1 an serait de

$$e_1 = \frac{(320 - 120) + (100 - 40)}{120} = 2{,}2 \text{ années.}$$

On doit souligner qu'il est d'usage, dans les tables de mortalité et dans le calcul de l'espérance de vie à la naissance, de diviser la première année en périodes plus courtes: 0-6 jours, 7-27 jours, 28 jours-5 mois, 6-11 mois. La première année de vie comporte un nombre élevé de décès dans les jours et semaines qui suivent de près la naissance. Ici, dans un désir de simplification, nous ne faisons pas ces distinctions.

ESPÉRANCE DE VIE POUR UNE GÉNÉRATION NON ENCORE ÉTEINTE

L'espérance de vie d'une génération réelle d'individus présente un inconvénient majeur: elle ne peut être calculée qu'une fois la génération éteinte. Ainsi, par exemple, peut-on déterminer l'espérance de vie d'une génération d'individus nés en 1880 ou en 1900. Mais, comment peut-on parler d'espérance de vie d'individus nés en 1985 alors que la

Tableau 5-1

Âge	Survivants	Décès	Années contribuées par les	
x	à l'âge *x*	entre *x* et x + *1*	survivants	décédés
0	200	80	120	40
1	120	30	90	15
2	90	20	70	10
3	70	30	40	15
4	40	40	0	20
5	0			
Total			**320**	**100**

génération est loin d'être éteinte? Et pourtant, en feuilletant des publications de statistiques sur la santé, on peut lire qu'en 1985, l'espérance de vie à la naissance en Sanpulie est de 75 ans pour les femmes, tandis qu'elle est de 71 ans pour les hommes. Voici comment, pour une population déterminée, un tel calcul peut être effectué, disons pour la population masculine sanpulienne.

On considère une cohorte fictive de taille initiale S0 que l'on fait évoluer sur les différents groupes d'âge (x) en la soumettant, pour chacun de ces groupes, au risque de décès réel de la population considérée. On pose, par commodité, S_0 = 100 000. Cette cohorte décroît sur un groupe d'âge en raison du risque de décès qui y prévaut. Ainsi de S_0, elle passera à une taille moindre de S_1 en raison du risque q_0 qui prévaut sur le premier groupe d'âge. De S_1 qu'elle est au début du second groupe d'âge, elle passe à S_2 à la fin de ce groupe d'âge en raison du risque q, qui y prévaut. De S_x, elle passera *à* $S_x + 1$ ... pour enfin s'éteindre sur le dernier groupe d'âge (k).

Calcul des risques de décès

Le calcul de l'espérance de vie pour une génération non encore éteinte passe d'abord par l'estimation des risques de décès par groupe d'âge. Cette estimation est faite à partir des taux de décès qui sont connus pour la population à laquelle appartient la génération. Nous rappelons cependant qu'il est facile, pour un groupe d'âge x d'intervalle h_x, d'estimer le risque (ou la probabilité) de décès (q_x) à partir du taux de décès (T_{Dx}) pour le même groupe d'âge. Il suffit de considérer la relation:

$$q_x = 1 - e^{-T_{Dx}h_x}$$

décrite au chapitre 4 sur les mesures de fréquence. (On trouve en annexe une autre estimation de q_x, le quotient de mortalité.) Ainsi, si le taux de décès pour le groupe d'âge 60-79 ans est de 30,0 décès par 1000 personnes-années, en posant

$$T_{D_{60\text{-}79}} = 0{,}030/\text{an}$$

et

$$h_{60\text{-}79} = 20 \text{ ans}$$

on a alors:

$$q_{60\text{-}79} = 1 - e^{-0{,}030 \times 20} = 1 - e^{-0{,}600} = 0{,}4512.$$

Pour la population masculine de Sanpulie, les taux de décès sont décrits au tableau 5-2. Contrairement à la pratique qui existe dans tous les pays (mais heureusement pour la simplification de nos calculs), les statistiques de mortalité en Sanpulie ne sont rapportées que pour les cinq grands groupes d'âge suivants : 0-19, 20-39, 40-59, 60-79, 80 ans et plus. Chaque taux annuel, exprimé par 1000 de population, est accompagné du risque correspondant q_x pour un individu de décéder dans l'intervalle d'âge considéré s'il a survécu jusqu'à cet intervalle. Les q_x sont estimés par la transformation $1 - e^{-T_{Dx}h}x$, *à* l'exception de celui du dernier groupe d'âge, puisque l'intervalle h_{80+} est indéterminé.

Tableau 5-2

Groupe d'âge (ans) x	Taux de décès par 1000 personnes-années T_{DX}	Risque de décès q_x
0-19	0,003	0,0582
20-39	0,001	0,0198
40-59	0,004	0,0769
60-79	0,030	0,4512
80 +	0,200	—

Bien que l'on ne puisse pas estimer pour le dernier groupe d'âge le risque de décès, le taux 0,200 par année (ou 200 décès par 1000 personnes-années) n'en demeure pas moins une information de première importance pour le calcul de l'espérance de vie. Cette donnée nous permet d'estimer facilement le nombre total d'années qu'il reste à vivre aux survivants des 80 premières années. En effet, sur ce groupe d'âge résiduel, l'inverse du taux de décès peut être considéré comme l'espérance de vie à 80 ans, ce qui permet alors d'estimer la durée moyenne de vie des individus qui ont atteint 80 ans.

$$e_{80} = \frac{1}{0{,}200} = 5 \text{ ans.}$$

On peut donc dire que les individus âgés de 80 ans vivront encore en moyenne cinq années. S'ils sont, par exemple, au nombre de 5000, ils cumuleront donc au total:

$$5 \text{ ans} \times 5000 \text{ personnes} = 25\,000 \text{ personnes-années.}$$

Calcul de l'espérance de vie

Une fois connus les risques de décès par groupe d'âge, il ne reste plus qu'à faire évoluer la cohorte fictive des 100 000 individus sur les différents groupes d'âge depuis la naissance jusqu'à l'extinction. Pour chaque groupe d'âge, on attribue à cette cohorte le risque de décès qui prévaut. Les personnes-années générées pour chaque groupe d'âge sont alors calculées et cumulées. Le total des personnes-années cumulées par la cohorte divisé par le nombre d'individus dans cette cohorte donne alors l'estimation de l'espérance de vie à la naissance e_0. De même, le total des personnes-années cumulées à partir d'un âge (x) divisé par le nombre d'individus qui ont survécu jusqu'à cet âge permet d'estimer e_x, soit l'espérance de vie à l'âge x. Ces données sont résumées dans un tableau qui peut comprendre huit colonnes. Le tableau 5-3, qui résume cette démarche pour la population sanpulienne de 1985, en est un exemple. Nous expliquons ici chacune des huit colonnes d'un tel tableau.

La colonne des intervalles (x):	elle donne pour chaque groupe d'âge les limites de l'intervalle.
La colonne des vivants (S_x):	elle indique le nombre de vivants au début de l'intervalle. Pour l'intervalle x, ce nombre est obtenu en soustrayant du nombre (S_{x-1}) le nombre de décès (D_{x-1}). L'indice ($x - 1$) réfère au groupe d'âge qui précède immédiatement le groupe (x). Le nombre S_0 de vivants au début du premier intervalle est arbitraire et souvent pris égal à 100 000.
La colonne des probabilités ou risques de décès (q_x):	elle donne pour l'intervalle considéré la probabilité de décéder durant cet intervalle. Cette probabilité est estimée, sauf pour le dernier groupe d'âge, par la relation $1 - e^{-T_{D_x} h_x}$.

La colonne des décès *(D_x) :* elle donne le nombre de décès dans l'intervalle considéré. Ce nombre est obtenu en multipliant le nombre de vivants S_x au début de l'intervalle par le risque de décès q_x dans l'intervalle. On a,

$$D_x = S_x q_x$$

La colonne du nombre de survivants à l'intervalle *(S_x — D_x) :* elle indique le nombre d'individus qui, parmi les vivants au début de l'intervalle, ont survécu à l'intervalle. Ce nombre est obtenu en soustrayant des vivants (S_x) les décès (D_x) dans l'intervalle.

La colonne des personnes-années ou années contribuées dans l'intervalle (C_x) : pour chaque intervalle, le nombre C_x totalise le nombre d'années vécues dans l'intervalle. Ce nombre est estimé en comptant un intervalle complet pour les survivants et un demi-intervalle pour les décédés. On a donc

$$C_x = (S_x - D_x)\, h_x + D_x \cdot h_x/2.$$

La colonne des personnes-années ou années contribuées dans l'intervalle x et dans tous les intervalles suivants (T_x) : ce nombre compte au numérateur du calcul de l'espérance de vie à l'âge x. Il est obtenu par $T_x = C_x + C_{x+1} + \ldots + C_{x+k}$ où $(x + k)$ désigne le dernier groupe d'âge.

La colonne des espérances de vie *(e_x) :* cette dernière colonne décrit pour chaque âge x (début de l'intervalle considéré) l'espérance de vie à cet âge. On a,

$$e_x = \frac{T_x}{S_x}$$

Le tableau 5-3 révèle une espérance de vie à la naissance estimée à 71,36 ans. Elle est obtenue en faisant le rapport des années totales (T_0) contribuées (7135740 années) au nombre total des individus (S_0) qui ont cumulé ces années (100 000). L'espérance de

Tableau 5-3

x	S_x	q_x	D_x	$S_x - D_x$	C_x	T_x	e_x
0-19	100 000	0,0582	5 820	94180	1941800	7 135 740	71,36
20-39	94180	0,0198	1865	92 315	1 864 950	5 193 940	55,15
40-59	92 315	0,0769	7 099	85 216	1 775 310	3 328 990	36,06
60-79	85 216	0,4512	38 448	46 768	1 319 840	1 553 680	18,23
80 +	46 768		46 768	0	233 840	233 840	5,0

vie à vingt ans est de 55,15 ans, c'est-à-dire le nombre total d'années contribuées à partir de 20 ans (T20 = 5193 940) cumulées par les 94 180 individus qui étaient encore vivants au début de l'intervalle d'âge 20-39 ans, et ainsi de suite pour les autres âges.

L'inverse de l'espérance de vie (e_0) doit correspondre à un *taux moyen de sorties* de la cohorte de 0,0140 décès par personne-année.

$$1/e_0 = 1/71{,}36 = 0{,}0140.$$

Concrètement, cela signifie que si la population était affectée pour chaque groupe d'âge d'un taux de décès de 0,0140 par année (ou 14 décès par 1000 personnes-années), on y observerait aussi une espérance de vie à la naissance de 71,36 ans. L'espérance de vie à la naissance est un résumé des taux de mortalité observés dans tous les groupes d'âge de la population où cette espérance de vie est calculée.

De même, $1/e_x$ correspond à un taux moyen de sorties calculé à partir de l'âge x pour la cohorte réduite aux individus qui ont atteint cet âge. Ainsi, pour e_{20} = 55,15 ans, $1/e_{20}$ = 0,0181 par année. Si, pour chaque groupe d'âge à partir de 20 ans, le taux de décès dans la population était de 0,0181 par année (ou 18,1 décès par 1000 personnes-années), on observerait aussi une espérance de vie à 20 ans de 55,15 ans. L'espérance de vie à 20 ans est un résumé des taux de mortalité observés pour tous les groupes d'âge, à partir de 20 ans, de la population où cette espérance de vie est calculée. Il en va ainsi pour les espérances de vie à un âge quelconque.

Influence des taux spécifiques de décès sur l'espérance de vie

Nous allons maintenant comparer entre elles les espérances de vie de quatre populations: les populations masculine et féminine de Sanpulie et celles Épidélie. On décrit, au tableau 5-4, les taux de mortalité respectifs de ces populations et, comme précédemment, ils sont donnés uniquement pour les cinq grands groupes d'âge.

Le tableau 5-5 donne pour chacune de ces quatre populations les espérances de vie à la naissance, à 20 ans, à 40 ans, à 60 ans et à 80 ans. En comparant les espérances de vie de ces différentes populations, on remarque qu'un taux de mortalité influence d'autant plus l'espérance de vie qu'il se situe dans les premiers groupes d'âge. Plus un taux est élevé et plus sa position

Tableau 5-4

Âge	Sanpulie		Épidélie	
(en années)	H	F	H	F
0-19	0,003	0,002	0,004	0,003
20-39	0,001	0,001	0,001	0,001
40-59	0,004	0,004	0,003	0,003
60-79	0,030	0,030	0,030	0,030
80 +	0,200	0,200	0,200	0,200

Tableau 5-5

	Sanpulie		Épidélie	
	H	F	H	F
e_0	71,36	72,59	70,62	71,84
e_{20}	55,15	55,15	55,67	55,67
e_{40}	36,06	36,06	36,59	36,59
e_{60}	18,23	18,23	18,23	18,23
e_{80}	5,00	5,00	5,00	5,00

dans l'échelle de l'âge est proche des premiers groupes d'âge, plus il marque, en la diminuant, l'espérance de vie. C'est ce qui explique, par exemple, que la population masculine Épidélie a une espérance de vie à la naissance (70,62 ans) plus basse que celle des autres populations. Avec un taux de 0,004/an dans le groupe d'âge 0-19 ans, cette population possède la plus faible espérance de vie à la naissance, bien que pour les autres groupes d'âge ses taux soient égaux ou inférieurs aux taux correspondants des autres populations. Qui plus est, sans considérer l'ordre des groupes d'âge, on peut dire que les deux populations masculines ont les mêmes taux de mortalité: 0,001, 0,003, 0,004, 0,030 et 0,200. Mais, la population Épidélie est affectée dès le premier groupe d'âge par un taux plus fort (0,004) que celle de Sanpulie (0,003).

On remarque aussi qu'une modification dans les taux se fait sentir plus fortement si elle se situe dans les premiers groupes d'âge. Les taux de mortalité des populations féminines de Sanpulie et d'Épidélie sont, sauf pour un seul groupe d'âge, identiques à ceux de la population masculine de Sanpulie: une différence (ou réduction) de 0,001 pour le premier groupe d'âge pour la population féminine de Sanpulie; une différence aussi de 0,001 pour le troisième groupe d'âge (40-59 ans) pour la population féminine d'Épidélie. Cette même différence conduit pourtant à un plus grand accroissement de l'espérance de vie à la naissance (72,59 ans) dans la population féminine de Sanpulie comparée à celle d'Épidélie (71,84 ans). Pour cette population de Sanpulie, la différence des taux (en termes de réduction) est observable dans le premier groupe d'âge.

Lonsidérons une autre situation où la population masculine de Sanpulie est comparée à la population masculine d'Onusie. Les taux de cette dernière sont essentiellement les mêmes que ceux de la population masculine de Sanpulie, la différence se situant au plan de l'échelle d'âge décrite. Le tableau 5-6 décrit pour la population onusienne les taux spécifiques par groupe d'âge retenus par cette population.

L'espérance de vie à la naissance de cette population (72,50 ans) est supérieure à celle de la population masculine de Sanpulie (71,36). On peut donc dire que l'influence d'un taux est aussi reliée à la longueur de la catégorie d'âge sur lequel il exerce son influence. Plus l'intervalle d'âge est important, plus l'influence du taux est marquée.

En résumé, chaque taux spécifique de décès (spécifique pour l'âge) peut influencer l'espérance de vie de trois façons:

— par l'importance qu'il a : un taux plus fort conduit à une plus forte diminution de l'espérance de vie;

— par la position qu'il occupe dans l'échelle d'âge : une position plus basse (plus près de l'âge 0) marque une plus grande influence;

Tableau 5-6

Âge	Taux de décès par 1000 personnes-années
x	T_{Dx}
0-9	0,003
10-39	0,001
40-59	0,004
60-79	0,030
80+	0,200

— par la longueur de l'intervalle d'âge dans lequel il se trouve : un plus long intervalle suppose une influence plus forte.

À. partir de ces considérations, on peut déduire que deux espérances de vie à un même âge ne peuvent véritablement être comparées que si elles ont été calculées pour des taux décrits dans une même échelle de classification pour l'âge.

RÉSUMÉ

L'âge moyen au décès d'une génération éteinte peut être vue comme l'espérance de vie à la naissance, laquelle est alors très facile à calculer. Par contre, lorsque la génération n'est pas encore éteinte, comme c'est souvent le cas en pratique, il faut utiliser des moyens détournés pour connaître l'espérance de vie à la naissance ou à tout autre âge. Une fois déterminés les risques de décès par groupe d'âge, on fait évoluer une cohorte fictive d'individus selon ces groupes d'âge depuis la naissance jusqu'à l'extinction. Cette pratique permet l'estimation de l'espérance de vie. Deux espérances de vie à un même âge, calculées sur deux populations différentes, ne peuvent vraiment être comparées que si les deux populations utilisent les mêmes groupes d'âge. Sous l'angle de la mortalité, l'espérance de vie est une sorte de résumé des taux de mortalité. Par exemple, l'espérance de vie à 20 ans est un résumé de tous les taux de mortalité observés dans les groupes d'âge à partir de 20 ans.

Symboles

e_0, e_x: espérance de vie à la naissance, à l'âge x

T_0, T_x: total des années vécues (contribuées) par les membres d'une génération depuis la naissance, depuis l'âge x.

S_0, S_x: (pour une génération) nombre de sujets à la naissance, nombre de sujets vivants au début de l'intervalle x

h_x: intervalle du groupe d'âge x

T_{D_x}: taux de décès dans le groupe d'âge x

q_x: risque ou probabilité de décès dans le groupe d'âge x ou intervalle x

D_x: nombre de décès dans le groupe d'âge x ou intervalle x

C_x: nombre d'années de vie contribuées par les membres du groupe d'âge x

Formules

$$e_0 = \frac{T_0}{S_0}.$$

$$e_x = \frac{T_x}{S_x}$$

$$T_x = C_x + C_{x+1} + \ldots + C_{x+k}$$

$$C_x = (S_x - D_x)\, h_x + D_x \cdot \frac{h_x}{2}$$

$$D_x = S_x \cdot q_x$$

$$q_x = 1 - e^{-T_{D_x}\ h_x}$$

LECTURES SUGGÉRÉES

1. PHILIPPE, P. ***Épidémiologie pratique,*** Montréal, Presses de l'Université de Montréal, 1985, chapitre 4, pp. 51-56.
2. RUMEAU-ROUQUETTE, C., BRÉART, G. et PADIEU, R. ***Méthodes en épidémiologie,*** Paris, Flammarion, 1985, chapitre XXIII, pp 284-301.

ANNEXE DU CHAPITRE 5

Quotient de mortalité

Selon l'hypothèse d'une répartition uniforme des décès sur l'intervalle d'âge, le risque de décès peut être estimé par le quotient de mortalité q_x:

$$q_x = \frac{2T_{D_x}h_x}{2 + T_{D_x}h_x}.$$

Établissons ce quotient de mortalité.

On désigne par S_x le nombre de vivants au début de l'intervalle x. Alors $S_x - S_{x+1}$ mesure le nombre de décès dans l'intervalle x. Ainsi,

$$q_x = \frac{S_x - S_{x+1}}{S_x}$$
$$= 1 - \frac{S_{x+1}}{S_x}$$

ou

$$\frac{S_{x+1}}{S_x} = 1 - q_x.$$

Si on suppose que les décès se répartissent uniformément sur l'intervalle x, on a

$$T_{D_x} = \frac{S_x - S_{x+1}}{\frac{(S_x + S_{x+1})}{2} \times h_x} = \frac{2(S_x - S_{x+1})}{(S_x + S_{x+1})h_x}$$

$$= \frac{2S_x\left(1 - \frac{S_{x+1}}{S_x}\right)}{S_x\left(1 + \frac{S_{x+1}}{S_x}\right)h_x}$$

$$= \frac{2[1 - (1 - q_x)]}{[1 + (1 - q_x)]h_x}$$

De cette dernière relation, en explicitant la valeur q_x, on trouve

$$q_x = \frac{2T_{D_x}h_x}{2 + T_{D_x}h_x}.$$

Selon l'hypothèse d'une répartition uniforme des décès sur l'intervalle d'âge, le risque de décès peut être estimé par le quotient de mortalité q_x:

$$q_x = \frac{2T_{D_x}h_x}{2 + T_{D_x}h_x}.$$

PARTIE III

Mesures composées

CHAPITRE 6

Mesures d'association

Dans ce chapitre, nous allons mesurer l'intensité de la liaison entre deux variables. Parmi les nombreuses mesures d'association, certaines sont plus spécifiques à l'épidémiologie: ce sont les mesures d'association entre un facteur et une maladie, comme le risque attribuable, le risque relatif et le rapport des cotes. D'autres mesures d'association y sont aussi d'usage répandu : ce sont les mesures de corrélation, celle par exemple entre deux variables quantitatives. On examinera le coefficient de corrélation linéaire.

MESURES D'ASSOCIATION ENTRE UN FACTEUR ET UNE MALADIE

La définition d'une mesure d'association entre un facteur *E* et une maladie *M* repose sur la comparaison quantitative des mesures de fréquence (taux d'incidence, taux de mortalité, incidence cumulative...) entre les différentes catégories du facteur (par exemple, les sujets exposés, les sujets non-exposés). L'une des catégories est choisie comme catégorie de référence (par exemple, la catégorie des sujets non-exposés). On compare les mesures de fréquence des autres catégories à celle de la catégorie de référence. Ces comparaisons sont le plus souvent faites à l'aide d'une différence ou d'un rapport de ces mesures.

Pour établir plus clairement les définitions des mesures d'association, nous allons distinguer les deux types d'études à visée étiologique à partir desquelles ces mesures sont généralement effectuées: les études de cohorte et les études cas-témoins. À des fins de simplification, pour chacune d'elles le facteur *E* et la maladie *M* seront considérés comme variables dichotomiques : le facteur *E* est présent (+) ou absent (—), la maladie est présente (+) ou absente (—).

Tableau 6-1

		E		
		+	—	Total
Événement (maladie ou décès)	+	*a*	*b*	
Personnes (personnes-temps)		N_1	N_0	*N*

Étude de cohorte

Le tableau 6-1 décrit, au plan formel, les résultats d'une étude de cohorte.

Dans ce tableau, le nombre *N* représente des personnes-temps à risque si les mesures de fréquence considérées sont des taux (de mortalité ou d'incidence). Par ailleurs, si les mesures de fréquence sont des incidences cumulatives, alors N représente la totalité des personnes de l'étude atteintes ou non de la maladie, avec ou sans le facteur. Les nombres N_1 et N_0 décrivent la répartition de *N* entre les sujets exposés et les sujets non-exposés.

On pose:

$$R_1 = a/N_1 \text{ et } R_0 = b/N_0.$$

Les valeurs R_1 et R_0 sont des mesures de fréquence de la maladie ou du décès respectivement chez les sujets exposés et chez les sujets non-exposés. Bien que ces mesures puissent être des taux d'incidence, des taux de mortalité ou des incidences cumulatives, nous utiliserons le terme générique « mesure du risque » pour les désigner.

DIFFÉRENCE DES MESURES DU RISQUE (RISQUE ATTRIBUABLE)

Une première mesure d'association entre le facteur E et la maladie M nous est donnée par la différence entre les deux mesures du risque R_1 et R_0. Cette différence, notée *DR* dans le texte, est souvent appelée *différence des risques.* (Il existe une notation et une appellation plus spécifiques: *DTi* pour différence entre deux taux d'incidence, DT_D pour différence entre deux taux de mortalité, *DIc*

pour différence entre deux incidences cumulatives.)

$$DR = R_1 - R_0$$

$$= a/N_1 - b/1$$

Dans le cas où le facteur *E* n'est pas associé à la maladie *M,* on a $R_1 = R_0$, donc $DR = 0$. Tout *DR* plus grand que 0, réellement supérieur à 0 au sens statistique, indique une *association positive* entre le facteur *E* et la maladie *M,* association d'autant plus forte que le *DR* est élevé.

Si l'on admet que le facteur est causal, la responsabilité de celui-ci quant au nombre de cas de maladie *M* est d'autant plus grande que le *DR* est élevé. Dans le cas d'un lien de causalité entre un facteur et une maladie, le terme *risque attribuable* et la notation *RA* sont utilisés, de préférence à différence des risques et *DR.*

RAPPORT DES MESURES DU RISQUE (RISQUE RELATIF)

Une autre mesure d'association nous est fournie par le rapport des deux mesures du risque R1 et R0. Ce rapport, généralement désigné par *RR,* est appelé *rapport des risques* ou *risque relatif.* (On trouve aussi une notation plus spécifique : *RTi* pour rapport des taux d'incidence, RT_D pour rapport des taux de mortalité, *Rlc* pour rapport des incidences cumulatives.)

$$RR =$$

$$=$$

$$=$$

Si le facteur E n'est pas associé à la maladie *M,* on a $R_1 = R_0$, donc $RR = 1$. Tout *RR* plus grand que 1, réellement supérieur à 1 au sens statistique, indique une association positive entre le facteur E et la maladie *M,* association d'autant plus grande que le *RR* est élevé. Si l'on admet un lien de causalité, on peut dire que le *RR* marque l'importance du facteur dans l'étiologie de la maladie, et cette responsabilité est d'autant plus grande que le *RR* est élevé.

Le tableau 6-2 décrit les résultats d'une étude dans une population masculine. On y trouve les taux de décès des gros fumeurs (25 cigarettes et plus par jour) et des non-fumeurs pour trois maladies (causes de décès). Ces taux sont exprimés par 100 000 personnes-années. Les risques attribuables *(RA)* et les risques relatifs *(RR)* sont aussi décrits.

Tableau 6-2

Cause de décès	Gros fumeurs	Non-fumeurs	*RA*	*RR*
Cancer du poumon	227	7	220	32,4
Bronchite chronique	106	5	101	21,2
Maladies cardio-vasculaires	993	732	261	1,36

Le risque attribuable *(RA)* est calculé, pour chacune des maladies en soustrayant R_0 de R_1. Par exemple, pour le cancer du poumon, on a :

R_1 = 227 décès par 100 000 personnes-années

et

R_0 = 7 décès par 100 000 personnes-années.

Ainsi, $RA = R_1 - R_0$ = 220 décès par 100 000 personnes-années.

Le risque relatif *(RR)* est calculé, pour chacune des maladies, par le rapport R_1/R_0. Ainsi, pour le cancer du poumon, on a

$$RR = R_1/R_0$$

$$= \frac{^{227}/_{100\,000} \text{ années}}{^{7}/_{100\,000} \text{ années}} = 32{,}4.$$

Le facteur E (fumer beaucoup) est un facteur très important dans l'étiologie du cancer du poumon *(RR = 32,4),* de moindre importance (quoique encore très important) dans l'étiologie de la bronchite chronique *(RR* = 21,2), enfin beaucoup moins important dans l'étiologie des maladies cardiovasculaires *(RR* = 1,36). Par contre, le facteur E est responsable chez les gros fumeurs d'un plus grand nombre de décès par maladie cardiovasculaire *(RA* = 261) que par cancer du poumon *(RA* = 220). Comme on peut le voir, le *RR* mesure l'association sur le plan étiologique, alors que le *RA* décrit la responsabilité du facteur quant au nombre de cas qu'il provoque. Le *RR* est un nombre pur, le *RA (ou DR)* s'exprime dans les mêmes unités que R_1 et R_0.

Étude cas-témoins (rapport des cotes)

Le tableau 6-3 présente, au plan formel, les résultats d'une étude cas-témoins.

Contrairement aux études de cohorte, il est plus difficile, dans les études cas-témoins, de calculer directement les mesures de risque R_1 et R_0. En effet, le rapport du nombre de cas M_1 au nombre de témoins M_0 relève en grande partie d'une décision du chercheur. Par conséquent, les proportions *a / (a* + c) et *bl (b + d)* sont d'une certaine façon arbitraires et ne représentent pas en général les mesures de risque R_1 et R_0. Pour s'en convaincre, voici un exemple.

Considérons une étude où l'on compare 200 cas de maladie *M* à 200 témoins, pour la présence du facteur E. Le tableau 6-4 décrit pour chaque groupe, cas et témoins, la répartition des sujets exposés et des sujets non-exposés.

Tableau 6-3

	E		
	+	−	Total
Cas	*a*	*b*	M_1
Témoins	*c*	*d*	M_0

Tableau 6-4

	E		
	+	−	Total
Cas	80	120	200
Témoins	50	150	200

Considérons une seconde étude cas-témoins identique à la première en tout sauf qu'elle utilise deux fois plus de témoins. Le tableau 6-5 résume les données de cette étude.

Selon que l'on utilise le tableau 6-4 ou le tableau 6-5, on obtient des résultats différents pour les proportions *a/(a + c) et b/(b + d).*

Au tableau 6-4, *a / (a* + c) = $^{80}/_{130}$ et *bl(b + d)* = $^{120}/_{270}$.

Au tableau 6-5, a/(a + c) = $^{80}/_{180}$ et b/ *(b + d)* = $^{120}/_{420}$.

Ces proportions, qui changent selon le rapport entre le nombre de cas et le nombre de témoins, suffisent à démontrer qu'elles ne peuvent correctement évaluer les risques R_1 et R_0. Donc, dans les études cas-témoins, la décision toute méthodologique de choisir des groupes de taille relative plus ou moins grande, influence directement les proportions a/(a + *c) et b/(b + d).*

À moins de connaître le risque total dans la population ou la fraction d'échantillonnage des témoins, les études cas-témoins ne fournissent pas des mesures correctes de R_1 et R_0; elles ne permettent pas non plus le calcul direct du *DR* et du *RR. Nous* verrons dans une section ultérieure comment on peut estimer, dans certaines conditions, les mesures du risque dans une étude cas-témoins. Pour le moment, nous définissons une mesure d'association de calcul très simple qui peut s'avérer une très bonne estimation du *RR*. C'est la mesure connue sous l'appellation anglaise de « *odds ratio* » et que l'on a traduit par *rapport des cotes*.

Tableau 6-5

	E		
	+	−	Total
Cas	80	120	200
Témoins	100	300	400

Au tableau 6-3, parmi les M_1 cas, a ont été exposés et b ne l'ont pas été. La proportion a/M_1 représente, en quelque sorte, la « chance » pour un cas d'avoir été exposé et b/M_1 celle de ne pas l'avoir été. Le rapport $\frac{a/M_1}{b/M_1}$ (ou $\frac{a}{b}$ après simplification) traduit la « chance » relative pour un cas d'avoir été exposé contre celle de ne pas l'avoir été. On peut dire que le rapport $\frac{a}{b}$ mesure la cote en faveur de l'exposition contre la non-exposition. Une cote égale à 2 indique que la « chance » d'avoir été exposé est deux fois plus grande que celle de ne pas l'avoir été. Du même tableau, on peut tirer que la cote d'exposition chez les témoins est égale à $\frac{c}{d}$.

Le rapport $\frac{a/b}{c/d}$ compare les cotes d'exposition entre les cas et les témoins. Ce rapport des cotes, noté *RC*, peut s'écrire $\frac{ad}{bc}$.

$$RC = \frac{a/b}{c/d} = \frac{ad}{bc}.$$

S'il n'y a pas d'association entre le facteur et la maladie, on doit s'attendre d'observer chez les deux groupes, cas et témoins, des cotes d'exposition semblables, sinon identiques: $a/b = c/d$, ce qui conduit à un rapport des cotes égal à 1. Si, par contre, il y a association positive entre le facteur et l'exposition, on doit s'attendre de trouver une cote

d'exposition plus forte chez les cas que chez les témoins, ce qui correspond à un rapport des cotes plus grand que 1:

$$RC = \frac{ad}{bc} > 1$$

Au tableau 6-4, les cotes d'exposition respectivement chez les cas et les témoins sont de $^2/_3$ ($^{80}/_{120}$) et $^1/_3$ ($^{50}/_{150}$). Le rapport des cotes est alors:

$$RC = \frac{^{80}/_{120}}{^{50}/_{150}} = \frac{80 \times 150}{50 \times 120} = 2$$

Au tableau 6-5,

$$RC = \frac{^{80}/_{120}}{^{100}/_{300}} = \frac{80 \times 300}{100 \times 120} = 2$$

Suivant certaines situations que nous résumons ci-dessous, la mesure d'association *RC* est ou bien la même mesure que le *RR* des études de cohorte, ou bien généralement une bonne estimation de celui-ci. Pour chacune des trois situations décrites nous supposons que le groupe témoin est un échantillon aléatoire de la population. On trouve dans l'annexe de ce chapitre une description plus formelle de ces trois situations qui permettent d'établir la correspondance entre le *RC* et le *RR*.

Situation 1. Si l'étude cas-témoins est menée auprès des cas incidents dans une population ouverte et stable considérée pour une période donnée, alors le *RC* calculé dans l'étude cas-témoins est la même mesure que le *RR* des études de cohorte calculé à partir du rapport des taux d'incidence chez les sujets exposés et chez les sujets non-exposés.

Situation 2. Si l'étude cas-témoins est faite sur les cas prévalents, si la population est ouverte et stable, si la maladie est en état d'équilibre et si la durée moyenne de la maladie chez les sujets exposés est la même que celle des sujets non exposés, alors le *RC* est égal au *RR* des études de cohorte calculé à partir du rapport des taux d'incidence.

Situation 3. Si l'étude cas-témoins est menée auprès des cas incidents cumulés en cours d'une période déterminée dans une population fermée, le *RC* est une surestimation du *RIc* calculé dans cette population pour la même période. Si la maladie est rare, le *RC* peut être considéré comme une bonne estimation du *Rlc.*

Pour terminer, soulignons que le rapport des cotes *RC* peut être utilisé comme mesure d'association dans les études de cohorte avec incidence cumulative ou dans les études transversales. Pour les premières, le *RC* se comporte face au *RIc* de même façon décrite dans la situation 3 plus haut. Pour les études transversales, le *RC* est une bonne estimation du rapport des taux d'incidence *RTi* s'il est calculé à partir de cas prévalents comme cela est décrit pour les études cas-témoins dans la situation 2. S'il est calculé à partir de cas prévalents cumulés dans une population fermée, il est alors une surestimation du *RIc* calculé dans cette même population fermée, considérée pour la période dans laquelle ont été cumulés les cas.

MESURE DE CORRÉLATION ENTRE DEUX VARIABLES QUANTITATIVES

La corrélation ou l'association dont il est question ici concerne la liaison entre des variables quantitatives. Nous n'évoquerons

dans ce chapitre que la corrélation entre deux variables. C'est la corrélation simple.

Pour qui regarde la corrélation entre deux variables quantitatives, il est intéressant d'examiner la forme, le sens et la force de la liaison entre ces deux variables. Quelle forme, quel sens, quelle force a la liaison entre la tension artérielle systolique et l'âge des individus, entre le poids et la taille des adultes, entre la proportion de médecins par région et le taux des interventions chirurgicales qui y sont pratiquées? Nous nous limiterons aux liaisons de forme linéaire ou linéarisable par transformation des variables.

La forme linéaire d'une liaison est reconnaissable à l'examen du diagramme de dispersion des données.

Tableau 6-6

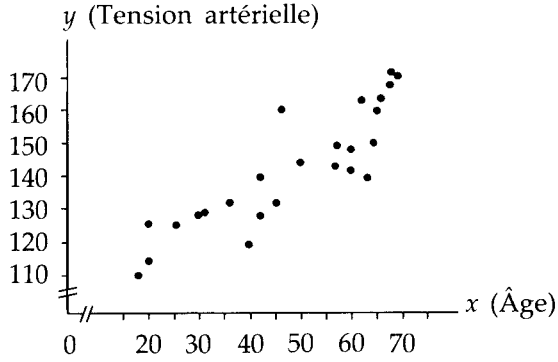
Individu	Âge en année	Tension artérielle en mmHg
i	x_i	y_i
1	25	125
2	62	163
3	18	110
4	65	160
5	42	128
6	64	150
7	45	132
8	60	142
9	60	148
10	20	114
11	63	140
12	30	128
13	46	160
14	68	172
15	20	125
16	68	168
17	57	143
18	69	171
19	57	150
20	35	112
21	42	140
22	40	120
23	50	144
24	36	132
25	66	164

Diagramme de dispersion

Supposons que l'on ait noté l'âge (x) et mesuré la tension artérielle systolique (y) de vingt-cinq individus. Les résultats apparaissent au tableau 6-6.

Ce tableau se présente comme un certain fouilli (numérique) qui laisse difficilement paraître la forme que peut prendre la liaison entre la tension artérielle et l'âge. Il est facile de construire un graphique qui en dévoilera la forme. Il suffit de représenter chaque couple observé de valeurs (x_i, y_i) par un point dans le plan cartésien. A chaque individu correspond ainsi un point dans le plan. La figure 6-1 reproduit l'ensemble des 25 points du tableau 6-6. Cette figure porte le nom de nuage de points ou de *diagramme de dispersion.*

Figure 6-1



Outre qu'il permet d'observer que les individus plus âgés ont en moyenne une tension artérielle systolique plus élevée, le diagramme de dispersion de la figure 6-1 révèle une liaison de forme linéaire. On peut imaginer que l'information contenue dans ce nuage de points peut être résumée par une droite qui le traverse. Il existe des formes non linéaires, par exemple paraboliques, comme à la figure 6-2.

D'autres formes, en apparence non linéaires, peuvent être linéarisées par transformation. C'est le cas de la forme exponentielle. En prenant le logarithme de y, le nuage de points (x_i, y_i) de forme exponentielle à la figure 6-3A est transformé à la figure 6-3B en nuage de points $(x_i, \log y_i)$ de forme linéaire.

En plus de la forme, le diagramme de dispersion permet d'apprécier le sens de la tendance entre y et x. Il y a tendance positive ou ascendante lorsque y augmente en moyenne lorsque x augmente. Si par contre, y diminue en moyenne lorsque x augmente, alors la tendance est négative ou descendante. A la figure 6-1, la liaison entre la tension artérielle et l'âge est de tendance positive. On observe une tendance négative à la figure 6-4 entre la quantité d'œstriol (x) chez la

Figure 6-2

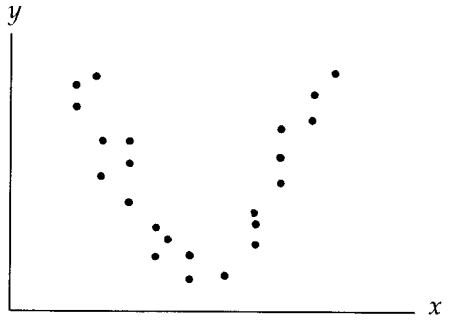


Figure 6-3A

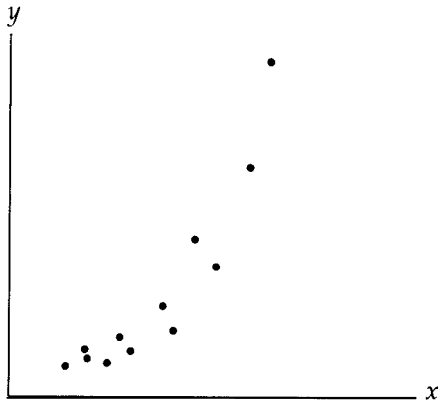


Figure 6-3B

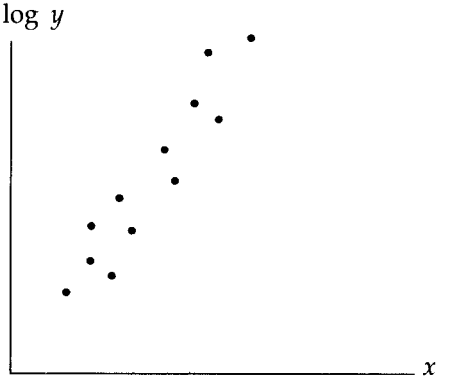
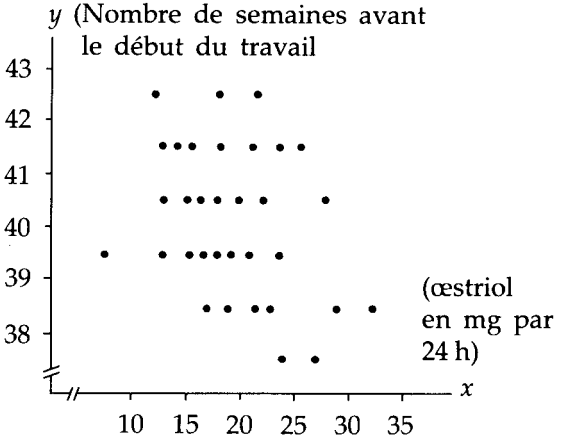


Figure 6-4



femme enceinte et le nombre de semaines (y) avant le début du travail. Bien que la liaison ne paraît pas très forte, il semble toutefois qu'une plus grande quantité d'œstriol corresponde en moyenne à un nombre plus faible de semaines avant le début du travail.

Corrélation linéaire

Le diagramme de dispersion ne permet pas d'évaluer de manière tout à fait satisfaisante, l'intensité de la liaison entre deux variables quantitatives. Pour en avoir une bonne idée numérique, nous allons définir une mesure de corrélation uniquement pour le cas des nuages de forme linéaire. C'est le coefficient de corrélation linéaire, habituellement dénoté par la lettre minuscule *r*, souvent appelé coefficient de Pearson, parfois de Bravais-Pearson. Pour mieux en comprendre la définition, il est préférable d'abord de définir la covariance.

COVARIANCE

Soit n couples de valeurs:

x_i	y_i
x_1	y_1
x_2	y_2
.	.
.	.
.	.
x_n	y_n

Chacune des deux variables, x et y, a sa variance, respectivement s_x^2 et s_y^2. Elles peuvent s'écrire:

$$s_x^2 = \text{Var}(x) = \frac{\Sigma (x_i - \bar{x})(x_i - \bar{x})}{n - 1}$$

$$s_y^2 = \text{Var}(y) = \frac{\Sigma (y_i - \bar{y})(y_i - \bar{y})}{n - 1}$$

Ces variances mesurent la dispersion des variables x et y prises séparément. Il découle assez naturellement que l'expression

$$\frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

puisse être vue comme une mesure de la dispersion entre les deux variables x et y prises simultanément. Cette expression est une sorte de variance, plus spécifiquement de covariance entre x et y.

$$\text{Cov}(x, y) = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

La *covariance* est une sorte de moyenne du produit des écarts à la moyenne arithmétique de chacune des deux variables. La covariance est une mesure intéressante puisqu'elle distingue par son signe le nuage de tendance positive de celui de tendance négative. Pour apprécier cette propriété de la covariance, considérons les nuages de points des figures 6-5A, 6-5B et 6-5C.

À la figure 6-5A, y tend à croître lorsque x croît. Les écarts $x_i - \bar{x}$ et $y_i - \bar{y}$ ont tendance à être du même signe. Le produit de ces écarts est le plus souvent positif et, en définitive, la covariance est positive. S'il n'y a pas de liaison linéaire, comme à la figure 6-5B, un écart positif $x_i - \bar{x}$ est aussi souvent associé à un écart $y_i - \bar{y}$ positif que négatif. La somme des produits de ces écarts tend vers zéro, ce qui entraîne une covariance qui s'approche de zéro. À la figure 6-5C, y tend à décroître lorsque x croît. Les écarts $x_i - \bar{x}$ et $y_i - \bar{y}$ ont tendance à être de signes contraires. Le produit de ces écarts est le plus souvent négatif et, en définitive, la covariance est négative. Va pour le sens de la liaison.

COEFFICIENT DE CORRÉLATION LINÉAIRE

La covariance a le grand inconvénient de dépendre des unités de mesure. La covariance entre la tension artérielle en mmHg et l'âge en années s'exprime en mmHg-années. Une unité de mesure plutôt ennuyeuse... Pour conserver les avantages de la covariance mais sans cet inconvénient, il suffit de la diviser par s_x et s_y. Les unités de mesure de s_x et s_y réunis étant celles de la covariance, la division proposée annule les unités. La nouvelle mesure obtenue est un nombre pur qu'on appelle le *coefficient de corrélation linéaire r.*

D'où $$r = \frac{\text{Cov}\,(x,\ y)}{s_x s_y},$$

ce qui peut aussi s'écrire

$$r = \frac{\Sigma\,(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\Sigma\,(x_i - \bar{x})^2\,\Sigma\,(y_i - \bar{y})^2}}$$

Si l'on se réfère à l'exemple du tableau 6-2 entre la tension artérielle et l'âge, on trouve r = 0,87.

Nous résumons maintenant les principales propriétés de r.

- Le coefficient de corrélation linéaire r est un nombre pur, sans dimension, indépendant des unités de mesure des deux variables. La comparaison de deux coefficients de corrélation est alors rendue plus facile.
- On peut démontrer que le coefficient de corrélation linéaire r prend des valeurs nécessairement comprises entre -1 et 1.

$$-1 \leq r \leq 1$$

Figure 6-5A

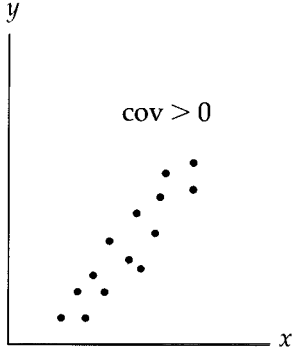


Figure 6-5B

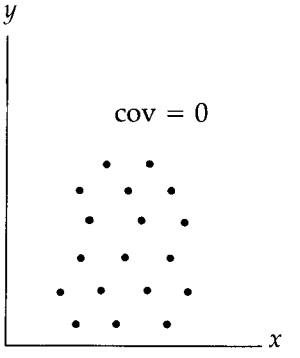
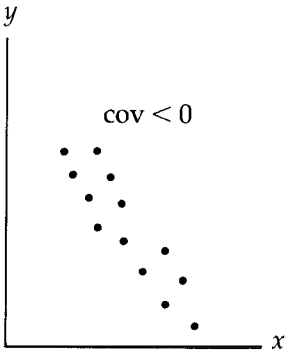


Figure 6-5C



Les valeurs r = -1 et r = 1 indiquent une liaison linéaire parfaite ou stricte comme celles que l'on trouve dans les figures 6-6A et 6-6B. De telles liaisons ne se rencontrent jamais (ou presque) dans les champs d'étude qui relèvent de l'épidémiologie ou de la santé publique ou communautaire. Les figures 6-6C et 6-6D correspondent à des liaisons plus lâches. La figure 6-6E illustre l'absence de liaison.

- Pour terminer, il est important d'insister sur le fait que r est une mesure de l'intensité de liaison linéaire. Si la corrélation est non-linéaire, la valeur de r peut être nulle, alors qu'il existe une parfaite corrélation entre x et y. Le diagramme de dispersion non-linéaire (parabolique) de la figure 6-7 le montre clairement. Il y a une parfaite corrélation et la covariance est nulle (et $r = 0$). Chaque terme $(x_i - \bar{x})(y_i - \bar{y})$ est annulé par son symétrique (de signe opposé). En effet, les valeurs $(x_i - \bar{x})(y_i - \bar{y})$ qui correspondent respectivement aux points A et B à la figure 6-7 s'annulent. Et ainsi de suite pour tous les points et leur symétrique, ce qui rend nulle la somme: $\Sigma (x_i - \bar{x})(y_i - \bar{y})$. Donc, $r = 0$.

Pour conserver la belle correspondance entre les valeurs de r qui passent de -1 à 0, puis de 0 à 1 avec la configuration d'une liaison négative parfaite vers un nuage sans liaison, ensuite vers une liaison positive par-

Figure 6-6

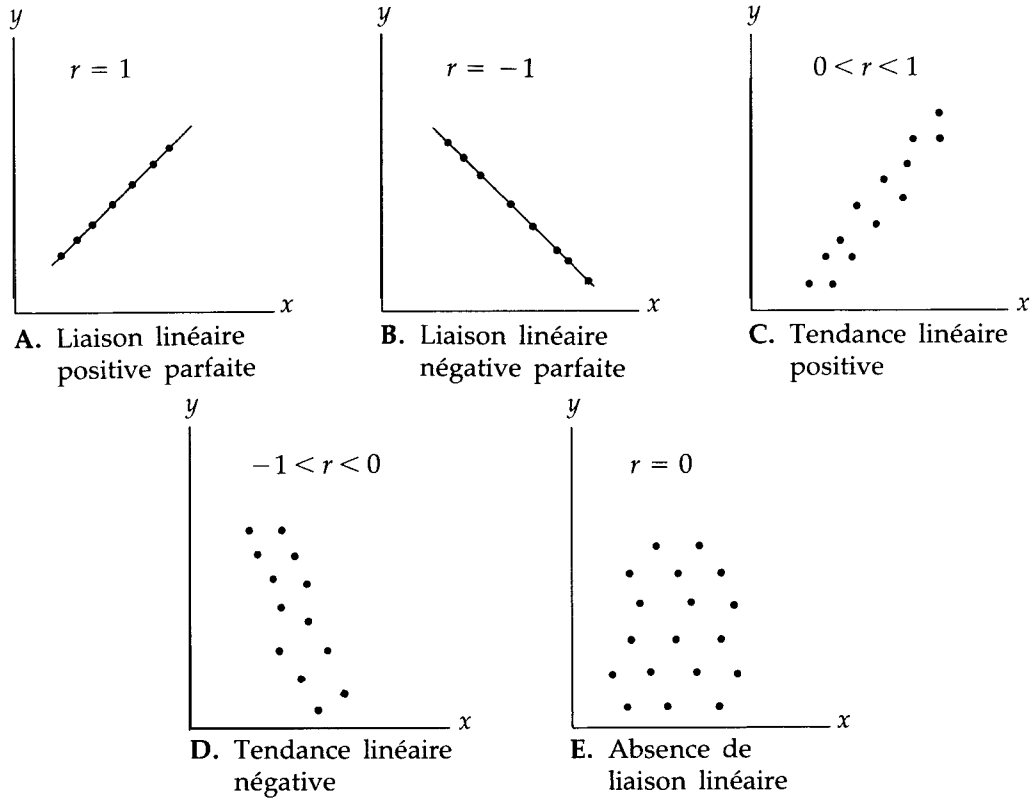
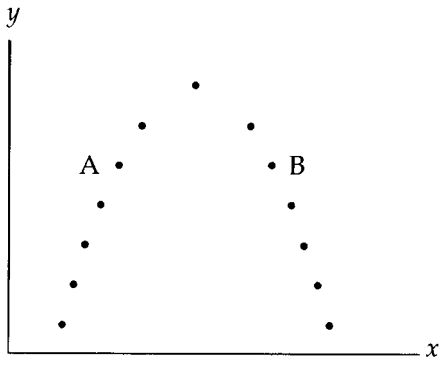


Figure 6-7



faite, il est nécessaire de comprendre le coefficient de corrélation *r* comme un coefficient de corrélation linéaire.

MESURES D'ASSOCIATION ET CAUSALITÉ

Selon l'état des connaissances actuelles, et en supposant qu'une seule particule virale viable soit suffisante pour causer la maladie, un humain non-vacciné, infecté par la morsure d'un animal enragé, développe nécessairement la rage. En revanche, si on observe la rage chez un humain, celui-ci a été infecté par le virus de la rage. Le lien entre l'infection par suite d'une morsure et la rage chez l'humain apparaît comme un lien de causalité de nature déterministe. Si la cause est présente, l'effet suit. Réciproquement, si on observe l'effet, la cause est présente au départ.

Généralement, la situation est différente en médecine. Le lien de causalité est plutôt de nature probabiliste. Le fait de fumer pour une personne n'entraîne pas nécessairement un cancer du poumon. Un non-fumeur peut, par ailleurs, développer un cancer du poumon. Le fait de fumer pour une personne augmente son risque d'être affectée par un cancer du poumon. Les mesures d'association comme le *RR* ou le *RC,* la mesure de corrélation linéaire *r,* n'indiquent pas nécessairement une relation de cause à effet. Elles témoignent plutôt d'une relation statistique.

Avant qu'une association observée entre un facteur et une maladie ne soit déclarée causale, certaines précautions doivent être prises pour établir un tel jugement. Il faut d'abord s'assurer qu'il n'existe dans l'étude aucun défaut d'importance, aucun biais qui puisse expliquer cette association. Une fois résolue la question des biais et celle de l'association statistique, peut-on admettre qu'il y a toujours un lien de causalité entre le facteur et la maladie? Pas nécessairement. Il est possible de mettre en évidence une association statistique entre la présence de doigts jaunis et le cancer du poumon. Aucun doute que les doigts jaunis ne sont pas la cause de ce type de cancer.

Pour qu'un investigateur passe de l'association statistique à la causalité, il peut faire appel à un certain nombre de critères qui vont le guider dans son jugement de causalité. Voici les principaux:

Constance de l'association observée. Des études conduites en des moments et des lieux différents sur d'autres populations ont produit des résultats semblables.

Intensité de l'association observée. L'action des biais sur l'association se fait vraisemblablement moins sentir lorsque l'intensité ou la force de l'association est grande. En ce sens, une plus forte intensité favorise le jugement de causalité.

Spécificité de l'association. Plus un facteur est exclusif par rapport à une maladie, plus l'interprétation causale est plausible. L'absence

de spécificité ne doit cependant pas être perçue comme un rejet de la causalité, étant donné que les facteurs sont souvent à effets multiples et que les maladies ont des causes multiples.

Cohérence chronologique. La cause doit précéder l'effet.

Présence d'une relation dose-effet. L'effet augmente lorsque la dose augmente.

Cohérence avec les connaissances bio-médicales. L'association observée ou les résultats obtenus doivent s'harmoniser avec l'histoire naturelle connue de la maladie et être en accord ou en accord possible avec les connaissances médicales. L'association observée doit être biologiquement plausible. Si la relation suggérée entre le facteur et la maladie s'accorde avec les connaissances médicales courantes, on accepte plus aisément l'interprétation causale.

Observer la même association chez l'animal, dans le cadre de l'expérimentation, est un appui important à une présomption de causalité.

En définitive, déclarer à partir de ces critères qu'un facteur est causal ne signifie pas qu'il y a preuve irréfutable de causalité, mais qu'il y a forte présomption en faveur de celle-ci.

RÉSUMÉ

Dans les études de cohorte, la différence des risques *(DR)* et le risque relatif *(RR)* sont les deux mesures que l'on utilise pour évaluer l'association entre un facteur et une maladie (décès compris). Dans le cas d'un lien de causalité entre un facteur et une maladie, le terme « risque attribuable

(RA) est utilisé de préférence à « différence des risques » *(DR)*. Le *RA* décrit la responsabilité du facteur quant au nombre de cas qu'il induit, tandis que le risque relatif *RR* mesure l'association sur le plan étiologique. Dans les études cas-témoins, la mesure d'association utilisée est le rapport des cotes (RC). En général, il est une bonne estimation du risque relatif. L'intensité de la liaison linéaire entre deux variables quantitatives est mesurée par le coefficient de corrélation (linéaire) *r*. Il prend des valeurs comprises entre -1 et +1. Ces deux valeurs extrêmes indiquent une liaison linéaire parfaite entre les deux variables. La valeur 0 pour *r* correspond à l'absence de liaison linéaire. Le signe de r indique le sens de la liaison.

Symboles

N_1, N_0: nombre de personnes (personnes-années) exposées, non-exposées

R_1, R_0: mesure du risque de maladie ou de décès chez les sujets exposés, chez les sujets non-exposés

DR : différence de deux risques *(RA:* risque attribuable)

DTi, DT_D, DIc: différence entre deux taux d'incidence, entre deux taux de mortalité, entre deux incidences cumulatives

RR: rapport de deux risques ou risque relatif

RTi, RT_D, RIc: rapport de deux taux d'incidence, de deux taux de mortalité, de deux incidences cumulatives

RC: rapport des cotes

s_x^2, s_y^2: variance de x, de y

$\text{cov}(x, y)$: covariance entre x et y

r: coefficient de corrélation linéaire

Formules

$$R_1 = a/N_1$$

$$R_0 = b/N_0$$

$$DR = R_1 - R_0;\ (RA = R_1 - R_0)$$

$$RR = \frac{R_1}{R_0}$$

$$RC = \frac{ad}{bc}$$

$$\text{cov}(x, y) = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$$r = \frac{\text{Cov}(x, y)}{s_x s_y} = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma (x_i - \bar{x})^2 \Sigma (y_i - \bar{y})^2}}$$

LECTURES SUGGÉRÉES

1. GOLDBERG, M. ***L'Épidémiologie sans peine,*** Paris, Editions médicales Roland Bettex, 1986, pp. 63-67.

2. KLEINBAUM, D.G., KUPPER, L.L. et MORGENSTERN, H. ***Epidemiologic Research,*** Belmont (USA), Lifetime Learning Publications, 1982, chapitre 8, pp. 140-158.

Sur la causalité

3. FLETCHER, R.H., FLETCHER, S.W. et WAGNER, E.H. ***Clinical Epidemiology,*** Baltimore, Williams and Wilkins, 1982, chapitre 11, pp. 185-202.

4. HOEY, J. et LAMBERT, *R.* ***Éléments d'épidémioIogie pour Ie clinicien,*** Paris, Editions du C.N.R.S., 1981, chapitre 8, pp. 139-149.

ANNEXE DU CHAPITRE 6

Correspondance entre le *RC* et le *RR* suivant différentes situations

Situation 1

Considérons une population ouverte et stable pour une période déterminée de durée Δt. Supposons que M_1 désigne le nombre de cas incidents de la maladie pour la période et qu'à chaque moment, depuis le début de la période d'induction de la maladie, la population des N individus à risque se divise en N_1 sujets exposés et en N_0 sujets non-exposés. Alors on a le tableau A6-1. Les majuscules A et B soulignent le fait qu'il s'agit de cas de maladie (ou de décès) au niveau d'une population.

Nous rappelons que

$$Ti_1 = \frac{A}{N_1 \triangle t}\ ,\ Ti_0 = \frac{B}{N_0 \triangle t}\ \text{ et }\ RTi = \frac{AN_0}{BN_1}.$$

Dans cette population, un investigateur décide de conduire une étude cas-témoins en utilisant les M_1 cas incidents (ou un échantillon de ceux-ci). Ainsi, il prélève:

—un échantillon aléatoire de cas, f_1M_1, *où f,* est une fraction d'échantillonnage; (si tous les cas sont pris, alors *f,* est égal à 1);

—un échantillon aléatoire de la population à risque, f_0N où f_0 est la fraction d'échantillonnage pour les témoins.

Les résultats de cette étude sont décrits dans le tableau A6-2.

Les échantillons de cas et de témoins étant aléatoires, on peut supposer que les répartitions de l'exposition dans ces échantillons choisis sont les mêmes que celles qui prévalent respectivement dans la population des cas et celle des individus à risque. Pour cette raison, on doit retrouver dans nos échantillons f_1A cas exposés contre f_1B cas non-exposés, $f_0\ N_1$ témoins exposés contre $f_0\ N_0$ témoins non-exposés. Ainsi,

$$a = f_1A$$

$$b = f_1B$$

$$c = f_0N_1$$

$$d = f_0N_0$$

Le tableau A6-2 prend donc la forme du tableau A6-3.

Le rapport des cotes RC se calcule donc comme:

$$RC = ad/bc = \frac{f_1A\ f_0N_0}{f_1B\ f_0N_1} = \frac{AN_0}{BN_1} = RTi.$$

Ainsi, le *RC* obtenu dans une étude cas-témoins menée auprès de cas incidents d'une population ouverte stable est la même mesure

Tableau A6-1

	E		
	+		Total
Cas incidents	*A*	*B*	M_1
Personnes	N_1	N_0	*N*

Tableau A6-2

	E		
	+	−	Total
Cas	f_1A	f_1B	f_1M_1
Témoins	f_0N_1	f_0N_0	f_0N

Tableau A6-3

	E		
	+	-	Total
Cas	*a*	*b*	f_1M_1
Témoins	c	*d*	f_0N

que le rapport des taux d'incidence *(RTi)* obtenu dans cette même population. Par exemple, considérons une certaine population ouverte stable de 1000 000 d'individus. On sait, à l'insu des chercheurs, que 40 % des personnes sont exposées au facteur *E,* que les taux d'incidence de la maladie M sont respectivement chez les sujets exposés et les sujets non-exposés de 2 et 1 cas pour 10 000 personnes-années. Au cours d'une période de deux ans, on a observé 280 nouveaux cas de la maladie M. La situation est résumée dans le tableau A6-4.

Les chercheurs qui s'intéressent à l'association qui peut exister entre le facteur *E* et la maladie *M,* décident de faire une étude cas-témoins en utilisant les *280* cas incidents ($f_1 = 1$). Ils les comparent à un échantillon aléatoire de 280 témoins ($f_0 = {}^{280}/_{100\ 0000}$). Les résultats de l'étude sont présentés dans le tableau A6-5.

Le *RC* calculé dans cette étude a exactement la même valeur que le *RTi* tel que l'on puisse le calculer dans une étude de cohorte menée dans cette population

$$RC = RTi = 2.$$

Tableau A6-4

	E		
	+	−	Total
Cas incidents	160	120	280
Personnes-années	800 000	1200 000	2000 000

$$RTi = \frac{160 \times 1200\,000}{120 \times 800\,000} = 2$$

Situation 2

Considérons, à un moment déterminé, une population ouverte et stable. À ce moment, il y a dans cette population N individus non-malades dont $N1$ sont exposés au facteur E et N_0 qui ne le sont pas. De même, il y a M_1 cas prévalents de la maladie M dont A sont exposés et B qui ne le sont pas. La situation se présente comme au tableau A6-6.

Si on suppose, en plus, que la maladie est en situation d'équilibre dans la population, alors on a:

$$\frac{Pr_1}{(1 - Pr_1)} = Ti_1 \times \bar{d}_1 \quad \text{et} \quad \frac{Pr_0}{(1 - Pr_0)} = Ti_0 \times \bar{d}_0.$$

D'où

$$\frac{[Pr_1/(1 - Pr_1)]}{[Pr_0/(1 - Pr_0)]} = \frac{Ti_1 \times \bar{d}_1}{Ti_0 \times \bar{d}_0}$$

$$= RTi\ \frac{\bar{d}_1}{\bar{d}_0}$$

Tableau A6-5

	E		
	+	−	Total
Cas	160	120	280
Témoins	112	168	280

$$RC = \frac{160 \times 168}{112 \times 120} = 2$$

Tableau A6-6

	E		
	+	−	Total
Cas prévalents	A	B	M_1
Non-cas	N_1	N_0	N

Si $\bar{d}_1 = \bar{d}_0$,

alors

$$\frac{[Pr_1/(1 - Pr_1)]}{[Pr_0/(1 - Pr_0)]} = RTi.$$

Puisque, par ailleurs

$$\frac{[Pr_1/(1 - Pr_1)]}{[Pr_0/(1 - Pr_0)]} = \frac{A/N_1}{B/N_0} = \frac{AN_0}{BN_1}$$

on a donc,

$$RTi = \frac{AN_0}{BN_1}$$

Un investigateur décide de mener une étude cas-témoins dans cette population en utilisant les cas prévalents ou un échantillon de ceux-ci.

Alors, il prélève

—un échantillon de cas prévalents, f_1M_1 ;

—un échantillon de non-cas, f_0N.

On obtient alors le tableau A6-7.

Tableau A6-7

	E		
	+	−	Total
Cas	f_1A	f_1B	f_1M_1
Non-cas	f_0N_1	f_0N_0	f_0N

ou encore le tableau A6-8 équivalent:

Tableau A6-8

	E		
	+	−	Total
Cas	a	b	f_1M_1
Témoins	c	d	f_0N

On a donc

$$RC = \frac{ad}{bc} = \frac{f_1A\, f_0N_0}{f_1B\, f_0N_1} = \frac{AN_0}{BN_1} = RTi.$$

Ainsi, le *RC* obtenu dans une étude cas-témoins menée auprès de cas prévalents d'une population ouverte et stable, la maladie en état d'équilibre, la durée moyenne de la maladie égale entre les sujets exposés et non-exposés est la même mesure que le rapport des taux d'incidence *(RTi)* obtenu dans cette même population.

Considérons l'exemple suivant. Dans une population ouverte mais stable, il y a à tout moment 875 cas prévalents d'une maladie M et 100 000 individus non-malades mais à risque de l'être. On sait, à l'insu des chercheurs, que 40 % des individus à risque ont été exposés au facteur *E* ainsi que 500 des 875 cas. La situation est résumée dans le tableau A6-9.

Si on suppose que la durée moyenne de la maladie M est la même chez les sujets exposés et chez les sujets non-exposés, alors

$$RTi = \frac{AN_0}{BN_1} = \frac{500 \times 60\,000}{375 \times 40\,000} = 2.$$

Un investigateur décide de faire une étude castémoins en utilisant les 875 cas prévalents ($f_1 = 1$). Il les compare, pour la présence du facteur *E,* à un échantillon aléatoire de 875

Tableau A6-9

	E		
	+	−	Total
Cas prévalents	500	375	875
Population à risque	40 000	60 000	100 000

témoins ($f_0 = {}^{875}/_{100\ 000}$). Les résultats de cette étude sont présentés dans le tableau A6-10.

Le rapport des cotes *RC* estimé dans cette étude est de 2 aussi.

$$RC = \frac{500 \times 525}{375 \times 350} = 2.$$

Situation 3

Considérons maintenant une population fermée de *N* individus pour une période déterminée de durée Δt. Supposons que cette population est saine au début de la période et qu'elle se partage en N_1 sujets exposés et N_0 sujets non-exposés.

En cours de période, il y a eu M_1 cas incidents cumulés. À la fin de la période, la population se présente comme au tableau A6-11.

Tableau A6-10

	E		
	+	−	Total
Cas	500	375	875
Témoins	350	525	875

Tableau A6-11

	E		
	+	−	Total
Cas incidents	A	B	M_1
Sujets non-affectés sur la période	$N_1 - A$	$N_0 - B$	M_0
Total	N_1	N_0	N

Nous rappelons que le rapport des incidences cumulatives est donné par:

$$RIc = \frac{AN_0}{BN_1}.$$

Un investigateur décide de conduire sur cette population une étude cas-témoins en utilisant les M_1 cas incidents ou un échantillon de ceux-ci. Alors, il prélève à la fin de la période

— un échantillon de f_1M_1 cas;

— un échantillon de f_0M_0 non-cas.

On obtient alors le tableau A6-12.

Tableau A6-12

	E		
	+	−	Total
Cas	f_1A	f_1B	f_1M_1
Non-cas	$f_0(N_1 - A)$	$f_0(N_0 - B)$	f_0M_0

ou encore le tableau A6-13 équivalent:

Tableau A6-13

	E		
	+	−	Total
Cas	a	b	f_1M_1
Témoins	c	d	f_0M_0

On a donc:

$$RC = \frac{ad}{bc} = \frac{f_1A\,f_0(N_0 - B)}{f_1B\,f_0(N_1 - A)}.$$

Si A est négligeable par rapport à N_1 et B par rapport à N_0, en d'autres termes si la maladie est rare, on a

$$RC \simeq AN_0/BN_1.$$

Ainsi, une étude cas-témoins sur des données d'incidence cumulative permet d'estimer le *RIc* à condition que la maladie soit rare.

Considérons l'exemple suivant. Une population fermée de 10 000 individus, non-affectés par la maladie *M* au début d'une période d'observation, enregistre 840 cas incidents de la maladie *M* sur cette période de deux ans. A l'insu des chercheurs, on sait que 40 % des individus ont été exposés à un facteur E, et que les incidences cumulatives à deux ans sont respectivement chez les sujets exposés et les sujets non-exposés de 12 et de 6 cas pour 100 individus. La situation de cette population à la fin de la période de deux ans est décrite dans le tableau A6-14.

Les chercheurs décident de conduire une étude cas-témoins en utilisant un échantillon aléatoire de 420 cas incidents (*f, = 0,50).* Ils les comparent à un échantillon aléatoire de *458* témoins ($f_0 = 0,05$) tiré du groupe des 9160 non-malades. Les résultats de cette étude sont résumés dans le tableau A6-15.

Le *RC* calculé dans cette étude cas-témoins est de *2,14,* alors que le *Rlc* réel est de *2,0.* Dans ces conditions, le *RC* est donc une surestimation du *Rlc.* La surestimation est d'autant plus grande que la fréquence de la maladie est plus importante, c'est-à-dire que les incidences cumulatives sont plus élevées.

Tableau A6-14

	E +	*E* −	Total
Malades	480	360	840
Non-malades	3520	5640	9 160
Total	4000	6000	10 000

$$RIc = \frac{480 \times 6000}{360 \times 4000} = 2.$$

Tableau A6-15

	E +	*E* −	Total
Cas	240	180	420
Témoins	176	282	458

$$RC = \frac{240 \times 282}{176 \times 180} = 2,14$$

CHAPITRE 7
Mesures d’impact

Ce chapitre distingue les mesures d'impact où un facteur a un effet nocif de celles où il a un effet préventif. Il y a donc, dans l'ordre, les fractions étiologiques et les fractions prévenues. On définit la fraction étiologique chez les sujets exposés et la fraction étiologique totale, de même la fraction prévenue chez les sujets exposés et la fraction prévenue totale. Pour chacune des quatre fractions, plusieurs formules, équivalentes entre elles, sont présentées.

Les mesures d'impact visent à quantifier l'effet d'un facteur sur la fréquence de la maladie ou du décès. Si le facteur a un effet positif, c'est-à-dire si sa présence est associée à une augmentation de la fréquence de la maladie *(RR* > 1), on parle alors de fraction étiologique. Cette mesure peut être utilisée en santé publique pour reconnaître l'effet d'un facteur nocif sur la santé d'une population. Si le facteur a un effet négatif, c'est-à-dire si sa présence est associée à une diminution de la fréquence de la maladie *(RR* < 1), on parle alors de fraction prévenue ou évitable. Dans ce cas, le facteur joue un rôle protecteur contre la maladie. Cette dernière mesure est souvent utilisée comme mesure d'effet dans les études ou essais comparatifs.

FRACTION ÉTIOLOGIQUE

La *fraction étiologique* mesure, en proportion ou en pourcentage, la responsabilité du facteur nocif *E* sur le nombre de cas de la maladie M (ou de décès par la maladie). La fraction étiologique est formellement le risque attribuable proportionnel et peut être calculée chez les cas exposés. On parle alors de la fraction étiologique chez les sujets exposés et on la note FE_1. Elle peut aussi être calculée chez les cas en totalité, on parle alors de la fraction étiologique totale (ou de population) et on la note FE_t.

Fraction étiologique chez les sujets exposés

La différence des risques $R_1 - R_0$ mesure, dans le contexte causal, la responsabilité du facteur nocif *E.* Supposons que l'on observe, pendant une année, 140 décès par cancer du poumon dans une population de 100 000 fumeurs $(R_1 = {}^{140}/_{100\,000})$. On reconnaît qu'une partie de ce risque est attribuable au facteur cigarette et que la partie restante du risque forme ce que l'on pourrait appeler le risque résiduel, R_0. L'estimation de R_0 ne peut se faire que sur une population de non-fumeurs. Supposons que l'on obtienne un risque R_0 égal à 10 pour 100 000 non-fumeurs. Alors, la différence $R_1 - R_0$ mesure la responsabilité de la cigarette dans le risque de décès par cancer du poumon chez les fumeurs, soit 130 décès par cancer du poumon par 100 000 fumeurs.

Alors que le risque chez les fumeurs est de R_1, la partie attribuable au facteur est de $R_1 - R_0$. On peut donc dire que le rapport $(R_1 - R_0)/R_1$ représente la fraction du risque attribuable au facteur. Ce rapport, noté FE_1, est appelé la *fraction étiologique chez les sujets exposés (ici* chez les fumeurs). Elle mesure donc, parmi les cas exposés, la proportion de ceux-ci attribuable au facteur *E.*

On a la formule

$$FE_1 = \frac{R_1 - R_0}{R_1}$$

ou encore, en divisant le numérateur et le dénominateur par R_0, on obtient l'expression équivalente

$$FE_1 = \frac{RR - 1}{RR}.$$

Pour l'exemple des fumeurs, avec un *RR* égal à 14, on peut donc estimer la fraction étiologique FE_1 à 0,93 ou 93 %. Chez les fumeurs, 93 % des décès par cancer du poumon sont attribuables au facteur cigarette.

La formulation *(RR* — 1)/*RR* est généralement la forme retenue pour l'expression de

la fraction étiologique chez les sujets exposés, car elle est formellement applicable aussi bien aux études de cohorte qu'aux études cas-témoins. Pour ces dernières études, on remplacerait l'expression *(RR — 1)/RR* par *(RC — 1)/RC.*

Le graphique à la figure 7-1 nous montre les variations de FE_1 en fonction du *RR.* Pour un risque relatif égal à 1 *(RR = 1),* la fraction étiologique est égale à zéro ($FE_1 = 0$). Lorsque *RR* augmente, FE_1 augmente aussi et tend à la limite vers 1 ou 100 %.

Une fraction étiologique chez les sujets exposés égale à 0 indique que le facteur n'a aucune responsabilité. Chez les sujets non-exposés, où le facteur n'est pas présent, la fraction étiologique est évidemment nulle.

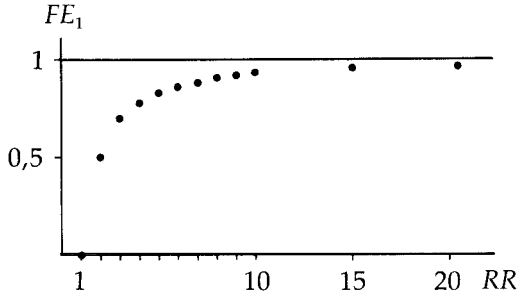
Fraction étiologique totale

Revenons à l'exemple du cancer du poumon associé à la cigarette. Supposons que les fumeurs représentent 40 % de la population totale. Le risque R_t de décès par cancer du poumon dans la population totale est la somme pondérée du risque chez les fumeurs et de celui chez les non-fumeurs:

$$R_t = 0{,}40 \times \frac{140}{100\,000} + 0{,}60 \times \frac{10}{100\,000}$$

$$= \frac{62}{100\,000}.$$

Figure 7-1



D'où la règle générale suivante:

$$R_t = p_1R_1 + p_0R_0$$

où p_i est la proportion de sujets exposés dans la population et $p_0 = 1 — p_1$, la proportion de sujets non-exposés.

Le rapport $(R_t — R_0)/R_t$ représente la fraction du risque total attribuable au facteur dans la population totale. Ce rapport, noté FE_t, est appelé la *fraction étiologique totale ou de population.* Il mesure, parmi tous les cas d'une population, la proportion de ceux dont la condition (de malade) est attribuable au facteur.

On a la formule :

$$FE_t = \frac{R_t - R_0}{R_t}$$

Pour l'exemple des fumeurs, on peut estimer la fraction étiologique totale FE_t à 0,84 ou 84 *%.* Dans la population totale considérée, 84 % des décès par cancer du poumon sont attribuables à la cigarette.

La valeur obtenue pour FE_t est inférieure à celle de FE_1. C'est un fait général. On a toujours l'inégalité:

$$FE_t \leq FE_1.$$

Voici la démonstration.

On a, $R_0 \leq R_t \leq R_1$,

d'où $\frac{R_0}{R_t} \geq \frac{R_0}{R_1}$;

ainsi, $1 - \frac{R_0}{R_t} \leq 1 - \frac{R_0}{R_1}$,

soit, $FE_t \leq FE_1$.

Il existe deux autres formules de la fraction étiologique totale FE_t, chacune équivalente à la formule (définition): $(R_t - R_0)/R_t$. Ces deux formules sont:

$$FE_t = \frac{p_1 (RR - 1)}{p_1 (RR - 1) + 1} \qquad [1]$$

où p_1 est la proportion de sujets exposés dans la population, et

$$FE_t = p_{c1} \times \frac{RR - 1}{RR} \qquad [2]$$

où pc, est la proportion de cas exposés parmi l'ensemble des cas.

Nous donnons ici le développement de ces deux formules et la démonstration de leur équivalence

Développement de ta première formule

Si, dans l'expression $(R_t - R_0)/R_t$, on remplace R_t par $p_1 R_1 + p_0 R_0$ où $p_1 + p_0 = 1$, alors on obtient:

$$FE_t = \frac{p_1 (R_1 - R_0)}{p_1 (R_1 - R_0) + R_0}$$

$$= \frac{p_1 (RR - 1)}{p_1 (RR - 1) + 1} \qquad [1]$$

On vérifie, pour l'exemple des fumeurs,

$$FE_t = \frac{0{,}40\,(14 - 1)}{0{,}40\,(14 - 1) + 1} = 0{,}84$$

Développement de la deuxième formule

Référons-nous à l'ensemble des cas dans la population. Une proportion p_{c1} des cas exposés au facteur et une proportion p_{co} de cas non-exposés, où, bien sûr, $p_{c1} + p_{co} = 1$. Comme on l'a vu, la FE_1 mesure la proportion attribuable au facteur chez les cas exposés seulement. On peut donc concevoir la proportion FE_t comme la somme pondérée des fractions étiologiques chez les exposés FE_t et celle chez les non-exposés FE_0. Les poids dans cette somme sont donnés par p_{c1} et p_{co}.

Ainsi, $FE_t = p_{c1}\ (FE_1) + p_{co}\ (FE_0)$. Mais puisque la fraction étiologique FE_0 chez les cas non-exposés ne peut être que nulle, on a donc:

$$FE_t = p_{C1}\,(FE_1)$$

$$= p_{C1} \times \frac{RR - 1}{RR}. \qquad [2]$$

Démonstration de l'équivalence

Les formules [1] et [2] sont équivalentes. Vérifions d'abord que

$$p_1 = \frac{p_{C1}}{p_{C1} + p_{C0} RR}. \qquad [3]$$

Nous savons que

$$p_1 = \frac{N_1}{N_1 + N_0}$$

$$p_{C1} = \frac{a}{M_1},\ p_{C0} = \frac{b}{M_1}$$

et

$$R_1 = \frac{a}{N_1},\ R_0 = \frac{b}{N_0}$$

Ainsi

$$p_1 = \frac{N_1}{N_1 + N_0}$$

peut s'écrire

$$p_1 = \frac{a/R_1}{a/R_1 + b/R_0}$$

$$= \frac{aR_0}{aR_0 + bR_1}$$

$$= \frac{a}{a + bRR}$$

$$= \frac{a/M_1}{a/M_1 + (b/M_1)RR}$$

$$= \frac{p_{c1}}{p_{c1} + p_{c0}RR}.$$

En substituant cette dernière expression de p_1 dans l'équation [1], on obtient l'équation [2], après quelques manipulations algébriques.

Nous obtenons en définitive trois formules équivalentes pour la fraction étiologique totale FE_t, qui sont, rappelons-le :

$$FE_t = \frac{R_t - R_0}{R_t}$$

$$= \frac{p_1(RR - 1)}{p_1(RR - 1) + 1}$$

$$= p_{c1} \times \frac{RR - 1}{RR}.$$

La dernière formulation : $p_{c1} \times \frac{RR - 1}{RR}$ est intéressante pour les études cas-témoins où elle est facile à calculer. Elle ne requiert que la connaissance du *RR*, donc du *RC*, et de la fraction p_{c1} directement fournie par l'étude cas-témoins si le groupe des cas constitue un échantillon représentatif des cas.

Deux cents cas d'une maladie M ont été comparés, pour la présence du facteur *E*, à un échantillon de 200 témoins tiré de la population générale. Les résultats de l'étude apparaissent au tableau 7-1.

Alors $$p_{c1} = \frac{160}{200} = 0{,}80$$

et $$RC = \frac{160 \times 100}{100 \times 40} = 4$$

On obtient

$$FE_1 = \frac{RC - 1}{RC} = \frac{4 - 1}{4} = 0{,}75$$

et

$$FE_t = p_{c1} \times \frac{RC - 1}{RC} = 0{,}80 \times \frac{4 - 1}{4} = 0{,}60.$$

Dans le groupe des cas exposés, 75 % de ceux-ci sont attribuables au facteur *E*. Dans la population générale, ce pourcentage n'est plus que de 60 %.

FRACTION PRÉVENUE (OU ÉVITABLE)

Supposons que l'on s'intéresse à un facteur *E*, par exemple un vaccin, qui protège contre la maladie M. Dans ce cas, le risque R_1 de la maladie chez les vaccinés est inférieur au risque *R0* des sujets non-vaccinés, c'est-à-dire le risque relatif *RR* a une valeur inférieure à 1. Pour que l'action désirée du vaccin soit décrite par une mesure d'association de valeur supérieure à 1, il serait alors préférable d'utiliser comme mesure l'inverse du *RR*, soit RR^{-1}. *Si* RR^{-1} a une valeur supérieure à 1, alors le facteur a un rôle protecteur. Plus

Tableau 7-1

	E		
	+	−	Total
Cas	160	40	200
Témoins	100	100	200

RR^{-1} est élevé, plus l'action du facteur est déterminante dans la protection contre la maladie *M*. Il est donc justifié de concevoir RR^{-1} comme une sorte de *mesure d'efficacité Ef* du facteur *E à* protéger contre la maladie *M*. On peut écrire :

$$Ef = RR^{-1} = \frac{R_0}{R_1}.$$

Supposons que l'on observe 50 cas de la maladie *M* dans un groupe de 1000 vaccinés contre cette maladie $(R_1 = {}^{50}/_{1000})$ et 200 cas dans un groupe de 1000 sujets non-vaccinés $(R_0 = {}^{200}/_{1000})$. Alors, on peut estimer l'efficacité de ce vaccin à 4, c'est-à-dire que les sujets non-vaccinés risquent quatre fois plus d'être atteints par la maladie *M*.

$$Ef = \frac{R_0}{R_1} = 4.$$

La mesure d'efficacité décrit la force qu'exerce le facteur dans la protection d'un individu contre la maladie.

Un facteur ou une intervention aura beau être efficace, il n'est pas garanti pour autant que son effet de protection contre la maladie soit grand dans le groupe où il est appliqué. Tout dépend, bien entendu, de la proportion des individus qui y sont exposés. Régulièrement, le planificateur d'intervention en santé s'intéresse à la question de l'effet que peut avoir l'intervention. Quel effet a produit ou produira l'intervention E dans la population? En d'autres termes, combien de cas de maladie M ont été ou seront évités à cause de l'intervention? Pour jauger cet effet, le planificateur utilise une mesure de l'effet que l'on appelle ici la fraction prévenue ou évitable. La *fraction prévenue (ou évitabIe)* mesure, en proportion ou en pourcentage, le nombre de cas que le facteur prévient. Comme pour la fraction étiologique, on distingue deux fractions prévenues : la fraction prévenue chez les sujets exposés et la fraction prévenue totale.

Fraction prévenue chez les sujets exposés

Reprenons notre exemple du vaccin où R_1 représente le risque de maladie M chez les vaccinés, et R_0 celui chez les sujets non-vaccinés. On reconnaît qu'une partie de la proportion $(1 - R_1)$ des sujets non-malades chez les vaccinés est due à l'action du vaccin et que la partie restante est attribuable à une protection naturelle que l'on estime par la proportion $(1 - R_0)$ des non-malades chez le groupe des non-vaccinés. La protection exercée est donc donnée par:

$$(1 - R_1) - (1 - R_0)$$

soit par,

$$R_0 - R_1.$$

Pour l'exemple des 50 cas de maladie M dans un groupe de 1000 vaccinés et 200 cas dans un groupe de 1000 sujets non-vaccinés, la protection *R0 — R1* attribuable au vaccin est donc de:

$$\frac{200}{1000} - \frac{50}{1000}$$

soit,

$$\frac{150}{1000}$$

La différence des risques $R_0 — R_1$ mesure, dans le contexte causal, la protection attribuable au facteur E, c'est-à-dire sa responsabilité quant au nombre de cas qu'il a permis

d'éviter parmi les personnes qui lui ont été exposées (dans notre exemple, 150 parmi 1000 vaccinés). Le risque R_1 de maladie chez les sujets exposés au facteur *E,* par comparaison au risque *R0* des sujets non-exposés, se trouve être réduit d'une quantité $R_0 - R_1$.

On peut donc dire que le rapport $(R_0 - R_1)/R_0$ mesure en proportion, parmi tous les cas potentiels (dans notre exemple, 200 cas pour 1000), le nombre de ceux qui ont pu être prévenus par l'action du facteur *E* (dans notre exemple, 150 cas pour 1000). Ce rapport, qui est une proportion, est appelé la *fraction prévenue chez les exposés.* Elle est dénotée FP_1.

On a ainsi la formule:

$$FP_1 = \frac{R_0 - R_1}{R_0}$$

Dans l'exemple des vaccinés, la fraction prévenue chez les sujets exposés (ici chez les vaccinés) est égale à 0,75.

$$FP_1 = \frac{200/1000 - 50/1000}{200/1000} = 0{,}75.$$

On peut donc dire que 75 % des cas potentiels de la maladie M chez les vaccinés ont pu être évités par l'inoculation du vaccin.

Il est possible de reformuler la fraction prévenue FP_1 en utilisant la relation $Ef = R_0/R_1$.

On obtient:

$$FP_1 = \frac{Ef - 1}{Ef}.$$

En se référant au même exemple sur le vaccin, où $Ef = 4$, on trouve :

$$FP_1 = \frac{4 - 1}{4} = 0{,}75.$$

Fraction prévenue totale

Continuons l'exemple du vaccin. Le risque R_1 de maladie *M* chez les vaccinés est de $^{50}/_{10000}$ et celui R_0 chez les sujets non-vaccinés de $^{200}/_{1000}$, ce qui confère une efficacité de 4 pour le vaccin. Supposons que le vaccin ait été injecté à 60 % des individus d'une population risquant d'être affectée par la maladie *M.* On observe alors, dans cette population, une fréquence totale (R_t) de 110 cas de maladie pour 1000 individus. On rappelle que R_t est la somme pondérée des risques R_1 et R_0:

$$R_t = p_1R_1 + p_0R_0$$

où p, représente la proportion de sujets exposés dans la population et $p_0 = (1 - p_1)$ la proportion de sujets non-exposés.

Dans l'exemple,

$$\frac{110}{1000} = 0{,}60 \times \frac{50}{1000} + 0{,}40 \times \frac{200}{1000}.$$

Comme $R_0 - R_1$ mesure la fréquence de cas protégés (ou protection) chez les vaccinés, on peut comprendre que $R_0 - R_t$ mesure la protection totale dans la population. Dans l'exemple, la protection totale est de

$$\frac{200}{1000} - \frac{110}{1000} = \frac{90}{1000}.$$

Pour mesurer, parmi tous les cas potentiels, la proportion de ceux qui ont pu être évités, nous établissons le rapport de la fréquence $(R_0 - R_t)$ de cas protégés sur la fréquence totale (R_0) de cas potentiels.

$$\frac{R_0 - R_t}{R_0}$$

Ce rapport est la *fraction prévenue totale ou de population,* notée FP_t. Elle mesure l'effet total qu'a le facteur E dans la population.

On a ainsi la formule :

$$FP_t = \frac{R_0 - R_t}{R_0}$$

Pour l'exemple du vaccin, la fraction prévenue totale est égale à 0,45.

$$FP_t = \frac{200/1000 - 110/1000}{200/1000} = 0{,}45 \text{ (ou 45 \%)}.$$

En d'autres termes, le vaccin a permis de prévenir, d'éviter, dans la population, 45 % de tous les cas potentiels de la maladie *M*.

Il existe deux autres formules pour la fraction prévenue totale FP_t, chacune équivalente à la formule (définition): $(R_0 - R_t)/R_0$. Ces deux formules sont:

$$FP_t = p_1 \frac{(Ef - 1)}{Ef} \qquad [4]$$

où p_1 est la proportion de sujets exposés dans la population, et

$$FP_t = \frac{p_{c1}(Ef - 1)}{p_{c1}(Ef - 1) + 1}. \qquad [5]$$

où p_{Cl} est la proportion de sujets exposés parmi les cas.

Nous donnons ici le développement de ces deux formules.

Développement de la première formule

Si, dans l'expression $(R_0 - R_t)/R_0$, on remplace R_t par $p_1R_1 + p_0R_0$, où $p_1 + p_0 = 1$, alors on

$$FP_t = p_1 \frac{(R_0 - R_1)}{R_0}.$$

$$= p_1 \frac{(Ef - 1)}{Ef}.$$

Développement de la deuxième formule

On peut facilement retrouver une autre formule équivalente en utilisant la relation [3] déjà décrite entre p_1 et p_{C1} et dans laquelle on remplace *RR* par 1 /*Ef*. La relation devient:

$$p_1 = \frac{p_{C1}Ef}{p_{C1}Ef + p_{C0}}$$

Maintenant, par substitution de p_i dans l'expression [4] et en se rappelant que $p_{Co} = 1 - p_{C1}$, on obtient:

$$FP_t = \frac{p_{c1}(Ef - 1)}{p_{c1}(Ef - 1) + 1}.$$

Chacune des expressions [4] et [5] de la fraction prévenue totale peut être utilisée suivant que l'on connaît ou bien la proportion p1 de sujets exposés dans la population, ou bien la proportion p_{c1} de sujets exposés parmi les malades. Nous allons illustrer l'utilisation des formules [4] et [5].

On veut évaluer la performance d'un programme de vaccination qui a été implanté depuis quelque temps dans une population pour contrôler une maladie déterminée. On reconnaît à priori pour le vaccin une efficacité de

4 (Ef = 4) et l'on sait par ailleurs que la proportion de vaccinés dans la population est de 0,60 (p_1 = 0,60). La fraction prévenue totale est alors:

$$FP_t = \frac{0{,}60\,(4-1)}{4} = 0{,}45.$$

Parmi 110 sujets choisis au hasard ayant eu la maladie, on observe 80 sujets non-vaccinés (p_{C1} = $^{30}/_{110}$). Alors, la fraction prévenue totale du programme est de

$$FP_t = \frac{^{30}/_{110} \times (4-1)}{^{30}/_{110} \times (4-1) + 1} = 0{,}45.$$

Nous obtenons en définitive trois formules équivalentes pour la fraction prévenue totale FP_t, qui sont, rappelons le :

$$FP_t = \frac{R_0 - R_t}{R_0}$$

$$= p_1 \frac{(Ef-1)}{Ef} \quad \text{ou} \quad p_1 \frac{(R_0 - R_1)}{R_0}$$

$$= \frac{p_{C1}\,(Ef-1)}{p_{C1}\,(Ef-1)+1}.$$

La deuxième relation montre que la fraction prévenue totale est fonction à la fois de l'efficacité et de la proportion de sujets exposés dans la population. Pour être plus concret, l'effet global d'une intervention préventive, comme un programme de vaccination, dépend autant de l'efficacité de l'intervention que de la proportion des individus atteints par celle-ci. Une efficacité plus grande veut dire un effet plus grand. Une proportion p_1 plus grande de sujets atteints par l'intervention veut dire aussi un plus grand effet sur la population. Dans le tableau 7-2, nous donnons la valeur de l'effet global *(FP_t)* pour différentes valeurs d'efficacité *(Ef)* et différentes proportions de sujets exposés (P_1).

Tableau 7-2

Ef	p_1	FP_t	*Ef*	p_1	FP_t
2	0,50	0,250	10	0,50	0,450
2	0,70	0,350	10	0,70	0,630
2	0,90	0,450	10	0,90	0,810
4	0,50	0,375	12	0,50	0,458
4	0,70	0,525	12	0,70	0,642
4	0,90	0,675	12	0,90	0,825
6	0,50	0,417	14	0,50	0,464
6	0,70	0,583	14	0,70	0,650
6	0,90	0,750	14	0,90	0,836
8	0,50	0,437	16	0,50	0,469
8	0,70	0,612	16	0,70	0,656
8	0,90	0,787	16	0,90	0,844

RÉSUMÉ

Les deux principales mesures d'impact sont la fraction étiologique et la fraction prévenue ou évitable. On parle de fraction étiologique si un facteur a un effet positif, en ce sens que la présence du facteur est associée à une augmentation du risque relatif. Par contre, on parle de fraction prévenue si un facteur a un effet négatif, en ce sens que la présence du facteur est associée à une diminution du risque relatif. Ces fractions, étiologiques et prévenues, peuvent être calculées chez les sujets exposés et dans la population totale. La fraction étiologique mesure, en proportion ou en pourcentage, la responsabilité du facteur sur le nombre de cas d'une maladie (ou du décès). La fraction prévenue mesure, en proportion ou en pourcentage, le nombre de cas qu'un facteur prévient, évite.

Symboles

FE_1, FE_t: fraction étiologique chez les sujets exposés, étiologique totale

R_t, R_1, R_0: risque de maladie (ou de décès) dans la population totale, chez les sujets exposés, chez les sujets non-exposés

P_1, p_0: proportion de sujets exposés dans la population, de sujets non-exposés dans la population

P_{c1} P_{c0}: proportion de sujets exposés parmi les cas, de sujets non-exposés parmi les cas

Ef: mesure d'efficacité

FP_1, FP_t: fraction prévenue (ou évitable) chez les sujets exposés, prévenue totale

Formules

$$FE_1 = \frac{R_1 - R_0}{R_1} = \frac{RR - 1}{RR} \left(= \frac{RC - 1}{RC}\right);$$

$$FP_1 = \frac{R_0 - R_1}{R_0} = \frac{Ef - 1}{Ef}$$

$$R_t = p_1 R_1 + p_0 R_0$$

$$FE_t = \frac{R_t - R_0}{R_t}; \qquad FP_t = \frac{R_0 - R_t}{R_0}$$

$$FE_t = \frac{p_1 (RR - 1)}{p_1 (RR - 1) + 1}; \; FP_t = \frac{p_{c1} (Ef - 1)}{p_{c1} (Ef - 1) + 1}$$

$$FE_t = p_{c1} \times \frac{RR - 1}{RR}; \qquad FP_t = p_1 \times \frac{Ef - 1}{Ef}$$

$$Ef = RR^{-1} = \frac{R_0}{R_1}$$

LECTURES SUGGÉRÉES

1. KLEINBAUM, D.G., KUPPER, L.L. et MORGENSTERN, *H.* ***EpidemioIogic Research,*** Belmont (USA), Lifetime Learning Publications, 1982, chapitre 9, pp. 159-180.

2. ROTHMAN, K.J. ***Modern EpidemioIogy,*** Boston, Little, Brown, 1986, chapitre 4, pp. 35-40.

ANNEXE DU CHAPITRE 7

Estimation de R_1 et R_O dans les études cas-témoins en fonction de la fraction étiologique totale FE_t

Nous rappelons que, dans les études cas-témoins, on ne connaît pas directement le risque R_1 chez les sujets exposés et le risque R_0 chez les sujets non-exposés. Ici, les risques R_1 et R_0 peuvent être pris comme taux d'incidence ou, dans l'hypothèse de la rareté de la maladie, comme incidence cumulative.

Si l'étude cas-témoins est faite sur la totalité des cas incidents, si la population à risque N est connue et si le groupe des témoins en constitue un échantillon représentatif, alors on peut estimer:

1) le risque total R_t:

$$R_t = \frac{M_1}{N\triangle t}, \text{ pour le taux d'incidence}$$

ou

$$R_t = \frac{M_1}{N}, \text{ pour l'incidence cumulative.}$$

2) le risque relatif RR:

$$RR = RC = \frac{ad}{bc}$$

3) la fraction étiologique totale FE_t:

$$FE_t = p_{C1} \times \frac{RC - 1}{RC}.$$

Puisque

$$FE_t = \frac{R_t - R_0}{R_t}$$

on déduit que:

$$R_0 = R_t(1 - FE_t).$$

De $R_1/R_0 = RR = RC$, on obtient:

$$R_1 = RR \times R_t(1 - FE_t).$$

Ces deux formules peuvent être utilisées dans les études cas-témoins pour estimer les risques R, et R_0 qui ne sont pas connus directement.

Une étude cas-témoins a été faite sur la totalité des 200 cas incidents survenus dans une population (stable) de 100 000 individus. Ces 200 cas ont été diagnostiqués au cours d'une période de deux ans. Dans l'étude, ils ont été comparés, pour la présence du facteur E, à un échantillon de 200 témoins tiré de la population. Les résultats sont présentés dans le tableau A7-1.

Tableau A7-1

	E		
	+	−	Total
Cas	160	40	200
Témoins	100	100	200

On a alors,

$$R_t = \frac{M_1}{N\triangle t}$$

$$= \frac{200 \text{ cas}}{100\,000 \text{ personnes} \times 2 \text{ ans}}$$

$$= 0{,}001/\text{année}.$$

$$RC = \frac{160 \times 100}{40 \times 100} = 4$$

$$FE_t = \frac{160}{200} \times \frac{4 - 1}{4} = 0{,}60.$$

Ainsi, on peut estimer R_1 et R_0. On trouve :

$$R_0 = 0{,}001/\text{année} \times (1 - 0{,}60)$$

$$= 0{,}0004/\text{année ou 4 cas par 10 000 personnes-années}$$

$$R_1 = 4 \times 0{,}001/\text{année} \times (1 - 0{,}60)$$

$$= 0{,}0016/\text{année ou 16 cas par 10 000 personnes-années.}$$

CHAPITRE 8

Mesures d'interaction

Ce chapitre distingue les notions d'interaction dans le modèle additif et dans le modèle multiplicatif. Le premier type d'interaction met en cause l'effet conjoint de deux facteurs sur la différence des risques (ou risque attribuable); de ce fait, il intéresse davantage la santé publique. Le deuxième type met en cause l'effet conjoint de deux facteurs sur le risque relatif et intéresse donc la recherche étiologique. On propose pour chacun des types une mesure d'interaction. Enfin, on suggère deux mesures de l'impact d'une interaction dans le modèle additif.

Souvent, on est appelé à s'interroger sur l'action combinée de deux ou plusieurs facteurs dans l'apparition d'une maladie. On s'intéresse, par exemple, à l'action combinée des expositions au tabac et aux poussières d'amiante dans l'apparition du cancer du poumon. On essaie de découvrir si la présence des deux facteurs de risque dans un groupe induit un plus grand nombre de cas que celui correspondant à la somme des cas induits séparément par chacun des facteurs. À un autre niveau d'investigation, on peut se demander si la présence simultanée des facteurs chez un individu ne produit pas un risque plus grand d'être affecté par la maladie que le produit des risques induits séparément par ces facteurs. Si leur combinaison conduit à un effet plus grand que celui correspondant à la somme de leurs effets pris séparément, on parle alors *d'interaction positive (ou synergie);* pour un effet plus petit, on parle *d'interaction négative (ou antagonisme).*

Nous éviterons les termes synergie et antagonisme, plus en relation avec la description de l'action des mécanismes biologiques dans l'apparition de la maladie. Nous sommes plutôt intéressés ici par la description quantitative, statistique, de l'action simultanée de facteurs de risque sur les données épidémiologiques observées.

Dans ce qui suit, nous allons distinguer les deux principaux types d'interaction considérés en épidémiologie. Le premier marque l'influence qu'exerce la combinaison des facteurs sur la différence des risques. Il est dit de *modèle additif.* Le second décrit l'influence qu'exerce la combinaison des facteurs sur le risque relatif et est dit de *modèle multiplicatif.* La présentation de ces notions se fera dans un cadre simplifié. Seule l'interaction entre deux facteurs dichotomiques sera considérée. Enfin, l'appellation risque dans les expressions « différence des risques » et « risque relatif » réfère aussi bien à un taux d'incidence, à un taux de décès, qu'à une incidence cumulative ou une prévalence relative, suivant le contexte.

INTERACTION DANS LE MODÈLE ADDITIF

Considérons une population dans laquelle sont présents les deux facteurs de risque A et B pour la maladie M. Le tableau 8-1 décrit la fréquence de la maladie pour les quatre catégories d'exposition désignées respectivement par A_0B_0 lorsque A et B sont absents, par A_1B_0 lorsque seul A est présent, par A_0B_1 lorsque seul B est présent et par A_1B_1 lorsque les deux facteurs sont présents.

Tableau 8-1

		Classe d'exposition				
		A_0B_0	A_1B_0	A_0B_1	A_1B_1	Total
Maladie	+	2	16	8	10	36
	Personnes	10 000	20 000	20 000	10 000	60 000

Dans ce tableau, on lit que :

1) le risque R_{00} chez les sujets exposés à aucun des deux facteurs est de $^{2}/_{10\,000}$;

2) le risque R_{10} chez les sujets exposés seulement à A est de $^{16}/_{20\,000}$ ou $^{8}/_{10\,000}$;

3) le risque R_{01} chez les sujets exposés seulement *à B* est de $^{8}/_{20\,000}$ ou $^{4}/_{10\,000}$;

4) le risque R_{11} chez les sujets exposés *à* la fois *à* A et *à B* est de $^{10}/_{10\,000}$.

Si on prend le groupe des sujets non exposés, ni *à* A ni *à B,* comme groupe de référence, on peut calculer les trois différences de risques suivantes.

$$DR_{10} = R_{10} - R_{00}\ (= {}^{6}/_{10\,000});$$
$$DR_{01} = R_{01} - R_{00}\ (= {}^{2}/_{10\,000});$$
$$DR_{11} = R_{11} - R_{00}\ (= {}^{8}/_{10\,000});$$

Chaque *DR* mesure l'effet additif de l'exposition à l'un ou à l'autre ou aux deux facteurs sur la fréquence de la maladie. Par exemple, DR_{11} marque l'effet additif de l'exposition aux deux facteurs A et B : cette double exposition augmente la fréquence de base R_{00} de 8 cas par 10 000 sujets ou, si l'on veut, suivant un jugement causal, la combinaison des deux facteurs est responsable de 8 cas par 10 000 sujets qui leur sont simultanément exposés. Rappelons que dans le cas d'un lien causal, le symbole *DR* pourrait être remplacé par RA (le risque attribuable).

On observe ici l'égalité :

$$DR_{11} = DR_{10} + DR_{01}$$
$$({}^{8}/_{10\,000} = {}^{6}/_{10\,000} + {}^{2}/_{10\,000})$$

Dans cet exemple, l'effet de la combinaison des deux facteurs équivaut à la somme de leurs effets pris séparément. Pour 10 000 personnes exposées, le nombre de cas attribuables à la double exposition est égal à la somme des cas attribuables d'une part à A et d'autre part à *B.* Il n'en va pas toujours ainsi.

Considérons une autre population dans laquelle sont aussi présents les deux facteurs de risque A et B pour la maladie M, mais où cette fois l'égalité précédente n'est plus vérifiée. Le tableau 8-2 décrit la fréquence de la maladie pour les quatre catégories d'exposition.

Dans ce tableau, on lit que :

$$R_{00} = {}^{2}/_{10\,000};\ R_{10} = {}^{8}/_{10\,000};$$
$$R_{01} = {}^{4}/_{10\,000};\ R_{11} = {}^{16}/_{10\,000}.$$

Les différences des risques sont maintenant :

$$DR_{10} = R_{10} - R_{00}\quad (= {}^{6}/_{10\,000});$$
$$DR_{01} = R_{01} - R_{00}\quad (= {}^{2}/_{10\,000});$$
$$DR_{11} = R_{11} - R_{00}\quad (= {}^{14}/_{10\,000}).$$

Tableau 8-2

		Classe d'exposition				
		A_0B_0	A_1B_0	A_0B_1	A_1B_1	Total
Maladie	+	4	8	4	32	48
	Personnes	20 000	10 000	10 000	20 000	60 000

On remarque qu'ici

$$DR_{11} > DR_{10} + DR_{01}$$

L'effet de la combinaison des facteurs A et *B* est plus grand que la somme de leurs effets pris séparément. Par la présence des deux facteurs, la fréquence de base R_{00} est augmentée de 6 cas par 10 000 à cause de A, de 2 cas par 10 000 à cause de *B,* et d'un autre 6 cas par 10 000 à cause de leur présence simultanée. C'est un peu comme si la combinaison des facteurs A et *B* engendrait un nouveau facteur dont l'effet correspond à 6 cas supplémentaires par 10 000 sujets qui lui sont exposés. C'est l'effet *d'interaction dans le modèle additif.* Les facteurs A et *B* agissent ici en interaction positive pour produire un nombre de cas supplémentaires à la somme de ceux produits par les facteurs pris séparément.

Dans l'exemple du tableau 8-1, la combinaison des facteurs n'a pas produit de cas supplémentaires à ceux produits par leur action séparée. Il n'y avait pas d'interaction. Rappelons qu'on avait alors, $DR_{11} = DR_{10} + DR_{01}$. On pourrait envisager des situations où l'interaction est négative; dans ce cas, l'effet de la combinaison des facteurs serait moindre que la somme de leurs effets séparés.

Résumons.

Si $DR_{11} > DR_{10} + DR_{01}$, il y a interaction positive;

Si $DR_{11} = DR_{10} + DR_{01}$, il y a absence d'interaction;

Si $DR_{11} < DR_{10} + DR_{01}$, il y a interaction négative.

Comme mesure de l'interaction dans le modèle additif, il est naturel de proposer la valeur

$$i_{AB} = DR_{11} - DR_{10} - DR_{01}$$

$$= R_{11} - R_{10} - R_{01} + R_{00}.$$

Cette mesure a cependant le défaut de ne pas toujours être applicable, notamment dans les études cas-témoins. Pour la rendre plus généralement utilisable, on peut la transformer de la façon suivante.

$$\frac{i_{AB}}{R_{00}} = i'_{AB} = RR_{11} - RR_{10} - RR_{01} + 1$$

où $RR_{11} = R_{11}/R_{00}$, $RR_{10} = R_{10}/R_{00}$ et $RR_{01} = R_{01}/R_{00}$.

Ainsi,

$i'_{AB} > 0$ indique une interaction positive;

$i'_{AB} = 0$ indique une absence d'interaction;

$i'_{AB} < 0$ indique une interaction négative.

Au tableau 8-2, on trouve $i_{AB} = {}^{6}/_{10\,000}$ et $i'_{AB} = 3$. Il y a bien interaction positive. La valeur 3 indique que la combinaison des facteurs A et B, outre leurs effets respectifs, produit un effet supplémentaire égal à trois fois la fréquence de base R_{00}. La valeur i'_{AB} mesure l'effet additif que la combinaison des facteurs, outre leurs effets respectifs, produit sur la fréquence de base R_{00}.

On peut schématiser le problème de l'interaction dans le modèle additif comme au tableau 8-3 et ainsi mieux comprendre l'appellation modèle additif pour qualifier ce type d'interaction. Les effets sont additifs.

Tableau 8-3

		Facteur A	
		A_0	A_1
Facteur B	B_0	R_{00}	$R_{00} + e_A$
	B_1	$R_{00} + e_B$	$R_{00} + e_A + e_B + i_{AB}$

où e_A *(=* $DR_{,10}$*)* désigne l'effet de *A,* e_B *(=* DR_{01}*)* désigne l'effet de *B,* i_{AB} désigne l'interaction entre A et *B.*

INTERACTION DANS LE MODÈLE MULTIPLICATIF

Dans l'exemple du tableau 8-2, les risques relatifs correspondant aux différents niveaux d'exposition sont:

$$RR_{10} = \frac{R_{10}}{R_{00}} = 4; \quad RR_{01} = \frac{R_{01}}{R_{00}} = 2;$$

$$RR_{11} = \frac{R_{11}}{R_{00}} = 8.$$

Chaque mesure *RR* permet de décrire l'effet multiplicateur de l'un ou l'autre ou les deux facteurs sur le risque de la maladie. Ainsi, par exemple, RR_{11}, (= 8) indique que l'exposition aux deux facteurs A et B multiplie par *8* le risque de la maladie.

On observe ici l'égalité:

$$RR_{11} = RR_{10} \cdot RR_{01} \quad (8 = 4 \times 2)$$

Dans cet exemple, l'effet de la combinaison des deux facteurs est équivalent au produit de leurs effets respectifs. Tout se passe comme si les deux facteurs, A et *B,* agissent de façon indépendante dans l'étiologie de la maladie. Il n'en va pas toujours ainsi.

Considérons une autre situation décrite au tableau 8-4 dans laquelle l'égalité précédente n'est pas respectée. Dans ce tableau, la présence de A seul multiplie par 4 le risque R_{00}, alors que la présence de *B* seul le multiplie par 2. Pourtant, lorsque les deux facteurs sont présents, le risque relatif observé est de 12 *(*RR_{11} = 12), soit 1,5 fois plus élevé que le produit des effets séparés (4 et 2). Tout se passe comme si, au plan étiologique, la combinaison des deux facteurs, A et *B,* engendrait un effet supplémentaire à leurs effets séparés. C'est l'effet *d'interaction dans le modèle multiplicatif.* Cette fois :

$$RR_{11} > RR_{10} \cdot RR_{01}$$

La présence de A multiplie le risque par 4, celle de B par 2 et l'effet d'interaction (qu'on désigne ici par I_{AB}*)* par 1,5.

Tableau 8-4

		Classe d'exposition				
		A_0B_0	A_1B_0	A_0B_1	A_1B_1	Total
Maladie	+	4	8	4	48	64
	Personnes	20 000	10 000	10 000	20 000	60 000

Il est naturel de proposer l'expression suivante comme mesure d'interaction dans le modèle multiplicatif:

$$I_{AB} = \frac{RR_{11}}{RR_{10} \cdot RR_{01}}.$$

Ainsi, dans le modèle multiplicatif,

$I_{AB} > 1$ indique une interaction positive;

$I_{AB} = 1$ indique l'absence d'interaction;

$I_{AB} < 1$ indique une interaction négative.

Dans l'exemple précédent du tableau 8-4, puisque

$$I_{AB} = \frac{12}{4 \times 2} = 1{,}5$$

il y a interaction positive entre les deux facteurs A et B.

Au tableau 8-1, elle est négative :

$$I_{AB} = \frac{5}{4 \times 2} = 0{,}625.$$

Au tableau 8-2, il n'y a pas d'interaction:

$$I_{AB} = \frac{8}{4 \times 2} = 1$$

Le problème de l'interaction dans le modèle multiplicatif est décrit au tableau 8-5. On comprend mieux l'appellation modèle multiplicatif pour décrire ce type d'interaction.

Tableau 8-5

		Facteur A	
		A_0	A_1
Facteur B	B_0	R_{00}	$e_A R_{00}$
	B_1	$e_B R_{00}$	$e_A e_B I_{AB} R_{00}$

où e_A (= RR_{10}) désigne ici l'effet multiplicatif de A, e_6 (= $RR0_1$) désigne celui de B, ^1AB désigne l'interaction entre A et B.

INTERACTION ET MODIFICATION

Interaction dans le modèle additif et modification de la différence des risques

L'interaction dans le modèle additif du tableau 8-2 peut facilement être mise en évidence si on dispose les données dans deux tableaux 2 x 2. Les données, stratifiées pour le facteur B, peuvent être décrites à l'aide des tableaux 8-6A lorsque B est absent et 8-6B lorsque B est présent. Nous désignons par $DR\ (A_1/B_0)$ et par $DR\ (A_l/B_1)$ la différence des risques par rapport à A respectivement pour B absent et pour B présent. Ces différences de risques changent ici d'une strate à l'autre. Il passe de $^6/_{10\ 000}$ quand B est absent à $^{12}/_{10\ 000}$ quand il est présent. Dans ce cas, la présence de B a modifié l'effet de A mesuré par la différence des risques. On dit que B est un *facteur modifiant* de la différence des risques qui mesure l'effet de A. Il y a *modification de la différence des risques.*

Tableau 8-6A

B absent

		Facteur *A*	
		A_0	A_1
M	+	4	8
Total		20 000	10 000

Tableau 8-6B

B présent

		Facteur *A*	
		A_0	A_1
M	+	4	32
Total		10 000	20 000

On peut facilement montrer que les notions d'interaction dans le modèle additif et de modification de la différence des risques sont équivalentes. En effet, si le facteur B est modifiant pour la différence des risques, on a:

$$DR(A_1/B_0) \neq DR(A_1/B_1)$$

c'est-à-dire,

$$R_{10} - R_{00} \neq R_{11} - R_{01}$$

$$R_{10} - R_{00} \neq (R_{11} - R_{00}) - (R_{01} - R_{00})$$

D'où

$$DR_{10} \neq DR_{11} - DR_{01}$$

ce qui correspond à

$$i_{AB} = DR_{11} - DR_{10} - DR_{01} \neq 0$$

Il y a donc présence d'interaction. En prenant à rebours la démonstration, on peut facilement démontrer l'inverse.

Interaction dans le modèle multiplicatif et modification du risque relatif

L'interaction dans le modèle multiplicatif du tableau 8-4 peut facilement être mise en évidence par les tableaux 8-7A et 8-7B.

Tableau 8-7A

B absent

		Facteur *A*	
		A_0	A_1
M	+	4	8
Total		20 000	10 000

Tableau 8-7B

B présent

		Facteur *A*	
		A_0	A_1
M	+	4	48
Total		10 000	20 000

Le risque relatif pour A change d'une strate à l'autre. Il passe de 4 lorsque *B* est absent à 6 lorsqu'il est présent. Le facteur *B* est *modifiant pour le risque relatif* de *A*. Il y a *modification du risque relatif.* On peut facilement montrer que les deux notions d'interaction dans le modèle multiplicatif et de modification du risque relatif sont équivalentes. En effet, si le facteur *B* est modifiant pour le risque relatif, on a

$RR(A_1/B_0) \neq RR(A_1/B_1)$, c'est-à-dire

$$\frac{R_{10}}{R_{00}} \neq \frac{R_{11}}{R_{01}}$$

Ainsi,

$$RR_{10} \neq \frac{R_{11}/R_{00}}{R_{01}/R_{00}}$$

ou encore

$$RR_{10} \neq \frac{RR_{11}}{RR_{01}} \quad \text{ou} \quad RR_{11} \neq RR_{10} \cdot RR_{01}.$$

On a donc,

$$I_{AB} = \frac{RR_{11}}{RR_{10} \cdot RR_{01}} \neq 1$$

c'est-à-dire présence d'interaction dans le modèle multiplicatif. Ici aussi, en prenant à rebours la démonstration, on peut facilement démontrer l'inverse. La notion de modification du risque relatif sera reprise dans le chapitre 16.

RELATION ENTRE L'INTERACTION DANS LE MODÈLE ADDITIF ET L'INTERACTION DANS LE MODÈLE MULTIPLICATIF

Deux facteurs peuvent être en interaction positive dans le modèle additif ($i'_{AB} > 0$) sans l'être dans le modèle multiplicatif ($I_{AB} = 1$) comme nous l'avons vu au tableau 8-2. Ce fait est la conséquence d'un résultat plus large. Quand $RR_{,0} > 1$ et $RR_{01} > 1$, c'est-à-dire quand A et B sont des facteurs de risque, l'absence d'interaction ou la présence d'interaction positive dans le modèle multiplicatif ($I_{AB} \geq 1$) implique qu'il y a interaction positive dans le modèle additif ($i'_{AB} > 0$). En voici la démonstration.

On a, $I_{AB} \geq 1$

c'est-à-dire $RR_{11} \geq RR_{10} \cdot RR_{01}$. [1]

Par ailleurs, puisque $RR_{10} > 1$, on a

$$RR_{10}(RR_{01} - 1) > RR_{01} - 1,$$

$$RR_{10} \cdot RR_{01} - RR_{10} > RR_{01} - 1$$

$$RR_{10} \cdot RR_{01} > RR_{10} + RR_{01} - 1 \quad [2]$$

En combinant les inégalités [1] et [2], on obtient:

$$RR_{11} > RR_{10} + RR_{01} - 1$$

d'où, $RR_{11} - RR_{10} - RR_{01} + 1 > 0$

c'est-à-dire $i'_{AB} > 0$.

Deux facteurs peuvent être en interaction négative dans le modèle multiplicatif ($I_{AB} < 1$) sans l'être dans le modèle additif ($i'_{AB} = 0$) comme nous l'avons constaté au tableau 8-1. Ce fait s'inscrit dans un résultat plus large. Quand $RR_{10} > 1$ et $RR_{01} > 1$, l'absence d'interaction ou la présence d'interaction négative dans le modèle additif ($i'_{AB} \leq 0$) implique la présence d'interaction négative dans le modèle multiplicatif ($I_{AB} < 1$). La démonstration est immédiate puisqu'il est impossible que $I_{AB} \geq 1$. Autrement, en vertu de la démonstration précédente, cela impliquerait que $i'_{AB} > 0$ et contredirait ainsi la supposition de départ, $i'_{AB} \leq 0$.

MESURE DE L'IMPACT D'UNE INTERACTION

Rappelons que les mesures d'impact visent à quantifier l'effet qu'a un facteur sur la fréquence de la maladie ou du décès. Ici, nous allons définir deux mesures d'impact de l'interaction de deux facteurs sur la fréquence de la maladie (ou du décès). La première correspond à la proportion des cas attribuables à l'interaction, parmi tous les cas exposés à la fois à A et à B; nous l'appellerons *fraction étiologique due à l'interaction* et la désignerons par *FEi.* La deuxième mesure correspond à la proportion de cas attribuables à l'interaction, parmi tous les cas attribuables à la présence des deux facteurs; nous la nommerons *fraction attribuable à l'interaction* et la désignerons par *FAi.* Ces deux mesures se

distinguent par leurs dénominateurs: le dénominateur de la première réfère à tous les cas exposés à la fois aux deux facteurs; celui de la deuxième renvoie aux seuls cas attribuables à cette double exposition.

Les définitions de mesures d'impact d'une interaction ne sont données qu'en référence au modèle additif. Une raison formelle justifie cette approche. La définition d'une mesure qui essaie de quantifier la responsabilité d'une interaction dans la fréquence des cas relève directement de la notion de nombre de cas attribuables. Ce nombre est évalué par la différence des risques, qui est en relation directe avec le modèle additif.

Essayons de formaliser la première mesure. Parmi tous les cas exposés aux deux facteurs, combien en proportion sont dus à leur interaction? Rappelons que, dans le modèle additif, la mesure i_{AB} correspond à la fréquence des cas attribuables à l'interaction entre les deux facteurs. Si on divise cette valeur par la fréquence des cas exposés aux deux facteurs (R_{11}), on obtiendra effectivement la proportion recherchée. Dans l'exemple du tableau 8-2, calculons cette proportion.

On a

$i_{AB} = 6/10\,000$ et $R_{11} = 16/10\,000$. Ainsi,

$$FEi = \frac{6/10\,000}{16/10\,000} = \frac{6}{16} = 0{,}375 \text{ ou } 37{,}5\,\%.$$

De tous les cas exposés à A et à B, 37,5 % sont attribuables à l'interaction de ces deux facteurs.

Dans sa forme générale, cette fraction s'écrit comme

$$FEi = \frac{DR_{11} - DR_{10} - DR_{01}}{R_{11}}$$

$$= \frac{RR_{11} - RR_{10} - RR_{01} + 1}{RR_{11}}$$

Examinons maintenant la deuxième mesure. Parmi tous les cas attribuables à la présence des deux facteurs, combien en proportion sont dus à leur interaction? Si on divise i_{AB} par la fréquence de cas attribuables à la présence des deux facteurs (DR_{11}), on obtiendra effectivement la proportion recherchée. Dans l'exemple du tableau 8-2, calculons cette proportion.

On a

$i_{AB} = 6/10\,000$ et $DR_{11} = 14/10\,000$. Ainsi,

$$FAi = \frac{6/10\,000}{14/10\,000} = \frac{6}{14} = 0{,}429 \text{ ou } 42{,}9\,\%.$$

De tous les cas attribuables à l'exposition simultanée à A et à B, 42,9 % le sont à l'interaction de ces deux facteurs.

Dans sa forme générale, la fraction FAi s'écrit comme

$$FAi = \frac{DR_{11} - DR_{10} - DR_{01}}{DR_{11}}$$

$$= \frac{RR_{11} - RR_{10} - RR_{01} + 1}{RR_{11} - 1}.$$

La mesure *FEi* est une fraction étiologique au même titre que la fraction étiologique FE_I définie dans le chapitre 7. Elle mesure la proportion des cas attribuables à l'interaction, parmi les cas exposés à la fois à A et à B. La deuxième mesure, qui lui est toujours supérieure (sauf en l'absence d'interaction) mesure la proportion de cas attribuables à

l'interaction, parmi les cas attribuables à l'ex-position simultanée à A et à B.

En l'absence d'interaction dans le modèle additif ($i_{AB} = 0$), on a bien entendu *FEi=FAi=O*

COMMENTAIRES

Au plan statistique, l'interaction se définit dans le cadre d'un modèle additif, multiplicatif ou autre. Ce modèle essaie d'exprimer le risque dû à la combinaison des facteurs en fonction de leurs effets séparés, de leur interaction et du risque de base. Cette description est généralement faite à l'aide d'une fonction f ayant la forme générale

$$R_{11} = f(e_0, e_A, e_B, i_{AB}).$$

Ainsi, pour le modèle additif, on a:

$$R_{11} = R_{00} + e_A + e_B + i_{AB};$$

pour le modèle multiplicatif, on a:

$$R_{11} = e_A \cdot e_B \cdot I_{AB} \cdot R_{00}.$$

Une interaction statistique, qu'elle soit décrite dans le modèle additif ou dans le modèle multiplicatif, n'est pas toujours facile à interpréter au plan biologique. En retour, il est difficile d'arriver à une définition précise de l'interaction biologique et même de s'entendre sur un énoncé comme « une interaction biologique décrit l'action interdépendante de deux facteurs pour produire la maladie ».

L'étude de l'interaction répond à deux préoccupations. D'une part, elle participe à l'élaboration de modèles qui essaient de décrire le mieux possible les processus étiologiques d'une maladie. D'autre part, elle vise à mesurer tant l'impact des facteurs sur la santé d'une population que l'évaluation des interventions en santé communautaire. La présentation des mesures d'impact dans la section précédente relève surtout de ce dernier intérêt.

RÉSUMÉ

Dans le modèle additif, les facteurs peuvent agir en interaction pour produire un effet plus grand (interaction positive) ou plus faible (interaction négative) que la somme des effets des facteurs pris séparément. Dans ce modèle, les effets des facteurs sont évalués par des risques attribuables ou différences des risques. Dans le modèle multiplicatif, les facteurs peuvent agir en interaction pour produire un effet plus grand (interaction positive) ou plus faible (interaction négative) que le produit des effets des facteurs pris séparément. Dans ce modèle, les effets des facteurs sont évalués par des risques relatifs. S'il y a absence d'interaction ou présence d'interaction positive dans le modèle multiplicatif, alors il y a présence d'interaction positive dans le modèle additif. Par ailleurs, s'il y a absence d'interaction ou présence d'interaction négative dans le modèle additif, alors il y a présence d'interaction négative dans le modèle multiplicatif. Les notions d'interaction dans le modèle additif et de modification de la différence des risques sont équivalentes. Il en va ainsi des deux notions d'interaction dans le modèle multiplicatif et de modification du risque relatif. La fraction étiologique due à l'interaction *(FEi)* mesure en proportion le nombre de cas attribuables à l'interaction parmi tous les cas exposés à la fois aux deux facteurs. La fraction attribuable *(FAi)* mesure en proportion le nombre de cas attribuables à l'interaction parmi les seuls cas attribuables aux deux facteurs.

Symboles

i_{AB} : mesure d'interaction dans le modèle additif

i'_{AB}: mesure d'interaction dans le modèle additif exprimée en fonction du *RR*

I_{AB}: mesure d'interaction dans le modèle multiplicatif

FEi: fraction étiologique due à l'interaction, désignant la proportion des cas attribuables à l'interaction parmi tous les cas exposés à la fois à *A* et à *B*

FAi : fraction attribuable à l'interaction, désignant la proportion des cas attribuables à l'interaction parmi tous les cas attribuables à l'exposition simultanée aux facteurs *A* et *B*.

Formules

$$i_{AB} = DR_{11} - DR_{10} - DR_{01} = R_{11} - R_{10} - R_{01} + R_{00}.$$

$$i'_{AB} = RR_{11} - RR_{10} - RR_{01} + 1$$

$$I_{AB} = \frac{RR_{11}}{RR_{10} \cdot RR_{01}}$$

$$FEi = \frac{RR_{11} - RR_{10} - RR_{01} + 1}{RR_{11}}$$

$$FAi = \frac{RR_{11} - RR_{10} - RR_{01} + 1}{RR_{11} - 1}$$

LECTURES SUGGÉRÉES

1. KLEINBAUM, D., KUPPER L.L. et MORGENSTERN H. ***Epidemiologic Research,*** Lifetime Learning Publications, Belmont (USA), 1982, chapitre 19, pp. 403-418.
2. ROTHMAN, K.J. ***Modern Epidemiology,*** Boston, Little, Brown, 1986, chapitre 15, pp.311-326.
3. SCHLESSELMAN, J.J. ***Case-Control Studies,*** Oxford University Press, Oxford, 1982, chapitre 2, pp. 63-68.

CHAPITRE 9

Mesures d'accord

Dans ce chapitre, on distingue d'abord les termes « association » et « accord D. Ensuite, on présente la mesure kappa utilisée pour évaluer l'accord entre les jugements de nature qualitative de deux observateurs indépendants, puis on définit le coefficient de corrélation intra-classe comme mesure d'accord entre deux observateurs indépendants lorsque les jugements sont de nature quantitative.

Deux ou plusieurs observateurs indépendants peuvent porter des jugements différents sur un même sujet en regard d'une caractéristique; ainsi, des radiologistes peuvent interpréter de façon différente une radiographie. Cette divergence d'appréciation est un phénomène bien connu, spécialement dans les disciplines médicales et épidémiologiques. Deux ou plusieurs sources d'information peuvent concorder ou diverger sur certains aspects des caractéristiques d'un individu. Pour les travailleurs d'une industrie, le registre syndical comporte-il la même information que le registre patronal? Ces deux registres concordent-ils?

Dans ce chapitre, nous allons aborder la question de la mesure de l'accord (ou du désaccord) entre deux procédures qui doivent conduire au classement d'un sujet suivant un critère déterminé. Les procédures peuvent être comprises comme deux observateurs, deux sources d'information (deux registres), deux tests diagnostiques, deux questionnaires, etc. A ce titre et sans perte de généralité, nous utiliserons le terme « observateur » pour désigner toute procédure de classification.

Nous allons distinguer le cas où le critère est de nature qualitative de celui où il est de nature quantitative. Le jugement est *qualitatif* lorsque les sujets sont classés dans des catégories mutuellement exclusives et collectivement exhaustives formant ainsi une échelle de classification nominale. C'est le cas notamment lorsqu'un observateur psychiatre doit poser un diagnostic sur l'état de santé mentale d'un individu en indiquant s'il est normal, neurotique ou psychotique. II est qualitatif aussi lorsqu'un orthopédiste doit se prononcer sur la présence ou l'absence de scoliose chez un individu. Le jugement est *quantitatif* quand l'observateur attribue à chaque sujet une valeur numérique (un score). L'observateur orthopédiste qui mesure le degré de courbure de la colonne vertébrale d'un individu porte un jugement quantitatif. Dans ce cas, le classement des sujets se fait par grandeur de courbure exprimée en degrés.

Nous examinerons d'abord la situation où l'appréciation des observateurs est qualitative, pour y définir la mesure d'accord kappa; l'appréciation quantitative retiendra ensuite notre attention. Nous présenterons alors la mesure d'accord définie par le coefficient de corrélation intra-classe.

MESURE DE L'ACCORD ENTRE DEUX OBSERVATEURS DANS LE CAS D'UNE APPRÉCIATION QUALITATIVE

Avant d'introduire la mesure d'accord kappa, il importe de bien distinguer les notions d'association et d'accord. Pour ce faire, nous utiliserons des exemples.

Le tableau 9-1 présente les résultats fictifs d'un classement fait par deux observateurs indépendants (O_1 et 0_2) : chacun devait classer 60 sujets en trois catégories (diagnostiques) C_1, C_2 ou C_3.

Tableau 9-1

		0_1			
		C_1	**C_2**	**C_3**	**Total**
0_2	C_1	0	20	0	**20**
	C_2	0	0	20	**20**
	C_3	20	0	0	**20**
	Total	**20**	**20**	**20**	**60**

Les deux observateurs ont toujours été en parfait désaccord; aucune concordance n'a été observée dans leurs jugements, du genre C_1C_1, C_2C_2 ou C_3C_3. La diagonale du tableau 9-1 est vide. S'il n'y a eu aucun accord entre les deux observateurs, par contre il y a par- faite association entre leurs jugements. Si l'on connaît le jugement de l'observateur O_1, celui de l'observateur O_2 est entièrement prévisible. Si l'observateur O_1 attribue la catégorie C_2 à un sujet, l'observateur O_2 lui assigne à coup sûr la catégorie C_1. L'exemple du tableau 9-1 illustre bien qu'il y a tout lieu de distinguer entre l'association et l'accord. On y observe une complète association, mais dans un désaccord total.

Il est possible, par ailleurs, de trouver la plus parfaite association avec le plus parfait accord entre deux observateurs, comme en témoigne l'exemple au tableau 9-2.

Le jugement de l'observateur O_2 est prévisible quand on connaît celui de l'observateur O_1; de plus c'est le même jugement qui est posé. Enfin il est impossible d'imaginer une situation où il y aurait accord sans association. L'accord est un cas spécial d'association.

Mesure d'accord *po*

La plus simple ou la plus primitive des mesures d'accord est la *proportion globale observée* p_O de jugements convergents ou concordants (l'indice O réfère au terme observée). Elle reflète le degré de concordance entre les jugements de deux observateurs. Au tableau 9-3, sur 20 jugements posés, les deux observateurs ont été au total d'accord 15 fois: 7 fois pour la catégorie C_1 et 8 fois pour la catégorie C_2.

En proportion, on a donc selon le tableau :

$$p_O = \frac{15}{20} = 0{,}75$$

On remarque que le numérateur de p_O est déterminé par les éléments de la diagonale du tableau. De façon générale, si le classement de n sujets par deux observateurs se fait suivant les k catégories C_1, C_2, ..., C_k, alors

$$p_O = \frac{\Sigma\, n_{ii}}{n}$$

ou encore,

$$p_O = \Sigma\, p_{ii}$$

Tableau 9-2

		O_1			
		C_1	C_2	C_3	**Total**
O_2	C_1	30	0	0	**30**
	C_2	0	20	0	**20**
	C_3	0	0	10	**10**
	Total	**30**	**20**	**10**	**60**

Tableau 9-3

		O_1		
		C_1	C_2	**Total**
O_2	C_1	7	3	**10**
	C_2	2	8	**10**
	Total	**9**	**11**	**20**

Les symboles n_{ii} et p_{ii} trouvent leur signification aux tableaux 9-4 et 9-5.

Mesure d'accord véritable $p_O - P_C$

On peut toujours se demander si une partie des jugements concordants n'est pas due au hasard. Il est en effet possible d'imaginer que deux observateurs indépendants qui utiliseraient des jugements différents pour classer les sujets connaissent quand même un certain accord: c'est un accord par chance, fruit du hasard.

Différentes opinions se sont manifestées quant à l'utilité de tenir compte de l'élément hasard dans l'évaluation de l'accord entre deux observateurs. Il s'agit au fond d'une question d'attitude. On peut considérer que la valeur p_O, en comprenant les accords attribuables au hasard, exagère la proportion d'accords véritables. Suivant ce point de vue, la proportion globale d'accords véritables entre les deux observateurs serait inférieure à la proportion globale observée p_O. Si l'on admet que le hasard puisse jouer sur l'accord entre deux observateurs indépendants, il faut, pour mieux faire ressortir l'accord véritable, penser une mesure qui tente de soustraire l'effet-hasard. Une telle mesure a été proposée. Elle essaie de répondre aux deux questions suivantes :

1) Quelle est la *proportion globale d'accords par chance, c*'est-à-dire que le hasard seul suffirait à engendrer? Cette proportion est notée p_C (l'indice C réfère au terme *chance).*

2) Comment prendre en compte la proportion *pc* pour se faire une idée des accords véritables?

Considérons d'abord la première question. Le calcul de *pc* est fait ici dans le cadre où deux observateurs attribuent à un même sujet, de manière tout à fait indépendante, la même catégorie C_i. L'observateur O_1 attribue la catégorie C_i suivant une certaine probabilité (qui lui est propre); il en va de même de l'observateur O_2. Le comportement des observateurs O_1 et O_2 dans le classement des sujets est d'une certaine façon caractérisé respectivement par les probabilités Prob (^1C_i) et Prob (^2C_i). Les notations ^1C_i et ^2C_i indiquent la catégorie C_i attribuée respectivement par les observateurs O_1 et O_2. La probabilité qu'un sujet soit classifié dans la même catégorie C_i par les deux observateurs est donnée par: Prob ($^1C_i{}^2C_i$).

Tableau 9-4

		O_1				
		C_1	C_2	...	C_k	**Total**
O_2	C_1	n_{11}	n_{12}	...	n_{1k}	$n_{1.}$
	C_2	n_{21}	n_{22}	...	n_{2k}	$n_{2.}$
	⋮	⋮	⋮		⋮	
	C_k	n_{k1}	n_{k2}	...	n_{kk}	$n_{k.}$
	Total	$n_{.1}$	$n_{.2}$	...	$n_{.k}$	n

Tableau 9-5

		O_1				
		C_1	C_2	...	C_k	**Total**
O_2	C_1	p_{11}	p_{12}	...	p_{1k}	$p_{1.}$
	C_2	p_{21}	p_{22}	...	p_{2k}	$p_{2.}$
	⋮	⋮	⋮		⋮	
	C_k	p_{k1}	p_{k2}	...	p_{kk}	$p_{k.}$
	Total	$p_{.1}$	$p_{.2}$	...	$p_{.k}$	1

Si ces deux observateurs jugent de façon indépendante, on a, en vertu de la règle de multiplication pour les événements indépendants :

$$\text{Prob}\,({}^1C_i{}^2C_i) = \text{Prob}\,({}^1C_i)\ \text{Prob}\,({}^2C_i) \quad [*]$$

(cette règle est présentée au chapitre 10).

Lorsque les deux observateurs attribuent à un même sujet la même catégorie C_i, leurs jugements forment une *paire concordante.* S'il y a k catégories de classification possibles pour les sujets, il y a aussi k catégories de paires concordantes: ${}^1C_1 2C_1$, ${}^1C2{}^2C2$, ${}^1C_k{}^2C_k$. La probabilité que les deux observateurs attribuent une même catégorie à un même sujet est donc donnée par:

$$\text{Prob}\,({}^1C_1{}^2C_1 \text{ ou } {}^1C_2{}^2C_2 \text{ ou } \ldots \text{ ou } {}^1C_k{}^2C_k).$$

Parce que les k catégories sont mutuellement exclusives et en vertu de la règle d'addition des probabilités (que nous verrons aussi au chapitre 10), cette probabilité est égale à:

$$\text{Prob}({}^1C_1{}^2C_1) + \text{Prob}({}^1C_2{}^2C_2) + \ldots + \text{Prob}({}^1C_k{}^2C_k),$$

ce qui s'écrit:

$$\sum_{i=1}^{k} \text{Prob}\,({}^1C_i{}^2C_i).$$

En vertu de la règle [*] cette dernière expression est équivalente à :

$$\sum_{i=1}^{k} \text{Prob}\,({}^1C_i)\ \text{Prob}\,({}^2C_i)$$

Cette probabilité-sommation permet d'estimer la proportion globale p_C, c'est-à-dire la proportion de paires qui peuvent concorder par chance. Pour les deux observateurs des tableaux 9-1, 9-2 et 9-3, on obtient:

$$p_C = \left(\frac{20}{60} \times \frac{20}{60}\right) + \left(\frac{20}{60} \times \frac{20}{60}\right) + \left(\frac{20}{60} \times \frac{20}{60}\right)$$
$$= \tfrac{1}{3} = 0{,}33 \qquad \text{[tableau 9-1]}$$

$$p_C = \left(\frac{30}{60} \times \frac{30}{60}\right) + \left(\frac{20}{60} \times \frac{20}{60}\right) + \left(\frac{10}{60} \times \frac{10}{60}\right)$$
$$= \tfrac{14}{36} = 0{,}39 \qquad \text{[tableau 9-2]}$$

$$p_C = \left(\frac{9}{20} \times \frac{10}{20}\right) + \left(\frac{11}{20} \times \frac{10}{20}\right)$$
$$= \tfrac{1}{2} = 0{,}50 \qquad \text{[tableau 9-3]}$$

Considérons maintenant la deuxième question relative à la prise en compte de p_C. Il existe différentes façons de corriger p_O de l'effet attribuable au hasard. Nous retenons celle, assez naturelle, de soustraire p_C de p_O. La différence :

$$P_O - P_C$$

est une mesure possible d'accords véritables. Pour les deux observateurs des tableaux 9-1, 9-2 et 9-3, on obtient:

$$p_O - p_C = 0 - 0{,}33 = -0{,}33 \quad \text{[tableau 9-1]}$$
$$= 1 - 0{,}39 = 0{,}61 \quad \text{[tableau 9-2]}$$
$$= 0{,}75 - 0{,}50 = 0{,}25 \quad \text{[tableau 9-3]}$$

Quand on compare p_O et p_C, trois possibilités peuvent se présenter :

1) $p_O > p_C$: Les deux observateurs font mieux que le hasard; il existe un certain accord véritable.

2) $p_O = p_C$: Les deux observateurs ne font pas mieux que le hasard; il n'y a pas d'accord véritable.

3) $p_O < p_C$: Les deux observateurs font pire que le hasard.

Nous allons discuter les valeurs maximales de p_O et $p_O - p_C$ en nous aidant des exemples des tableaux 9-2 et 9-3.

La valeur de p_O est obligatoirement située entre 0 et 1. La valeur maximale de p_O dépend de la composition des marges du tableau. La valeur 1 ne peut être atteinte qu'en présence de marges identiques; c'est le cas au tableau 9-2. Au tableau 9-3, la valeur observée de po est 0,75. Dans ce tableau, la valeur maximale de po est inférieure à 1; elle est égale à 0,95 comme l'illustre le tableau 9-6 (ce tableau reprend le tableau 9-3 en gardant les marges fixes et en maximisant les accords entre les deux observateurs).

On rappelle que la valeur de p_C dépend aussi des deux marges, comme il a été fait pour les tableaux 9-1, 9-2 et 9-3. La proportion p_C est calculée aux conditions des marges fixes.

Tableau 9-6

		O_1		
		C_1	C_2	**Total**
O_2	C_1	9	1	**10**
	C_2	0	10	**10**
	Total	**9**	**11**	**20**

Puisque p_O doit se situer obligatoirement entre 0 et 1, quelles que soient les conditions, la valeur de $p_O - p_C$ doit être comprise entre les valeurs $-p_C$ et $1 - p_C$. Cette valeur $1 - p_C$ n'est atteinte qu'en présence de marges identiques; c'est le cas au tableau 9-2. Autrement, la valeur maximale que peut atteindre $p_O - p_C$ est inférieure à $1 - p_C$ et varie suivant la composition des marges. Au tableau 9-3, la valeur maximale de p_O étant 0,95, celle de $p_O - p_C$ est égale à 0,45 (0,95 – 0,50).

Mesure d'accord kappa (x)

Au tableau 9-3, nous avons obtenu une mesure d'accord véritable $p_O - p_C$ égale à 0,25 (0,75 – 0,50). La valeur est positive. Il y a un certain accord véritable. Est-ce un bon accord? Pour en juger numériquement, il serait intéressant de comparer la valeur obtenue 0,25 à la valeur maximale 0,45.

Comme on l'a vu, la valeur maximale de $p_O - p_C$ varie avec la composition des marges. Cela rend son calcul difficile. Le problème est d'autant plus complexe qu'il y a de catégories de classification. Il sera plus simple de comparer la valeur obtenue de $p_O - p_C$ à la valeur maximale dans toute condition, c'est-à-dire à $1 - p_C$.

Pour l'exemple du tableau 9-3:

$$\varkappa = \frac{0{,}75 - 0{,}50}{1 - 0{,}50} = 0{,}50.$$

Cette mesure x nous renseigne sur la fraction d'accords véritables. Pour un tableau donné, le dénominateur 1 — p_C est généralement supérieur à la valeur maximale de p_O — p_C. De ce fait, x peut être considéré comme une mesure conservatrice de la fraction d'accords véritables.

Puisque $p_O - p_C$ varie entre $-p_C$ et $1 - p_C$ dans toute condition, la mesure $\varkappa$ prend des valeurs comprises entre $\frac{-p_C}{1 - p_C}$ et 1. Les valeurs de kappa peuvent être négatives, nulles ou positives. Si l'une ou l'autre des deux premières possibilités se présente, le niveau d'accord véritable est très mauvais. Si la valeur est positive, il passe par toutes les nuances. Proche de 0, le niveau d'accord véritable est mauvais; proche de 1, il est excellent. Le seuil acceptable pour la valeur x dépend du degré de fiabilité que l'on veut atteindre dans le processus de classification des individus. Ce degré de fiabilité est lui-même fonction du critère de classification choisi. Un critère objectif devrait conduire à une grande fiabilité, donc à un seuil élevé pour x Pour qu'un niveau d'accord véritable soit qualifié d'excellent, avec un critère subjectif, il pourra peut-être suffire que la valeur de x soit égale ou supérieure à 0,75. Par ailleurs, avec un critère plus objectif, il faudra peut-être atteindre une valeur plus grande avant de considérer le niveau acceptable. En complément au calcul de la valeur x, il peut être intéressant pour un chercheur d'examiner les caractéristiques des paires discordantes, d'investiguer les raisons de désaccord entre les observateurs.

MESURE DE L'ACCORD ENTRE DEUX OBSERVATEURS DANS LE CAS D'UNE APPRÉCIATION QUANTITATIVE

Là aussi, comme pour une appréciation qualitative, il ne faut pas confondre l'association et l'accord. Au tableau 9-7, les jugements quantitatifs (nous dirons éventuellement scores) des deux observateurs portés sur chacun des cinq sujets présentent un certain désaccord quoique leur association soit parfaite. Si l'on connaît le score attribué à un sujet par l'observateur O_1, celui donné par l'observateur O_2 au même sujet est entièrement prévisible. C'est toujours 4 en moins. Par ailleurs, il y a désaccord puisque les scores entre les observateurs ne coïncident jamais.

Mesure d'accord par le coefficient de corrélation intra-classe

Il faut éliminer comme mesure du degré d'accord le coefficient de corrélation linéaire r défini au chapitre 6. Ce coefficient est incapable de distinguer les degrés d'accord pourtant différents que l'on trouve par exemple aux tableaux 9-7 et 9-8, où r prend la même valeur 1. Le coefficient r rend uniquement compte du degré d'association, ici de l'association parfaite de par sa valeur 1.

Tableau 9-7

Sujet	O_1	O_2
1	69	65
2	79	75
3	74	70
4	80	76
5	68	64

Une des mesures du degré d'accord est le coefficient de corrélation intra-classe dont la définition fait appel à l'idée de décomposition de la variance. Pour apprécier cette idée, revenons au tableau 9-7. On peut imaginer une décomposition de la variation totale des 10 scores. Dans les limites de la décomposition la plus simple, on peut identifier le sujet comme source de variation. Quel que soit l'observateur, les scores varient d'un sujet à l'autre. Ainsi, une partie de la variation totale entre les 10 scores est attribuable aux sujets. C'est la *variation inter-sujet,* composante de la variation totale. Par ailleurs, tout le reste de la variation totale qui ne s'explique pas par la variation inter-sujet est appelé *variation résiduelle.*

La variation inter-sujet correspond à une sorte d'effet-sujet dont le calcul, pour chacun des sujets, nécessite au préalable que soient neutralisées les différences entre observateurs. Cela est rendu possible en considérant par exemple la moyenne des scores par sujet. Selon une façon très habituelle, la différence entre score moyen par sujet et score moyen général peut servir à mesurer l'effet-sujet. Il est égal à 67 – 72, soit (-5), pour le premier sujet au tableau 9-7.

Tableau 9-8

Sujet	O_1	O_2
1	69	69
2	79	79
3	74	74
4	80	80
5	68	68

Le score attribué à un sujet par un observateur se développe ainsi autour d'un score moyen général auquel s'ajoutent l'effet-sujet et une part résiduelle. Toujours au tableau 9-7, le score 69 (sujet n° 1, observateur O_1) est décomposable comme suit:

$$69 = 72 - 5 + 2.$$

De façon générale, on a la décomposition additive :

score = score moyen général + effet-sujet + résidu.

Au tableau 9-8, où l'accord est parfait, le résidu est nul quel que soit le sujet. Le score 69 (sujet n° 1, observateur O_1) se décompose comme suit:

$$69 = 74 - 5 + 0.$$

Le degré d'accord entre deux observateurs se reflète donc dans le rapport qui peut exister entre la composante « sujet » (effet-sujet) et la composante « résidu ». En première analyse, on peut comprendre que si la composante « sujet » explique la totalité de la variation, l'accord entre les deux observateurs est parfait. Une plus grande part de la variation résiduelle dans l'explication de la variation totale correspond à une moins bonne concordance entre les deux observateurs. Pour en juger davantage, nous allons considérer les trois exemples des tableaux 9-9A, 9-9B et 9-9C, différents par leur degré d'accord. Chaque tableau indique les effets-sujet et les valeurs résiduelles.

Au tableau 9-9A, où l'accord est parfait, la variation totale des scores est entièrement absorbée par la variation inter-sujet. Au tableau 9-9B, où l'accord est plutôt très mau-

vais, la variation totale n'est en rien expliquée par la variation inter-sujet; la variation des résidus englobe complètement la variation des scores. Au tableau 9-9c, la variation totale des scores est portée pour une part par la variation inter-sujet, pour le reste par celle des résidus.

Tableau 9-9A (accord parfait)

Score			Effet-sujet		Résidus	
Sujet	O_1	O_2	O_1	O_2	O_1	O_2
1	18	18	2	2	0	0
2	19	19	3	3	0	0
3	15	15	-1	-1	0	0
4	12	12	-4	-4	0	0
5	16	16	0	0	0	0

(Moyenne générale = 16)

Tableau 9-9B (désaccord total)

	Score		Effet-sujet		Résidus	
Sujet	O_1	O_2	O_1	O_2	O_1	O_2
1	18	4	0	0	7	-7
2	4	18	0	0	-7	7
3	18	4	0	0	7	-7
4	18	4	0	0	7	-7
5	4	18	0	0	-7	7

(Moyenne générale = 11)

Tableau 9-9C (accord quelconque)

Score			Effet-sujet		Résidus	
Sujet	O_1	O_2	O_1	O_2	O_1	O_2
1	18	12	0	0	3	-3
2	19	19	4	4	0	0
3	15	13	-1	-1	1	-1
4	12	14	-2	-2	-1	1
5	16	12	-1	-1	2	-2

(Moyenne générale = 15)

On observe que le degré d'absorption de la variation totale (variation inter-sujet + variation résiduelle) par la variation inter-sujet suit le degré d'accord. Cette constatation permet d'imaginer le rapport entre la variation inter-sujet et la variation totale comme une mesure d'accord. La formule suivante exprime un tel rapport.

$$\varrho_I = \frac{\text{variation inter-sujet}}{\text{variation inter-sujet} + \text{variation résiduelle}}$$

Le rapport ϱ_I définit une mesure d'accord connue comme le *coefficient de corrélation intra-classe*. Cette appellation peut surprendre. C'est qu'il existe une autre définition de ϱ_I qui le présente comme un coefficient de corrélation d'un certain genre. Donner l'autre définition nous obligerait à expliquer ce qu'il faut entendre par tableau intra-classe (le mot classe renvoie dans notre contexte à sujet: à chaque sujet correspondent plusieurs scores). L'exercice serait superflu. De toute manière, on peut montrer formellement l'équivalence des deux définitions.

Le coefficient ϱ_I prend des valeurs théoriquement comprises entre 0 et 1. Plus sa valeur est proche de 1, meilleur est l'accord entre les deux observateurs. À l'inverse, l'accord est d'autant plus mauvais que la valeur de ϱ_I est proche de 0. Comme nous allons maintenant le voir, l'estimation de ϱ_I peut toutefois en pratique donner lieu à des valeurs négatives. C'est tout simplement le signe d'un très mauvais accord.

Calcul d'une estimation pour ϱ_I

C'est à des calculs de variances en tant qu'elles sont des mesures statistiques de la

variation, qu'il faut se prêter pour obtenir une estimation du coefficient de corrélation intra-classe. Rappelons qu'une variance se compose au numérateur d'une somme des carrés des écarts à une moyenne (SC) et au dénominateur d'un nombre qu'on appelle le degré de liberté (dl).

Pour l'ensemble des 10 scores au tableau 9-9C, la somme des carrés totale *(SCT)* est égale à 74. En effet:

$$SCT = (18 - 15)^2 + (19 - 15)^2 + \ldots + (12 - 15)^2 = 74.$$

Suivant les deux sources de variation déjà identifiées, la somme *SCT* est décomposable en SCE (la somme des carrés des effets-sujet) et en SCR (la somme des carrés des résidus). On peut noter cette décomposition au tableau 9-9C où :

$$SCE = 2(0)^2 + 2(4)^2 + 2(-1)^2 + 2(-2)^2 + 2(-1)^2 = 44 = \Sigma \text{ carré des effets-sujets;}$$

$$SCR = (3)^2 + (-3)^2 + \ldots + (2)^2 + (-2)^2 = 30 = \Sigma \text{ carré des résidus.}$$

Ce résultat est général. On a toujours $SCT = SCE + SCR$.

Le terme au dénominateur d'une variance est appelé *degré de liberté ou nombre de degrés de liberté.* C'est en premier lieu un nom. Par exemple, si d'une série de trois valeurs on connaît la moyenne et deux valeurs, alors la troisième valeur de cette série est entièrement déterminée. Si la moyenne est égale à 35 et deux des valeurs égales à 30 et 44, la troisième valeur est obligatoirement 31. Une des trois valeurs n'est pas libre. On dit de la série des trois valeurs qu'elle a deux degrés de liberté. L'analogie suivante peut aider à comprendre davantage. Soient trois chaises libres dans une salle. Le premier individu à pénétrer dans la salle peut choisir sa chaise. Un choix est encore possible pour le deuxième individu, par contre, la troisième personne n'a plus de choix. Les trois chaises forment un ensemble, pourrions-nous dire, à deux degrés de liberté.

Quels sont les degrés de liberté rattachés aux sommes SCT, SCE et SCR? Avec n sujets et 2 observations par sujet, donc $2n$ scores, les degrés de liberté sont $2n - 1$ pour SCT, $n - 1$ pour SCE et n pour SCR. On peut facilement s'en rendre compte en considérant à nouveau l'exemple au tableau 9-9C. Les 10 scores (2 x 5) de moyenne connue forment une série à 9 degrés de liberté (dl = 2 x 5 — 1). Relativement à la somme SCE, il y a 4 degrés de liberté ($dl = 5 - 1$). La série des 10 effets-sujet est entièrement déterminée dès que sont connus 4 d'entre eux. Si on connaît les 4 valeurs encadrées du tableau 9-10 qui reprend les effets-sujet du tableau 9-9C, les 6 autres sont connues. Pour la somme SCR, il y a 5 degrés de liberté *(dl = 5).* Les résidus s'échafaudent sujet par sujet. Pour chaque sujet, il y a deux scores, donc 1 degré de liberté. Pour les cinq sujets, c'est 5 x 1 degrés de liberté ($dl = 5$).

Tableau 9-10

Effet-sujet	
O_1	O_2
0	0
4	4
−1	−1
−2	−2
−1	−1

Comme pour les sommes SC, les degrés de liberté s'additionnent.

$$dl_{SCT} = dl_{SCE} + dl_{SCR}$$

$$(2n - 1) = (n - 1) + n$$

La contribution respective des deux composantes de la variation totale est souvent présentée dans un tableau dit d'analyse de variance, comme au tableau 9-11. C'est ici un tableau à un facteur, car le sujet est le seul facteur qui a été formellement identifié comme source de variation.

Les éléments *CME* et *CMR* entrent dans le calcul pour l'estimation des variances inter-sujet et résiduelle. C'est par un détour un peu théorique qu'on arrive à savoir comment utiliser les carrés moyens dans le calcul de ces variances. Nous vous invitons à croire, sans démonstration, que les variances inter-sujet et résiduelle sont respectivement estimées par $\frac{CME - CMR}{2}$ et *CMR*. On en déduit que le coefficient de corrélation intra-classe ϱ_I est estimé par:

$$r_I = \frac{\frac{CME - CMR}{2}}{\frac{CME - CMR}{2} + CMR}$$

soit,

$$r_I = \frac{CME - CMR}{CME + CMR}$$

Tableau 9-11 (analyse de variance)

Source de variation	*SC*	*dl*	*CM* (carré moyen)
Inter-sujet *SCE*		*n–1*	*CME = SCEI* (*n* -1)
Résiduelle *SCR*		*n*	*CMR = SCR/n*
Total	***SCT***	***2n—1***	***CMT = SCT/ (2n* -1)**

Les tableaux d'analyse de variance 9-12A, 9-12B et 9-12C se rapportent, dans l'ordre, aux exemples des tableaux 9-9A, 9-9B et 9-9C.

Le tableau 9-12A, où le carré moyen *CMR = 0,* correspond à l'exemple des deux observateurs en parfait accord. Leur accord parfait conduit précisément à l'absence de variation résiduelle et en conséquence, $r_I = 1$.

Le tableau 9-12B, où on a plutôt *CME* = 0, découle de l'exemple du désaccord total entre les deux observateurs. Le désaccord est

Tableau 9-12A (analyse de variance)

Source de variation	*SC*	*dl*	*CM*	
Inter-sujet	60	4	15	$r_I = \frac{15 - 0}{15 + 0} = 1$
Résiduelle	0	5	0	
Total	**60**	**9**		

Tableau 9-12B

Source de variation	SC	*dl*	*CM*	
Inter-sujet	0	4	0	$r_I = \frac{0 - 98}{0 + 98} = -1$
Résiduelle	490	5	98	
Total	**490**	**9**		

Tableau 9-12C

Source de variation	*SC*	*dl*	*CM*	
Inter-sujet	44	4	11	$r_I = \frac{11 - 6}{11 + 6} = 0{,}294$
Résiduelle	30	5	6	
Total	74	**9**		

total lorsque la variation résiduelle absorbe complètement la variation totale. En parfait désaccord, $r_I = -1$.

Ces deux valeurs, – 1 et 1, limitent inférieurement et supérieurement le coefficient de corrélation intra-classe observé r_I. Les valeurs négatives ou nulles pour r_I sont l'écho d'un très mauvais accord puisqu'elles résultent d'un carré moyen *CME* au plus égal au carré *CMR,* ce qui témoigne d'une variation résiduelle numériquement plus importante que la variation inter-sujet. Pour un coefficient r_I entre 0 et 1, tous les degrés d'accord sont permis, du plus médiocre à l'excellent en passant par un degré modéré. Au tableau 9-12C où $r_I = 0{,}294$, l'accord penche du côté médiocre. Mais où se termine l'accord médiocre, où commence l'excellent accord? Qu'entend-on par accord modéré? Les réponses à ces questions dépendent du problème étudié.

RÉSUMÉ

Le jugement de deux observateurs indépendants a été distingué selon qu'il est qualitatif ou quantitatif. Il est qualitatif lorsque les sujets sont classés dans des catégories fixées, mutuellement exclusives et collectivement exhaustives. Le jugement est quantitatif quand l'observateur attribue à chaque sujet une valeur numérique (un score). L'accord de jugements entre deux observateurs ne doit pas être confondu avec l'association de leurs jugements. On peut observer une complète association de leurs jugements sans qu'il existe un quelconque accord. Pour mesurer l'accord entre deux observateurs lorsque leur appréciation est qualitative, la mesure kappa (*x*) est proposée; elle nous renseigne sur une fraction d'accords véritables (épurés des accords par chance) à laquelle sont parvenus les deux observateurs. Si $\varkappa$ prend une valeur négative ou nulle, le niveau d'accord véritable est très mauvais; si la valeur est positive, le niveau peut aller de médiocre à excellent (pour $\varkappa$ proche de 1). Pour mesurer l'accord entre deux observateurs indépendants lorsque leur appréciation est quantitative, la mesure du coefficient de corrélation intra-classe ϱ_I est proposée. Ce coefficient fait appel à l'idée de la décomposition de la variation totale des scores attribués aux différents sujets. L'accord est parfait si r_I est égal à 1. Le désaccord est total lorsque r_I est égal à -1. Les valeurs négatives ou nulles pour r_I sont l'écho d'un très mauvais accord. Pour un coefficient r_I compris entre 0 et 1, les degrés d'accord vont de médiocre à excellent.

Symboles

p_O: proportion globale observée de jugements convergents (proportion d'accords observés)

p_C: proportion globale d'accords par chance

$\varkappa$: mesure d'accord kappa

ϱ_I, r_I: coefficient de corrélation intra-classe théorique, son estimation

SCE: somme des carrés (des écarts) des effets-sujet

SCR: somme des carrés (des écarts) des résidus

SCT: somme des carrés (des écarts) totale

Formules

$$\varkappa = \frac{p_O - p_C}{1 - p_C}$$

$$\varrho_I = \frac{\text{variation inter-sujet}}{\text{variation inter-sujet} + \text{variation résiduelle}}$$

$$r_I = \frac{CME - CMR}{CME + CMR}$$

$$où \quad CME = \frac{SCE}{n - 1}$$

$$et \quad CMR = \frac{SCR}{n}$$

LECTURE SUGGÉRÉE

FLEISS, J.L. ***Statistical Methods for Rates and Proportions,*** New York, John Wiley and Sons, 1981, chapitre 13, pp.212-236.

PARTIE IV

Mesures de probabilité

CHAPITRE 10

Mesure de probabilité

Les phénomènes aléatoires se rencontrent fréquemment en médecine. Leur compréhension conduit aux notions d'événement et de probabilité. Par exemple, on peut s'intéresser à la probabilité pour un patient de survivre cinq ans à une certaine maladie. Ce chapitre traite des notions d'expérience aléatoire et d'événement, de mesure de probabilité en général. On y explique l'interprétation fréquentiste de la probabilité, on y établit les règles d'addition et de multiplication du calcul de probabilités, ainsi que la formule sur la probabilité des causes, ou formule de Bayes.

L'exposition d'un individu à un contaminant peut conduire à l'apparition d'un cancer. Certains individus développeront le cancer, d'autres pas.

La survie à cinq ans pour un individu atteint d'une certaine maladie n'est pas assurée d'avance. Le type de maladie, son stade d'évolution, l'état physique et psychique du malade sont autant de facteurs qui pourront influencer la survie de ce dernier.

Un traitement peut avoir des effets pas toujours souhaités. Certains patients réagiront positivement à la médication, tandis que la condition de certains autres pourra continuer à se détériorer, allant même jusqu'au décès.

Plusieurs possibilités diagnostiques peuvent exister pour un même problème ressenti chez un patient. L'état d'un patient, la précision des symptômes rapportés par celui-ci, la qualité et la pertinence des tests de laboratoire, conjugués à l'expérience du médecin, sont autant de facteurs qui peuvent agir sur le diagnostic qui sera posé. Établir qu'un patient a telle maladie est un acte souvent nuancé.

Toutes ces situations médicales décrites sont caractérisées par l'incertitude. D'un certain point de vue, on peut les comprendre comme des expériences dont le résultat comporte une certaine part d'imprévisible. Un individu exposé à un certain contaminant développera-t-il le cancer? Le traitement appliqué au patient produira-t-il l'effet désiré? Le calcul des probabilités concerne l'ensemble des lois et des règles qui permettent justement de mesurer le degré de certitude (ou d'incertitude) qui accompagne un résultat (événement) incertain. Ces outils sont indispensables tant au raisonnement du chercheur en médecine qu'au discernement du clinicien dans sa pratique; ils sont aussi nécessaires à la compréhension des phénomènes aléatoires liés à la maladie ou à la santé.

EXPÉRIENCE ALÉATOIRE

Chaque situation de recherche épidémiologique, d'expérimentation ou d'observation, chaque situation clinique, d'observation ou de prise de décision, peut être comprise comme une certaine *expérience aléatoire,* c'est-à-dire dont le résultat comporte de l'incertitude. Il ne faut pas se choquer du terme expérience si on l'entend comme un processus qui implique une intervention de la part de l'homme. Cette intervention peut être de l'ordre de l'expérimentation ou de la simple observation du phénomène produit par la nature.

Caractéristiques d'une expérience aléatoire

Observer un patient depuis le début d'un traitement pour savoir s'il va survivre cinq ans à sa maladie relève d'un phénomène aléatoire. Les phénomènes ou expériences aléatoires sont essentiellement caractérisés par l'absence d'une complète certitude quant à la réalisation des événements qui s'y rapportent. Aléatoire s'oppose à certain. Les expériences aléatoires ont toutes cette propriété commune qui les caractérise : répétée un certain nombre de fois dans des conditions identiques, l'expérience n'engendre pas nécessairement le même résultat d'un essai à l'autre. Un phénomène qui produit toujours le même résultat correspond, d'un certain point de vue, à une situation extrême. Ce phénomène est *déterministe.*

L'expérience aléatoire est donc caractérisée par les deux conditions suivantes:

1) on ne peut pas prévoir avec certitude (avant l'expérience) le résultat de l'expérience, mais ce résultat est clairement identifiable;
2) on peut décrire (avant l'expérience) l'ensemble de tous les résultats possibles.

Résultat possible

Cette caractérisation de l'expérience aléatoire implique la notion de résultat possible. L'identification ou la définition des résultats possibles pour une expérience aléatoire dépend essentiellement du but que l'on se propose d'atteindre dans la conduite de cette expérience. On observe deux patients qui ont la maladie *M,* depuis le moment du diagnostic de M jusqu'au décès de chacun. Cette expérience aléatoire peut conduire à différents types de résultats. Certains intéressent plus particulièrement l'investigateur.

S'intéresse-t-on au nombre entier de jours de la durée de survie la plus longue? On appelle alors résultat tout nombre entier compris entre 0 et un certain nombre entier positif limite *n* au delà duquel il est raisonnablement impossible d'observer une valeur de durée de survie, (peut-être 20 000 jours). Suivant cette définition de résultat, cette expérience engendre un nombre fini n de résultats possibles.

Peut-être s'intéresse-t-on au décès par la maladie *M?* Alors on appelle résultat pour cette expérience un des couples (x,y) où x et y représentent respectivement pour le premier et le deuxième patient l'un des deux événements: décédé par *M (D +),* non-décédé par *M* (D –). Suivant cette définition de résultat, cette expérience aléatoire a quatre résultats possibles:

1) *(D+, D+)*
2) *(D+, D–)*
3) *(D–, D+)*
4) *(D–, D–)*

On peut choisir un exemple encore plus classique, celui du lancer d'un dé. Si l'expérience aléatoire consiste à lancer un dé deux fois de suite, on peut alors définir plusieurs types de résultats. Si on s'intéresse au plus grand nombre de points obtenus sur le dé lors de l'un de ces deux lancers, on appelle résultat possible l'un des nombres de l'ensemble {l, 2, 3, 4, 5, 6}. Si on s'intéresse à la somme des points obtenus sur le dé pour les deux lancers, on appelle résultat possible l'un des nombres de l'ensemble {2, 3, 4, ..., 10, 11, 12}. Si on s'intéresse au nombre de points obtenus sur le dé lors de chacun des lancers, on appelle résultat possible tout couple de nombre (x,y) où x et *y* sont des éléments de l'ensemble {1, 2, 3, 4, 5, 6}. On dénombre alors 36 résultats possibles :

(1,1), (1,2), ..., (6,5), (6,6).

Ensemble fondamental

Pour une expérience aléatoire, l'ensemble de tous les résultats possibles est appelé *ensemble fondamental* et est généralement désigné par 11. Pour l'exemple du double lancer d'un dé, on a, suivant la question d'intérêt, les ensembles fondamentaux suivants :

$\Omega_1 = \{1, 2, 3, 4, 5, 6\}$
$\Omega_2 = \{2, 3, 4, \ldots, 10, 11, 12\}$
$\Omega_3 = \{(1,1), (1,2), (1,3), \ldots, (6,4), (6,5), (6,6)\}.$

Pour l'exemple des deux patients, on a, suivant la question d'intérêt, les ensembles fondamentaux suivants :

$$\Omega_{\text{jours}} = \{0, 1, 2, \ldots, n\}$$
$$\Omega_{\text{décès}} = \{(D+, D+), (D+, D-), (D-, D+), (D-, D-)\}.$$

ÉVÉNEMENT

Quand on observe un phénomène ou que l'on étudie une expérience aléatoire, souvent on s'intéresse non pas à un résultat particulier, mais plutôt à une combinaison de plusieurs résultats possibles. Par exemple, dans l'observation des deux patients, on peut s'intéresser au fait qu'au moins l'un des deux décède de la maladie *M.* Dans le double lancer d'un dé qui conduit aux résultats simples des couples (x,y), on peut s'intéresser aux couples dont x est pair et *y* est impair: (2,1), (2,3), ..., (6,3), (6,5).

Définition d'événement

On est ainsi conduit à la définition d'un *événement* comme tout sous-ensemble de l'ensemble fondamental. L'ensemble

$$\{(D+, D+), (D+, D-), (D-, D+)\}$$

décrit l'événement « qu'au moins l'un des deux patients soit décédé par *M* ». Cet événement est composé de résultats de l'ensemble fondamental $\Omega_{\text{décès}}$.

L'ensemble {(2,1), (2,3), ..., (6,3), (6,5)} décrit l'événement « le premier lancer du dé est un nombre pair et le second un nombre impair ». Cet événement est composé de résultats de l'ensemble fondamental Ω_3.

Ainsi, peut-on distinguer deux grands types d'événements: ceux qui ne se composent que d'un seul résultat possible (ce sont les événements élémentaires) et ceux qui se composent de plusieurs résultats élémentaires (ce sont les événements composés). Pour l'exemple du double lancer d'un dé où l'on s'intéresse aux couples de nombres (x,y) du premier et du second lancer, l'événement {(1,1)} est un événement élémentaire. Il ne se réalise que si le résultat (1,1) se réalise. Par contre, l'événement « observer des nombres égaux au premier et au second lancer » est un événement composé des résultats (1,1), (2,2), (3,3), (4,4), (5,5) et (6,6). Si l'un de ces résultats se réalise, alors on peut dire que l'événement se réalise.

Avant d'aborder plus à fond la notion d'événement, il est important d'identifier de façon particulière certains événements spécifiques, non pas à cause de l'intérêt qu'ils suscitent du point de vue de l'observateur, mais plutôt à cause du rôle important qu'ils jouent dans la composition des événements et, conséquemment, dans le calcul des probabilités. Il s'agit de l'événement nul (ou impossible) et de l'événement certain. L'événement nul, que l'on désigne par Ø correspond à l'événement qui ne se réalise jamais. Il ne renferme aucun résultat possible. L'événement certain correspond à l'événement qui se réalise toujours. Il renferme la totalité des résultats possibles et coïncide donc avec l'ensemble fondamental Ω.

Composition d'événements

Il est possible, à partir d'événements et d'opérations de base, de composer de nouveaux événements. Ces opérations de base utilisent des mots du langage courant comme *ou, et, ne pas, si.*

COMPOSITION PAR DISJONCTION

La composition par disjonction est celle faite à partir du mot clé « ou ». Pour l'expérience du double lancer d'un dé, « obtenir des nombres égaux au premier et au second lancer » est un événement composé à partir de résultats élémentaires et de la particule « ou ». Obtenir des nombres égaux, c'est obtenir (1,1) ou (2,2) ou... ou (6,6). Autrement dit, la particule « ou » a servi à composer l'événement « obtenir des nombres égaux au premier et second lancer ». De façon générale, si *A* et *B* représentent deux événements, *(A* ou *B)* est l'événement qui se réalise si au moins un des deux événements *A* et *B* se réalise. C'est l'événement disjonction; il renferme tous les résultats élémentaires qui appartiennent à au moins l'un des deux événements *A* et *B.* Un investigateur est intéressé à l'événement pour une personne d'être atteinte de la maladie *A* ou de la maladie *B.* C'est l'événement *(A* ou *B).* Le « ou » doit être pris dans un sens inclusif, c'est-à-dire qu'une personne peut développer la maladie *A* uniquement, la maladie *B* uniquement, ou les deux.

COMPOSITION PAR CONJONCTION

La composition par conjonction se fait par la particule « et ». L'événement conjonction de *A* et de *B,* désigné par *(A* et *B),* est celui qui se réalise lorsque et *A* et *B* se réalisent. C'est l'événement conjonction; il renferme tous les résultats élémentaires qui appartiennent à la fois à *A* et à *B.* Être positif à un premier test *et* l'être à un deuxième au cours d'un examen clinique en est un exemple.

COMPOSITION PAR NÉGATION

Étant donné un événement *A,* la négation « ne pas » permet de définir l'événement non *A,* habituellement noté *A.* C'est l'événement qui se réalise si et seulement si *A* ne se réalise pas; il renferme tous les résultats élémentaires qui ne sont pas dans *A.* Si *A* est l'événement « être positif à un test diagnostique » alors l'événement *A* est donné par être négatif au test ». Si l'événement *A* est celui « d'être encore vivant après une certaine période d'observation », alors l'événement *A* est celui « d'être décédé au cours de la même période ».

COMPOSITION PAR RÉSTRICTION

Une autre composition est possible avec l'utilisation de la particule « si ». Admettons que l'on a deux événements *A* et *B,* l'événement « *A* si *B » ou « A* sachant *B* », souvent noté *(A/B),* désigne l'événement *A* dans la condition préalable de la réalisation de *B.* C'est comme si l'ensemble fondamental était restreint à l'événement *B* et que la réalisation de *A* n'était jugée que sur *B* (la condition). Être affecté d'un cancer du poumon si on est un fumeur en est un exemple. C'est un *événement conditionnel,* la condition étant d'être un fumeur. On peut dire que cet événement est composé par restriction, la restriction étant d'être un fumeur.

Toutes les opérations de composition dont il a été question, à l'exception de la négation (on le comprendra), peuvent être pratiquées avec deux événements ou plus.

Quelques relations entre événements

Il existe entre les événements certaines relations fondamentales qui peuvent déterminer leur réalisation. Ce sont principalement les relations de complémentarité, d'incompatibilité et de dépendance (ou d'indépendance). Nous en donnons ici des définitions intuitives, qui seront reprises plus loin dans un langage plus formel.

COMPLÉMENTARITÉ

Deux événements sont *complémentaires si* la seule façon pour que l'un se réalise est que l'autre ne se réalise pas. On parle aussi *d'événements contraires.* C'est le cas de l'événement certain et de l'événement nul. Puisque l'événement certain se réalise toujours, l'événement nul ne se réalise jamais. C'est le cas aussi de A et $\bar{A}$: si A ne s'est pas réalisé, on est sûr que $\bar{A}$ s'est réalisé et inversement. Si une naissance est masculine, on est sûr qu'elle n'est pas féminine. Si on observe une face paire au lancer d'un dé, on est sûr qu'elle n'est pas impaire. L'événement A est composé de tous les résultats possibles qui ne sont pas dans A.

INCOMPATIBILITÉ

Deux événements sont dits *incompatibles ou mutuellement exclusifs* s'ils ne peuvent pas se réaliser en même temps. Si l'un se réalise, on est sûr que l'autre ne s'est pas réalisé. Mais il se peut très bien que ni l'un ni l'autre ne se réalise. Être femme et souffrir du cancer de la prostate sont deux événements incompatibles. Au lancer d'un dé, observer un nombre inférieur à 3 et observer un nombre supérieur à 4 sont deux événements incompatibles. Si l'on observe 2, alors le premier événement s'est réalisé et non le second. Par contre, si l'on observe 3, ni le premier ni le second ne s'est réalisé. Deux événements incompatibles correspondent à des ensembles de résultats possibles complètement différents: si un résultat possible se retrouve dans un événement, il ne peut pas être dans l'autre événement qui lui est incompatible. Un cas particulier: deux événements contraires ou complémentaires sont toujours incompatibles. Les événements A et $\bar{A}$ ne peuvent pas se réaliser en même temps.

DÉPENDANCE (OU INDÉPENDANCE)

Deux événements sont dits *dépendants si* la réalisation de l'un est influencée par la réalisation de l'autre, autrement ils sont dits *indépendants.* Observer un fumeur et observer un cancer du poumon sont deux événements dépendants. La réalisation de l'événement cancer du poumon est modifiée par le fait que l'observation est faite chez un fumeur. On a une plus grande chance d'observer un cancer du poumon chez un fumeur que d'observer ce même cancer chez un individu tiré de la population générale. Par contre, suivant les connaissances actuelles, observer un buveur de jus d'orange et observer un cancer du poumon ne sont pas des événements dépendants. Les chances d'observer un cancer du poumon chez un buveur de jus d'orange ne sont pas différentes de celles de l'observer chez un individu quelconque (en comprenant bien que ces deux individus sont comparables sous tout autre rapport). Observer face au premier lancer d'une pièce de monnaie bien équilibrée n'influence pas le résultat du second lancer. Un cas particulier: deux événements incompatibles sont nécessairement dépendants. Si l'un se réalise, l'autre n'a aucune chance de se réaliser.

PROBABILITÉ

Comme on l'a vu, il est dans la nature du phénomène aléatoire de produire des résultats dont la prévision reste incertaine. Par contre, si les résultats ne peuvent pas être prévus avec certitude, il est possible d'une certaine façon d'en mesurer la vraisemblance. Cette mesure de la vraisemblance d'un résultat ou d'un événement incertain est ce que l'on appelle la *probabilité* de cet événement. Dans la présente section, on s'intéresse à la définition de cette mesure et aux principales règles de calcul qui la régissent.

Définition fréquentiste

Il existe plusieurs interprétations ou définitions du concept de probabilité. Mais la seule considérée ici est l'interprétation *fréquentiste.* Pour la comprendre, considérons l'expérience aléatoire du lancer d'une pièce de monnaie. Chaque lancer de la pièce est une expérience (un essai ou une épreuve). A chaque essai correspondent ici deux résultats possibles (élémentaires) : pile ou face. Portons notre attention sur l'événement *E:* obtenir le côté pile de la pièce. Si l'on procède à 100 essais faits dans des conditions identiques, l'événement *E* parfois se réalise, parfois ne se réalise pas. S'il s'est réalisé 54 fois au cours de 100 essais aléatoires, le rapport $^{54}/_{100}$ en est la fréquence relative, c'est-à-dire le rapport du nombre de cas favorables à l'événement au nombre d'essais. Pour toute autre suite de 100 essais faite avec la même pièce de monnaie, la fréquence relative pourra s'écarter plus ou moins de la valeur précédente $^{54}/_{100}$. La fréquence relative d'un événement *E* peut varier d'une suite d'essais à une autre. Il est cependant un fait d'expérience, que les fluctuations de la fréquence relative tendent à être de plus en plus faibles à mesure qu'augmente le nombre *n* d'essais. Pour une pièce bien équilibrée, la fréquence relative ou proportion de piles tend à se stabiliser autour de la valeur 0,5 quand *n* augmente. Pour une autre pièce, celle-là non équilibrée, la fréquence relative de piles pourrait se stabiliser autour de 0,6 par exemple.

La valeur autour de laquelle vient se stabiliser la fréquence relative de piles est interprétée comme la probabilité de l'événement E considéré. Si cette valeur est 0,5, celle-ci est prise comme la probabilité fréquentiste d'obtenir pile au lancer de la pièce de monnaie en question. Cela veut dire que, sur 1000 lancers, par exemple, on doit s'attendre à obtenir en moyenne 500 piles. Sur un très grand nombre d'essais, la fréquence relative observée s'éloigne peu de la valeur 0,5. La *probabilité fréquentiste* est une proportion limite ou fréquence relative limite, en ce sens que son calcul repose sur la répétition d'un très grand nombre d'essais. La probabilité est donc une mesure des chances qu'a un événement de se produire, par exemple, d'obtenir pile au lancer d'une pièce de monnaie, d'être victime d'un accident pour un piéton, de naître avec une malformation, etc.

Certaines mesures de fréquence utilisées en épidémiologie sont des probabilités, par exemple, la prévalence relative, l'incidence cumulative, la létalité. La prévalence relative d'une maladie ou d'une caractéristique au sein d'une population mesure la probabilité qu'un individu, tiré au hasard de cette population, présente cette maladie ou la caractéristique. L'incidence cumulative d'une maladie mesure la probabilité qu'un individu, appartenant à une cohorte déterminée d'individus sains, soit atteint de cette maladie

dans une période spécifiée. La létalité d'une maladie mesure la probabilité pour un malade de décéder dans une période déterminée, mesurée depuis le moment du diagnostic.

La probabilité fréquentiste n'est jamais connue exactement étant donné qu'on ne peut exécuter ou observer une suite infinie d'essais. Même grand, le nombre d'essais est forcément limité. On obtient donc une valeur approximative de la probabilité, qui est d'autant meilleure que le nombre d'essais est élevé.

La probabilité qu'un nouveau-né soit de sexe masculin ne peut être connue à partir de l'observation d'une seule naissance ou même de quelques-unes seulement. En ne considérant qu'une seule naissance (un seul essai), l'estimation de cette probabilité est 0 ou 1, ce qui ne correspond d'aucune manière à la réalité. Peut-on davantage l'estimer à partir d'une famille de 5 enfants qui comprend 4 filles et 1 garçon? Si, par contre, on a observé au cours d'une année dans la population 86 161 naissances vivantes dont 44 568 étaient masculines, il est raisonnable d'estimer la probabilité d'avoir un garçon à la naissance comme:

$$\frac{44\,568}{86\,161} = 0{,}5173,$$

ce qui signifie qu'en moyenne, on peut s'attendre à trouver 517 garçons par 1000 naissances vivantes.

Puisque jamais le nombre de cas favorables à l'événement ne dépasse celui des essais, la probabilité est au plus égale à 1. Aussi, puisque jamais le nombre de cas favorables est négatif, la probabilité est obligatoirement égale ou supérieure à 0. Symboliquement, on a la propriété fondamentale (de toute probabilité, quelle que soit d'ailleurs son interprétation) :

$$0 \leqslant \text{Prob}\,(E) \leqslant 1$$

où Prob *(E)* désigne la probabilité de l'événement *E*. Une probabilité égale à 1 renvoie à l'événement certain, une probabilité nulle à l'événement nul ou impossible. Plus un événement a des chances de se produire, plus sa probabilité est proche de 1.

La valeur d'une probabilité n'est pas éternelle. Elle est liée non seulement à l'événement futur qui nous intéresse mais aussi à notre connaissance actuelle du phénomène en cause. La probabilité de guérison d'une certaine maladie n'est pas définitive; sa valeur pourrait, par exemple, changer par suite de l'acquisition de nouvelles connaissances. Par contre, les règles de calcul des probabilités restent immuables. La probabilité qu'un dé présente une face paire en tombant est toujours égale à la somme des trois probabilités: d'obtenir la face 2, la face 4, la face 6, quelles que soient les valeurs de chacune de ces trois probabilités.

Calcul de probabilités

Nous allons maintenant exposer les règles de base du calcul des probabilités, celles en particulier de multiplication et d'addition. Nous présentons aussi la formule (ou règle) de Bayes, qui est dite règle de la probabilité des causes.

RÈGLE DÉ MULTIPLICATION

La règle de multiplication se rapporte au calcul de Prob (*A* et *B*), c'est-à-dire la proba-

bilité que les deux événements, *A* et *B,* se réalisent.

Il est faux de penser, de façon générale que

Prob *(A* et *B)* = Prob *(A)* Prob *(B).*

On peut facilement s'en rendre compte à partir d'un exemple. Considérons les deux événements

— *A* : avoir l'œil droit bleu

— *B :* avoir l'œil gauche bleu.

Si l'on accepte que la prévalence relative des yeux bleus dans la population sanpulienne est 0,30, on peut écrire Prob *(A* et *B) = 0,30.* Tous conviendrons que Prob *(A)* = Prob *(B) — 0,30.* Il devient alors évident que:

$$\text{Prob}\,(A \text{ et } B) \neq \text{Prob}\,(A)\ \text{Prob}\,(i$$
$$0{,}30 \neq 0{,}30 \times 0{,}30$$

En fait, la règle de multiplication est plus complexe. Considérons une suite aléatoire de *n* essais pour lesquels *A* et *B* sont des événements quelconques possibles. Les symboles n_A et n_B représentent respectivement le nombre d'essais où A s'est réalisé et le nombre d'essais où *B* s'est réalisé au cours des *n* essais. Le symbole n_A *et B* représente le nombre d'essais où les deux événements *A* et *B* se sont réalisés. On peut algébriquement écrire :

$$\frac{n_{A \text{ et } B}}{n} = \frac{n_{A \text{ et } B}}{n_B}$$

pourvu que n_B soit différent de zéro, c'est-à-dire à la condition que l'événement B soit possible. Si *n* devient de plus en plus grand, l'on voit se dessiner (fort de l'interprétation fréquentiste) la *règle de multiplication :*

Prob *(A et B)* = **Prob** *(A/B)* **Prob** ***(B).***

On aurait pu tout aussi bien obtenir la forme :

Prob ***(A*** **et** ***B)*** = **Prob** *(B/A)* **Prob (A).**

Cette fois la règle marche si on l'applique à l'exemple des yeux bleus. En effet,

Prob *(A* et *B) = 0,30*

Prob *(B/A)* Prob *(A)* = 1 *x 0,30 = 0,30.*

On comprend que, dans l'exemple, Prob *(B/A)* soit pratiquement 1 puisque les événements A et *B* sont fortement dépendants dans le sens que si A (avoir l'œil droit bleu) s'est produit, il est fort probable que *B* (avoir l'œil gauche bleu) se produise aussi.

La règle de multiplication se simplifie à :

Prob (*A* et *B)* = Prob *(A)* Prob *(B)*

lorsque les deux événements sont indépendants. D'ailleurs, la relation formelle qui caractérise l'indépendance entre deux événements *A* et *B* est donnée par:

Prob *(A/B)* = Prob *(A)*

ce qui traduit le fait que la réalisation de *B* n'influence pas la réalisation (probabilité) de *A.* En situation d'indépendance, on a aussi:

Prob *(B/A)* = Prob *(B)*

et de façon équivalente, la relation

Prob *(A* et *B)* = Prob *(A)* Prob *(B)*

RÈGLE D'ADDITION

La règle d'addition a trait au calcul de Prob *(A ou B),* c'est-à-dire de la probabilité que l'événement *(A ou B)* se réalise, en d'autres termes qu'au moins un des deux événements *A* et *B* se réalise.

Il est faux de croire qu'en règle générale

Prob *(A* ou *B)* = Prob *(A)* + Prob *(B).*

Pour s'en convaincre, il suffit de se rendre compte que Prob *(A* ou *B)* dépasserait la valeur 1 si, par exemple, Prob *(A)* et Prob *(B)* devaient valoir respectivement 0,8 et 0,3. En fait, la règle est plus complexe.

Nous n'allons pas la démontrer formelle ment, mais seulement la faire apparaître en utilisant les fréquences relatives. Considérons à nouveau une suite aléatoire de n essais pour lesquels A et B sont des événements quelconques possibles. Alors, au cours d'un essai, A peut se réaliser et il en est de même pour B. En additionnant le nombre n_A d'essais où A s'est réalisé, au nombre n_B d'essais où B s'est réalisé, l'on compte deux fois le nombre n_{A} et B d'essais où les deux événements se sont produits. Par conséquent, pour connaître n_A ou B, c'est-à-dire le nombre d'essais où au moins un des deux événements s'est réalisé, il faut soustraire une fois la valeur n_A et B.

D'où

$$n_{A \text{ ou } B} = n_A + n_B - n_{A \text{ et } B}$$

En divisant par n, on obtient,

$$\frac{n_{A \text{ ou } B}}{n} = \frac{n_A}{n} + \frac{n_B}{n} - \frac{n_{A \text{ et } B}}{n}.$$

Si n devient de plus en plus grand, l'on voit apparaître la *règle d'addition :*

Prob (A ou *B)* = Prob (A) + Prob *(B)* — Prob *(A* et *B).*

Cette règle se réduit à :

Prob *(A* ou *B)* = Prob *(A)* + **Prob *(B)***

lorsque les deux événements sont incompatibles, puisqu'en effet Prob *(A* et *B) = 0.* La probabilité que deux événements incompatibles apparaissent en même temps est nulle.

Considérons dans la population sanpulienne les deux caractéristiques génétiques A et *B* (disons « le groupe sanguin A » et « les yeux bleus »). Supposons que ces caractéristiques sont indépendantes (hypothèse raisonnable n'est-ce pas?) et que leurs prévalences relatives sont respectivement de 0,10 et 0,30. On cherche la prévalence relative des individus qui ont au moins l'une des deux caractéristiques. En d'autres termes, on veut connaître Prob *(A* ou *B).*

On a :

$$\begin{aligned}\text{Prob}(A \text{ ou } B) &= \text{Prob}(A) + \text{Prob}(B) \\ &\quad - \text{Prob}(A \text{ et } B) \\ &= \text{Prob}(A) + \text{Prob}(B) \\ &\quad - \text{Prob}(A)\,\text{Prob}(B) \\ &\quad (\text{indépendance de } A \text{ et de } B) \\ &= 0{,}10 + 0{,}30 - 0{,}10 \times 0{,}30 \\ &= 0{,}37.\end{aligned}$$

La règle d'addition est généralisable à plus de deux événements. Pour trois :

$$\begin{aligned}\text{Prob}(A \text{ ou } B \text{ ou } C) &= \text{Prob}(A) + \text{Prob}(B) \\ &\quad + \text{Prob}(C) - \text{Prob}(A \text{ et } B) \\ &\quad - \text{Prob}(A \text{ et } C) \\ &\quad - \text{Prob}(B \text{ et } C) \\ &\quad + \text{Prob}(A \text{ et } B \text{ et } C).\end{aligned}$$

Si les événements sont incompatibles deux à deux, on a :

Prob *(A* ou *B* ou C) = Prob *(A)* + Prob *(B)* + Prob *(C)*

RÈGLE DE COMPLÉMENTARITÉ

De la règle d'addition pour les événements incompatibles, on a :

$$\text{Prob } (A \text{ ou } \bar{A}) = \text{Prob } (A) + \text{Prob } (\bar{A})$$

Par ailleurs, du fait que (A ou Ā) est un événement certain, on a aussi :

$$\text{Prob } (A \text{ ou } \bar{A}) = 1$$

Ainsi, $\text{Prob } (A) + \text{Prob } (\bar{A}) = 1$

d'où la *règle de complémentarité:*

$$\textbf{Prob } (\bar{A}) = 1 - \textbf{Prob } (A)$$

FORMULE (THÉORÈME OU RÈGLE) DE BAYES

Considérons l'événement A « être hospitalisé ». Supposons que l'hospitalisation peut être entraînée de manière exclusive et exhaustive soit par la présence de la maladie B_1, soit par B_2 ..., soit encore par la maladie B_k. Les causes B_i d'hospitalisation forment une partition. Ayant observé l'hospitalisation A chez un individu, un investigateur veut connaître la probabilité que la maladie présente chez cet individu soit B_i. Il veut donc calculer la probabilité conditionnelle :

$$\text{Prob } (B_i/A).$$

La quantité Prob (B_i/A) est dite *probabilité ā posteriori,* connue aussi sous l'appellation de la *probabilité des causes.* Ayant été hospitalisé, quelle est la probabilité que la cause en soit la maladie B_i? La quantité Prob (B_i) est dite *probabilité à priori,* c'est-à-dire la probabilité calculée indépendamment de la réalisation ou non de l'événement A (de l'hospitalisation).

Si on connaît la probabilité Prob (B_i) qu'un individu soit affecté par la maladie B_i, de quelle façon l'information supplémentaire qu'il est hospitalisé » modifie-t-elle l'estimation de cette probabilité? C'est à cette question que la formule de Bayes permet de répondre. Si on connaît Prob (B_i) et Prob (A/B_i), alors il est possible de déterminer la valeur de Prob (B_i/A). En termes concrets, si on connaît la probabilité générale qu'un individu soit affecté par la maladie B_i et celle qu'un individu soit hospitalisé lorsqu'il est affecté par cette maladie, alors on peut déduire dans sa condition d'hospitalisé la probabilité qu'il soit affecté par B_i. Cette relation entre, d'une part Prob (B_i/A), et, d'autre part Prob (B_i) et Prob (A/B_i), est donnée par la formule de Bayes:

$$\textbf{Prob } (B_i/A) = \frac{\textbf{Prob } (A/B_i)\ \textbf{Prob } (B_i)}{\Sigma\ \textbf{Prob } (A/B_i)\ \textbf{Prob } (B_i)}$$

Voici la démonstration de cette formule. De la règle de multiplication, on a l'une ou l'autre des formes :

$$\text{Prob } (A \text{ et } B_i) = \text{Prob } (A/B_i)\ \text{Prob } (B_i) = \text{Prob } (B_i/A)\ \text{Prob } (A)$$

D'où

$$\text{Prob } (B_i/A) = \frac{\text{Prob } (A/B_i)\ \text{Prob } (B_i)}{\text{Prob } (A)} \qquad [1]$$

Le dénominateur Prob (A) peut s'écrire autrement. En effet, l'événement A se réalise en présence soit de B_1, soit de B_2 ..., soit de B_k.

Ainsi

$A = (A \text{ et } B_1)$ ou $(A \text{ et } B_2)$ ou ... ou $(A \text{ et } B_k)$.

D'où, suivant la règle d'addition pour les événements incompatibles:

Prob (A) = Prob *[(A et B_1) ou (A et B_2)*
ou ... ou *(A et B_k)]*
= Prob (*A* et B_1) + Prob (A *et B_2)* +
... + Prob (A et B_k).

et selon *la* règle de multiplication,

= Prob *(A/B,)* Prob (B_1) +
Prob *(A/B_1)* Prob *(B_2)* + ... +
Prob *(A/B_k)* Prob (B_k).
= Σ Prob (*A*/B_i) Prob *(Bi)* [2]

En substituant l'expression [2] dans l'équation [1], on obtient la formule de Bayes.

$$\text{Prob}\,(B_i/A) = \frac{\text{Prob}\,(A/B_i)\,\text{Prob}\,(B_i)}{\Sigma\,\text{Prob}\,(A/B_i)\,\text{Prob}\,(B_i)}$$

Pour illustrer la formule de Bayes, considérons l'accident de la route chez les conducteurs de véhicules automobiles. Supposons que ceux-ci soient répartis en trois groupes d'âge: les jeunes (moins de 25 ans), les conducteurs d'âge moyen (25-64 ans) et les conducteurs âgés (65 ans et plus). Leur distribution dans la population se lit comme suit :

Âge (années)	Proportion
<25	0,25
25-64	0,60
≥65	0,15

Cette distribution nous permet d'estimer les probabilités à priori Prob *(B_i) :*

Prob *(B_1)* = Prob (jeune) = 0,25;

Prob *(B_2)* = Prob (moyen) = 0,60;

Prob *(B_3)* = Prob (âgé) = 0,15.

Notons au passage que les B_i sont des événements incompatibles deux à deux, ce qui est une exigence essentielle à l'utilisation de la formule de Bayes. Supposons que les probabilités d'avoir un accident de la route chez les jeunes, les conducteurs d'âge moyen et les conducteurs âgés sont respectivement 0,10, 0,01 et 0,03. On peut écrire:

Prob *(A/B_1)* = Prob (accident/jeune) = 0,10

Prob *(A/B_2)* = Prob (accident/moyen) = 0,01

Prob *(A/B_3)* = Prob (accident/âgé) = 0,03

On peut se demander maintenant quelle est la probabilité, face à un accident ayant eu lieu, que le conducteur en cause soit un jeune. Quelle est la valeur de Prob *(B_1/A)*?

$$\text{Prob (jeune/acc.)} = \frac{0{,}25 \times 0{,}10}{0{,}25 \times 0{,}10 + 0{,}60 \times 0{,}01 + 0{,}15 \times 0{,}03} = 0{,}70$$

RISQUE COMME PROBABILITÉ

Dans le langage courant, la notion de risque réfère, pour l'individu, à l'éventualité d'un événement fâcheux qui, en pratique, coïncide avec la perte d'un bien. Le risque d'incendie pour le propriétaire d'une maison correspond effectivement à l'éventualité que sa maison flambe. La réalisation de cet événement est influencée par un certain nombre de conditions : maison en matière inflammable, circuit électrique défectueux, système de chauffage défectueux, absence de détecteur de fumée, imprudence du propriétaire ou des occupants, etc. En épidémiologie, la notion de risque réfère également à l'éventualité d'un événement fâcheux, en l'occurrence la maladie (blessures) ou le décès. En définitive, le risque est une probabilité.

RÉSUMÉ

Une expérience aléatoire est caractérisée par le fait que tout résultat de l'expérience ne peut être prévu avec certitude. La probabilité est une mesure de la vraisemblance de résultat dont la prévision est incertaine. La probabilité a plusieurs interprétations. L'interprétation fréquentiste correspond au concept de proportion limite ou de fréquence relative limite. Au sens fréquentiste, la probabilité qu'un nouveau-né soit de sexe masculin, 0,51 par exemple, veut dire qu'en moyenne sur 100 naissances, 51 sont masculines. Certaines des mesures de fréquence utilisées en épidémiologie sont des mesures de probabilité. C'est le cas notamment de la prévalence relative, de l'incidence cumulative, de la létalité. Les probabilités, quelles que soient leur interprétation, obéissent à des règles immuables. Les plus connues sont les règles de multiplication et d'addition. La première concerne le calcul de Prob (*A* et *B*), la deuxième celui de Prob (*A* ou *B*). Une autre règle intéressante est celle de Bayes, qui permet de calculer Prob(B_i/*A*) pour un i donné quand on connaît Prob (B_i) et Prob (A/B_i) pour tout *i*.

Symboles

Ω : ensemble fondamental

A ou *B:* événement disjonction

A et *B:* événement conjonction

$\bar{A}$: événement non *A*

A/B: événement *A* conditionnel à *B*

Prob *(E) :* probabilité de l'événement *E*

Formules

- *Règle de multiplication*

  Prob (*A* et *B*) = Prob *(A/B)* Prob (*B*)

  = Prob (*B/A*) Prob (*A*)

  Cas particulier des événements indépendants

  Prob (*A* et *B)* = Prob (*A*)-Prob *(B)*

- *Règle d'addition*

  Prob (*A* ou *B*) = Prob (*A*) + Prob (*B*)

  – Prob (*A* et *B*)

  Cas particulier des événements incompatibles

  Prob (*A* ou *B*) = Prob (*A*) + Prob (*B*)

- *Règle de complémentarité*

  Prob ($\bar{A}$) = 1 – Prob (*A*)

- *Formule de Bayes*

$$\text{Prob}\,(B_i/A) = \frac{\text{Prob}\,(A/B_i)\ \text{Prob}\,(B_i)}{\Sigma\ \text{Prob}\,(A/B_i)\ \text{Prob}\,(B_i)}$$

LECTURES SUGGÉRÉES

1. LAZAR, P. et SCHWARTZ, D. ***Éléments de probabilité et statistique,*** Paris, Flammarion, 1967, chapitres II et III, pp. 15-32.

2. MARTEL, J.-M. et NADEAU, R. ***Probabilités en gestion et en économie,*** Chicoutimi, Gaétan Morin éditeur, 1980, chapitre 2, pp. 11-60.

CHAPITRE 11

Mesure de la probabilité de survie

Ce chapitre traite du calcul de la probabilité de survie à un événement. La durée de survie à une maladie ou à un traitement varie d'un individu à un autre. On définit d'abord le concept de fonction de survie, puis la question de l'estimation de cette fonction en présence de données de survie censurées est commentée. On y présente deux méthodes d'estimation ou de calcul de probabilités de survie; nommément, il s'agit de la méthode de Kaplan-Meier et de la méthode actuarielle.

Étant donné un patient atteint d'une certaine maladie, on peut s'interroger sur la probabilité qu'il survive à une certaine période. Comment mesurer une probabilité de survie? Nous nous proposons dans ce chapitre, après avoir expliqué les concepts de durée et fonction de survie, de présenter les méthodes les plus courantes d'estimation d'une probabilité de survie.

DURÉE DE SURVIE

Un individu est victime d'un premier infarctus le 27 septembre d'une certaine année. Conséquemment à cet infarctus, il décède le 15 janvier de l'année suivante. Le malade a survécu 110 jours à l'événement premier infarctus ». On dira que sa durée de survie, ou simplement sa survie, a été de 110 jours. On a pratiqué le 12 octobre d'une certaine année une greffe cardiaque sur un patient. Il décède le 28 décembre de la même année. Sa survie depuis cette transplantation a été de 77 jours. La *durée de survie à un* événement pour un individu est en définitive le temps écoulé depuis la date d'apparition de cet événement jusqu'à la date de décès de l'individu.

Ce concept de survie n'est pas seulement applicable aux phénomènes qui impliquent le décès clinique d'individus. Ce concept recouvre des situations variées. On peut s'intéresser à la durée de tolérance à une greffe, de soulagement par suite de l'injection d'un analgésique, de rémission avant la première récidive après l'administration d'un traitement, ou encore à la durée d'abstinence chez un alcoolique soumis à une cure de désintoxication. Ces situations réfèrent toutes à l'observation chez un individu de la persistance d'un état depuis son apparition jusqu'à sa défaillance. Dans ce qui suit, comme le font tous les auteurs, nous utiliserons systématiquement les termes vivants et décédés pour traduire respectivement les situations qui correspondent à la persistance et à la défaillance d'un état.

Dans un groupe d'individus, on conçoit facilement que la durée de survie à un événement varie d'un individu à l'autre. La survie est une variable, en l'occurrence quantitative, continue. Suivant la nature du phénomène, la durée de survie est mesurée en années, mois, semaines, jours ... Il est important de souligner que l'*événement-origine,* c'est-à-dire à partir duquel est mesurée la durée de survie, doit être clairement défini. Dans le cas d'un cancer, la survie est-elle mesurée à partir de la première admission à l'hôpital, ou du début de l'administration d'un traitement?

FONCTION DE SURVIE

Chez 20 patients atteints d'un cancer du pancréas, un clinicien a noté leur survie en jours depuis leur première admission à l'hôpital. Tous sont décédés, un après 133 jours, un autre après 55 jours et ainsi de suite comme l'indiquent les résultats suivants:
133 jours, 55, 247, 22, 119, 88, 171, 103, 239, 212, 331, 35, 284, 61, 285, 186, 393, *61,* 47, 171

Pour mieux apprécier le profil de la survie de ces 20 patients, le clinicien peut choisir de présenter (dans un tableau ou un graphique) les proportions de patients encore vivants après certains délais ou durées. La présentation la plus immédiate est celle où les délais correspondent aux durées de survie observées. On trouve une telle représentation au

tableau 11-1 et à la figure 11-1. La distribution au tableau 11-1 est celle de fréquences cumulées décroissantes.

On peut facilement imaginer qu'avec un plus grand nombre de patients, le diagramme en gradins de la figure 11-1 se rapprocherait d'une courbe plus lisse comme à la figure 11-2. Cette courbe (de survie) caractérise graphiquement ce qu'on peut appeler une *fonction de survie,* dénotée S(t):

$$t \longrightarrow S(t).$$

Tableau 11-1

Durée de survie *t* (en jours)	Nombre de patients ayant une durée de survie supérieure à *t*	Proportion (en %) de patients ayant une durée de survie à *t*
0	20	100
22	19	95
35	18	90
47	17	85
55	16	80
61	14	70
88	13	65
103	12	60
119	11	55
133	10	50
171	8	40
186	7	35
212	6	30
239	5	25
247	4	20
284	3	15
285	2	10
331	1	5
393	0	0

À chaque durée de survie t observée correspond une proportion S*(t)* de survivants. On définit habituellement S*(t)* comme la probabilité que la durée de survie T soit supérieure à une valeur donnée t, ce qui peut s'écrire symboliquement:

$$S(t) = \text{Prob}\,(T > t).$$

Dans l'exemple, au tableau 11-1 (ou à la figure 11-1), S(35) = Prob (T > 35) = 0,90.

Figure 11-1

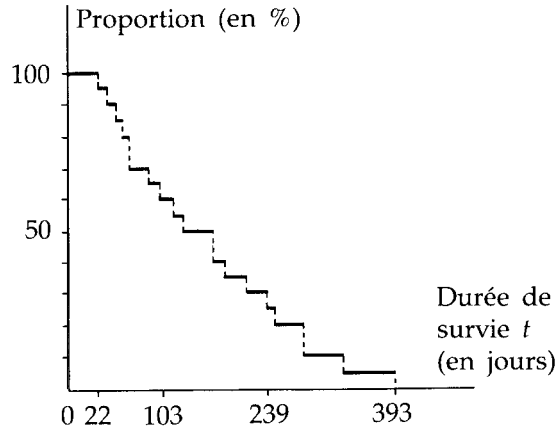
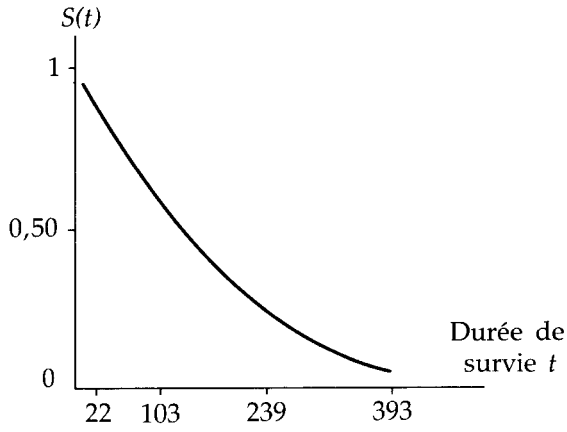


Figure 11-2



Une fonction de survie S(*t*) se caractérise par les trois propriétés suivantes:

1) *S(0)* = 1	Tous les individus sont vivants au moment de leur admission dans une étude de survie.
2) S(∞)= 0	Après un temps d'observation suffisamment long, tous les individus sont décédés.
S(*t*) est une fonction décroissante.	Du groupe initial étudié, le nombre d'individus encore vivants tend à décroître à mesure que la durée d'observation s'allonge.

Soulignons que la fonction complémentaire de S(t), dénotée le plus souvent par *F(t)* et appelée fonction de répartition, représente la probabilité que la durée de survie T soit au plus égale à une valeur donnée *t.* (Pensons à la létalité d'une maladie). On a:

$$F(t) = 1 - S(t) = \text{Prob}\ (T\ t)$$

et dans l'exemple :

$$F(35) = 1 - S(35) = \text{Prob}\ (T \leq 35) = 0{,}10.$$

Estimation en contexte laboratoire ou clinique

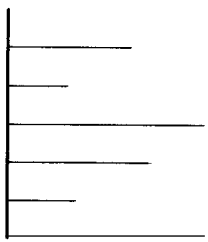
Pour estimer une fonction de survie, il est nécessaire d'observer un certain nombre d'individus à partir d'un événement-origine défini de façon identique pour chacun d'entre eux.

Si les individus sont tous observés à partir d'une même date du calendrier, le contexte dans lequel est menée l'étude de survie peut être qualifié de *contexte laboratoire.* La date d'entrée dans l'étude est fixée par l'investigateur.

Ce contexte correspond à l'idée d'une entrée fixe pour tous les individus; on le trouve fréquemment en expérimentation animale. Par exemple, pour étudier l'effet carcinogène d'une certaine substance, toutes les souris d'un groupe reçoivent en même temps cette substance par inoculation. La figure 11-3 décrit un contexte laboratoire.

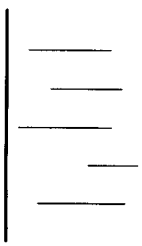
Dans le *contexte clinique,* la date d'admission dans une étude n'est pas en général la même pour tous les individus. Dans ce cas, la date d'entrée dans l'étude est variable. Si, par exemple, un clinicien décide d'observer un groupe de patients à partir du début de leur traitement, on conçoit aisément que la date du début de traitement, et donc la date d'admission du patient dans l'étude, varie d'un patient à l'autre. Le début du traitement d'un patient pourra être le 12 juillet, alors qu'il sera le 26 septembre pour un autre. La figure 11-4 décrit un contexte clinique.

Figure 11-3 **Contexte laboratoire**



Entrée fixe

Figure 11-4 **Contexte clinique**



Entrée variable

Pour disposer d'un certain nombre d'individus, l'investigateur en contexte clinique choisit une période du calendrier au cours de laquelle sont admis les individus qui feront partie de son étude. C'est la *période* dite *d'entrée dans l'étude.* Elle est limitée par deux dates du calendrier, la première étant la *date de début d'entrée dans l'étude* et la deuxième, la *date de fin d'entrée dans l'étude.* L'investigateur convient parfois d'une troisième date qui marque la fin des observations pour l'ensemble des individus. C'est la *date de fin d'étude.* Cette date, dont la détermination relève d'une décision de l'investigateur, est toujours postérieure à la date de fin d'entrée dans l'étude. Théoriquement, à la limite, elle peut lui être égale. La *période d'étude, ou période d'observation,* est celle qui s'étend de la date du début d'entrée dans l'étude à la date de fin d'étude. La figure 11-5 illustre les différentes dates dont il a été question ici.

Nous allons dans la suite supposer que les caractéristiques des individus, admis à des dates différentes, ne changent pas durant la période d'étude. Cette supposition permet de considérer comme équivalents les schémas des figures 11-5 et 11-6. En d'autres termes, cette hypothèse d'homogénéité permet de ramener le contexte clinique au contexte laboratoire.

Figure 11-5

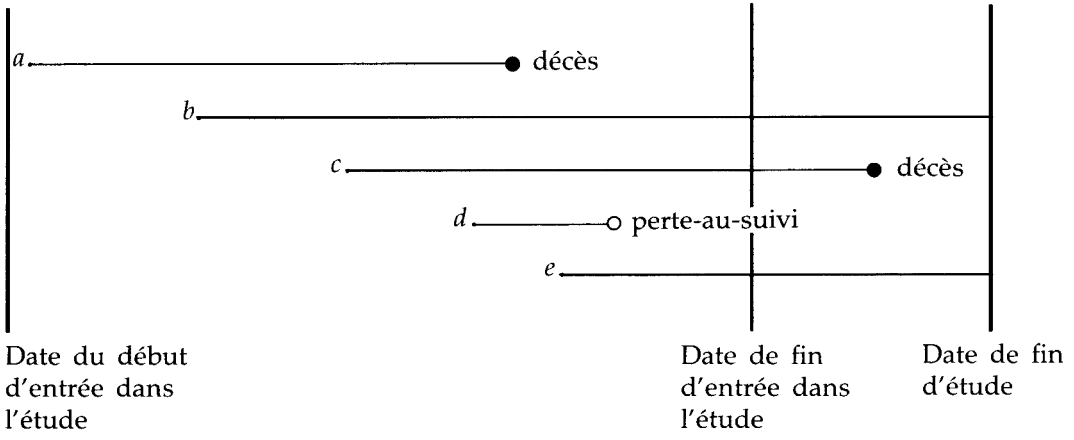
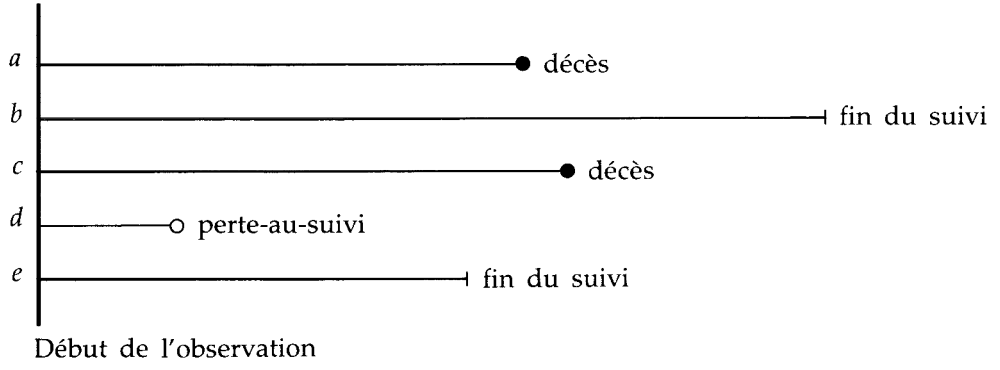


Figure 11-6



Estimation avec des données censurées

L'estimation d'une fonction de survie est une opération simple ou moins immédiate selon que les données de survie sont complètes ou incomplètes.

Si tous les patients ou individus sont suivis, sans exception, jusqu'au décès du dernier, les données de survie forment une *série complète.* Cela suppose que l'investigateur n'a perdu de vue aucun de ses patients en cours d'observation et qu'il les a suivis sur une période suffisamment longue pour observer le décès chez tous. Les données se rapportant à l'exemple des 20 patients atteints d'un cancer du pancréas, forment une série complète. Dans un tel cas, l'estimation de la fonction de survie se fait aisément. Au tableau 11-1, on a par exemple :

$$S(239) = {}^{5}/_{20} = 0{,}25$$

L'estimation s'obtient directement de point en point. De façon générale, si les durées de survie de n individus, disposées en ordre croissant, sont:

$$t_1, t_2, \ldots, t_i, \ldots, t_n,$$

alors

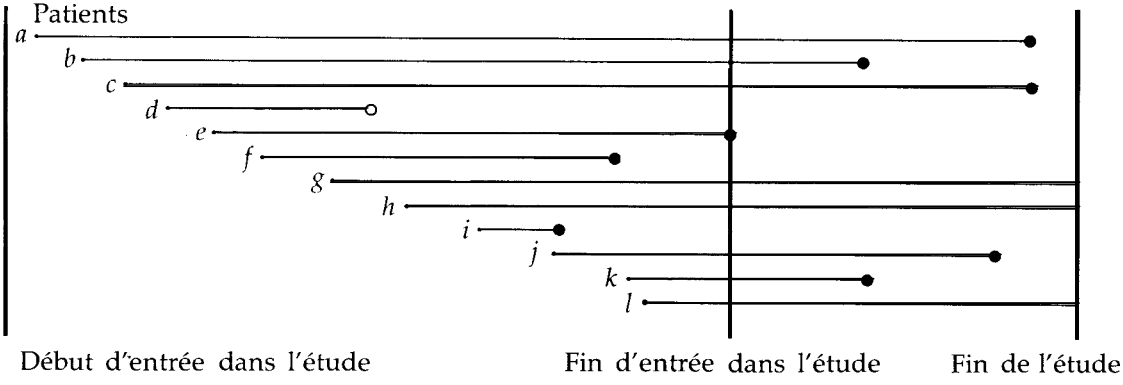
$$S(t_i) = \frac{\text{nombre de survivants au temps } t_i}{\text{nombre d'individus } (n)}$$

Par contre, si l'investigateur perd de vue des patients en cours d'observation ou fixe une date de fin d'observation qui fait en sorte qu'à cette date tous ne sont pas encore décédés, alors les données sont dites incomplètes. C'est le cas, à la figure 11-5, pour les données de survie des patients *b, d* et *e.* Dans une *série incomplète,* la date du décès n'est donc pas connue pour un certain nombre de patients.

La date de décès pour un patient peut être inconnue soit parce qu'il a été perdu de vue en cours d'observation *(perdu-au-suivi),* soit parce qu'il n'était pas encore décédé à la date de fin d'étude *(exclu-vivant).* Un exclu-vivant est vivant à la date de fin d'étude, mais devient exclu de l'étude à partir de cette date. C'est le cas des patients *b et e* à la figure 11-5. Tant pour les perdus-au-suivi que pour les exclus-vivants, les données de survie sont d'une certaine manière *censurées.*

La présence de données censurées rend plus difficile l'estimation d'une fonction de survie. Nous allons illustrer cette difficulté par un exemple. Considérons 12 patients

Figure 11-7



atteints d'un même cancer dont la survie depuis le diagnostic est décrite à la figure 11-7.

Les données relatives aux patients *d, g, h* et *l* sont censurées. Le patient *d* est un perdu-au-suivi, alors que les patients *g, h* et *l* sont des exclus-vivants. Quant aux autres, ils sont décédés en cours d'étude. Selon l'ordre d'entrée, les données de la figure 11-7 forment la série incomplète suivante :

525 jours, 413, 477, 105 + , 273, 189, 399 + , 367 + , 42, 232, 126, 232 + ,

où le signe + indique une donnée censurée. Si l'on dispose les données en ordre croissant, la série devient :

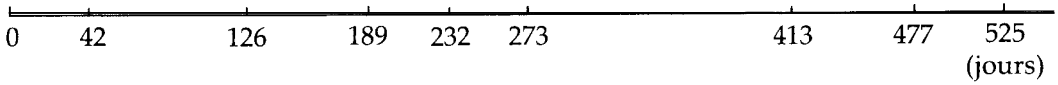
42 jours, 105 + , 126, 189, 232, 232 + , 273, 367 + , 399 + , 413, 477, 525.

Supposons que l'on veuille calculer la survie au 189^{e} jour, c'est-à-dire $S(189)$. Il y a une difficulté. On ne peut dire que $S(189)$ égale 8 survivants sur 12 puisque nous ignorons si le patient *d* à la figure 11-7 est encore vivant au 189^{e} jour. Il a été perdu après 105 jours d'observation et, de ce fait, exclu de l'étude à partir de ce moment. Le dénominateur de $S(189)$ ne serait donc plus 12, mais 11.

Nous allons aborder les procédures d'estimation d'une fonction de survie sur une série incomplète, c'est-à-dire lorsqu'il y a des données censurées. Ces procédures relèvent de la *méthode* dite des *tables de survie.* Notre souci étant d'expliquer ces procédures, nous allons les appliquer à un petit groupe de patients, conscients que les probabilités qui en résultent seront assez approximatives.

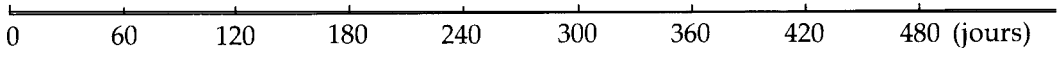
TABLES DE SURVIE

Deux approches peuvent être envisagées dans l'estimation d'une fonction de survie par la méthode des tables de survie. Pour l'une, l'investigateur décide d'estimer la fonction de survie en des points ou moments qui correspondent aux temps de décès tels qu'ils furent observés. C'est l'approche que l'on peut désigner à *intervalles variables.* Dans l'exemple des 12 patients cité plus haut, l'investigateur calculera suivant cette approche $S(42)$, $S(126)$, $S(189)$, etc.



Dans le cas de l'autre approche, l'investigateur décide d'estimer la fonction de survie en des points particuliers qu'il a choisis à l'avance. Il partage la période d'étude ou d'observation en intervalles de temps généralement de même

On peut qualifier cette approche comme étant celle *à intervalles fixes.* Se référant au même exemple, l'investigateur peut convenir d'estimer $S(60)$, $S(120)$, $S(180)$, etc.



La table de survie actuarielle est la principale méthode utilisée lorsque les intervalles sont fixés par l'investigateur.

Pour les intervalles variables, la méthode employée est celle dite de Kaplan-Meier (ou du produit limite).

Méthode de Kaplan-Meier
(ou du produit limite)

Rappelons les données de survie des 12 patients, disposées en ordre croissant:
42 jours, 105 + , 126, 189, 232, 232 + , 273,
367 + , 399 + , 413, 477, 525.

Lorsqu'un retrait (par exclusion ou par perte) et un décès se produisent au même moment, on considère que le décès est survenu avant le retrait. Le décès au 232e jour a préséance sur le retrait au même jour. C'est une règle à suivre.

Dans ce même groupe de 12 patients, on note que le premier décès survient au 42e jour. On peut dès maintenant estimer la fonction de survie à 42 jours:

$$S(42) = \frac{11}{12} \qquad (= 0{,}9167).$$

Par suite de ce décès, le groupe est réduit à 11 patients. Après le 105e jour, le retrait 105 + réduit le groupe à 10 patients. Le décès suivant survient au 126e jour. Pour qu'un patient ait une survie supérieure à 126 jours, il doit survivre aux 42 premiers jours et de la fin du 42e jusqu'à la fin du 126e jour. Ainsi :

$$S(126) = S(42) \times \frac{9}{10} \qquad \left(= \frac{11}{12} \times \frac{9}{10} = 0{,}8250\right),$$

en vertu de la règle de multiplication.

Le même raisonnement conduit aux estimations suivantes :

$$S(189) = S(126) \times \frac{8}{9} = 0{,}7333$$

$$S(232) = S(189) \times \frac{7}{8} = 0{,}6417$$

$$S(273) = S(232) \times \frac{5}{6} = 0{,}5347$$

$$S(413) = S(273) \times \frac{2}{3} = 0{,}3565$$

$$S(477) = S(413) \times \frac{1}{2} = 0{,}1782$$

$$S(525) = S(477) \times \frac{0}{1} = 0{,}0000$$

Sur un plan formel, si les temps de décès observés sont, en ordre croissant,

$$t_1, t_2, t_3, \ldots, t_i, \ldots, t_k,$$

alors l'estimation de la fonction de survie $S(t)$ en chacun de ces points est obtenue à partir de l'expression suivante :

$$S(t_i) = \left(1 - \frac{D_0}{O_0}\right)\left(1 - \frac{D_1}{O_1}\right)\left(1 - \frac{D_2}{O_2}\right)\ldots\left(1 - \frac{D_{i-1}}{O_{i-1}}\right),$$

D_i et O_i dénotant respectivement le nombre de décès et le nombre de personnes en observation dans l'intervalle $(t_i, t_i + i)$, pour $i = 1, 2, 3$ .... La limite t_0 correspond au temps 0, c'est-à-dire au début de l'étude, avec bien sûr $S(t_0) = 1$.

Ces estimations ponctuelles décrivent l'expérience (quantitative) se rapportant à la survie d'un groupe de patients. Les calculs conduisant aux estimations de la fonction de survie $S(t)$ en des points particuliers sont habituellement présentés dans un tableau. Celui-ci comprend généralement huit colonnes, la dernière étant justement celle où figurent les estimations de la fonction de sur-

vie. Ce tableau est communément appelé table de survie. Le tableau 11-2 est un exemple de la table de survie pour les douze patients atteints du cancer.

Nous allons maintenant expliquer chacune des colonnes de la table de survie pour la méthode de Kaplan-Meier.

Intervalles *(Int)* : limite inférieure ou début des intervalles.

Vivants *(V)* : nombre d'individus encore vivants au début de chaque intervalle. Au début du premier intervalle figure le nombre total de patients admis dans l'étude.

Décédés *(D)* : nombre de décès survenus dans chaque intervalle.

Exclus-vivants *(EV) :* nombre de patients exclus-vivants au début de chaque intervalle. Cette colonne intègre ici les perdus-au-suivi *(PV)* aux exclus-vivants.

Individus en observation *(O)*: nombre d'individus en observation pendant l'intervalle complet et ce pour chaque intervalle. Ce nombre est obtenu en faisant la différence *V - EV*.

Probabilité de décès dans l'intervalle *(q)*: proportion de décès dans chaque intervalle. C'est une estimation de la probabilité de décès $q = D/O$.

Probabilité de survie sur l'intervalle *(p)*: parmi les patients vivants au début de l'intervalle, proportion de ceux encore vivants à la fin de l'intervalle. C'est une estimation de la probabilité de survie à l'intervalle. On a pour chaque intervalle $p = 1 - q$.

Estimation de la fonction de survie *(S)* : estimation de la proportion de patients encore vivants au début de chaque intervalle. Si chaque intervalle est numéroté dans l'ordre 0, 1, 2, ..., *i, ... n,* et si ti indique le début de l'intervalle i, alors *S(ti)* est le résultat du produit:

$$P_0 \, x \, p_1 \text{ x } ... \text{ x } pi - 1$$

En effet, pour survivre jusqu'au début de l'intervalle i, un patient doit survivre aux intervalles 0, 1, 2, ..., *i* - 1. D'où le produit des probabilités d'après la règle de multiplication. $S(t_0)$ mesure la probabilité d'être vivant au début de l'étude; elle est égale à 1. Se rapportant à l'exemple des 12 patients, on a :

Tableau 11-2: Table de survie

Int (jours)	*V*	*D*	*EV*	*O*	*q*	*p*	*S*
0	12	1	0	12	0,0833	0,9167	1
42	11	1	1	10	0,1000	0,9000	0,9167
126	9	1	0	9	0,1111	0,8889	0,8250
189	8	1	0	8	0,1250	0,8750	0,7333
232	7	1	1	6	0,1667	0,8333	0,6417
273	5	1	2	3	0,3333	0,6667	0,5347
413	2	1	0	2	0,5000	0,5000	0,3565
477	1	1	0	1	1,0000	0,0000	0,1782
525							0

Si les intervalles sont nombreux, on peut simplifier la présentation de la table de survie en définissant des intervalles plus larges. On peut choisir comme limites d'intervalle, et c'est une convention courante, les temps correspondants à chacun des décès qui précède immédiatement une donnée censurée. Dans l'exemple, ces limites sont alors 42, 232 et 273. Avec ces nouvelles limites, le tableau 11-2 devient le tableau 11-3.

Dans la méthode Kaplan-Meier, le souci est d'estimer la fonction de survie aux différents temps de décès observés. A cet égard, cette méthode colle davantage à la description des données observées.

Tableau 11-3: Table de survie

Int	*V*	*D*	*EV*	*O*	*q*	*p*	*S*
0	12	1	0	12	0,0833	0,9167	1
42	11	3	1	10	0,3000	0,7000	0,9167
232	7	1	1	6	0,1667	0,8333	0,6417
273	5	3	2	3	1,0000	0,0000	0,5347
525	0						0

Méthode actuarielle

La méthode actuarielle est une procédure courante d'estimation de la fonction de survie lorsque l'on choisit d'utiliser des intervalles de temps fixes. Ce qui est surtout spécifique à cette méthode, c'est la manière dont elle tient compte des exclus-vivants (*EV*). La méthode de Kaplan-Meier rapporte les exclus-vivants *au début* de l'intervalle, alors que la méthode actuarielle les rapporte *pendant* l'intervalle.

Nous allons expliquer la façon actuarielle de traiter les exclus-vivants en rappelant les données de l'exemple des 12 patients, disposées en ordre croissant:

42, 105 + , 126, 189, 232, 232 + , 273, 367 + , 399 + , 413, 477, 525,

et pour lesquelles on a convenu des intervalles fixes suivants:

0-60-120-180-240-300-360 420-480-540.

Au début de l'intervalle 360 - , il reste encore 5 survivants (367 + , 399 + , 413, 477, 525). Dans cet intervalle, il y a 2 exclus-vivants, soit au 367^{e} et au 399^{e} jour. A combien peut-on alors estimer le nombre d'individus en observation (O) pendant cet intervalle? Ce nombre est au plus 5 et au moins 3. La méthode actuarielle suppose que les exclus-vivants dans un intervalle ont, en moyenne, été exposés au risque de décès pendant la moitié de l'intervalle ou, de façon équivalente, que la moitié des exclus-vivants ont été exposés au risque de décès pendant l'intervalle complet. En conséquence, on estime le nombre d'individus en observation *(O)* en soustrayant du nombre de vivants au début de l'intervalle *(V)* la moitié du nombre des exclus-vivants dans l'intervalle *(EV)*. Ainsi:

$$O = V - \frac{1}{2}EV$$

Dans notre exemple: $O = 5 - \frac{1}{2} \times 2.$

En tenant compte de ces modifications, la table de survie actuarielle pour les 12 patients se présente comme au tableau 11-4.

Tableau 11-4: Table de survie actuarielle

Int	*V*	*D*	*EV*	*O*	*q*	*p*	*S*
0	12	1	0	12,0	0,0833	0,9167	1,0000
60	11	0	1	10,5	0,0000	1,0000	0,9167
120	10	1	0	10,0	0,1000	0,9000	0,9167
180	9	2	1	8,5	0,2353	0,7647	0,8250
240	6	1	0	6,0	0,1667	0,8333	0,6309
300	5	0	0	5,0	0,0000	1,0000	0,5257
360	5	1	2	4,0	0,2500	0,7500	0,5257
420	2	1	0	2,0	0,5000	0,5000	0,3943
480	1	1	0	1,0	1,0000	0,0000	0,1972
540	0						0

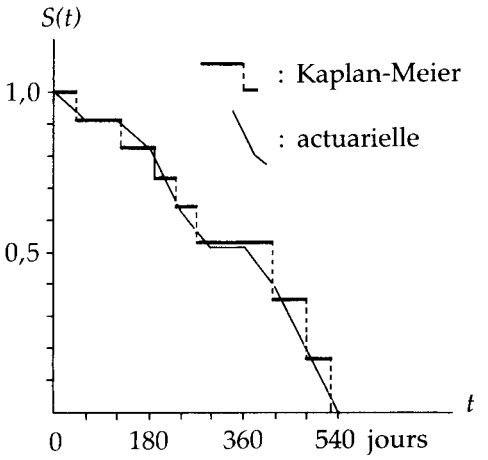
L'estimation de *S(t)* se fait, comme pour la méthode de Kaplan-Meier, en multipliant les probabilités *(p_i)* de survie.

$$S(t_i) = P_0 \, x \, P_1 \, x \ldots \text{ x } p_i \text{ — } 1$$

On note qu'en chaque point où la fonction de survie a été estimée, par l'une ou l'autre des deux méthodes, le calcul est le résultat d'un produit de probabilités. Conséquemment, on peut comprendre *S*(60), par exemple, comme la *probabilité cumulative* de survie à 60 jours.

La figure 11-8 représente la fonction de survie estimée respectivement par la méthode actuarielle (tableau 11-4) et celle de Kaplan-Meier (tableau 11-2).

Figure 11-8



Dans le cas de la méthode actuarielle, ou plus généralement celles à intervalles fixes, l'investigateur a le souci d'estimer la fonction de survie en des points qui l'intéressent. Quelle est, par exemple, la probabilité cumulative de survie à 1 an, 2 ans, etc.? De ce point de vue, la méthode actuarielle se prête bien à des comparaisons entre études différentes si les mêmes intervalles ont été choisis. La méthode actuarielle est surtout utilisée quand on a un grand nombre de patients.

La méthode actuarielle n'est pas la seule procédure qui utilise des intervalles fixés. Elle tire sa particularité de la façon dont elle traite les exclus-vivants en comptant pour chacun un demi-intervalle d'observation. On peut imaginer bien d'autres façons de tenir compte des exclus-vivants. Une façon est d'éliminer les exclus-vivants au début de l'intervalle considéré. Dans ce cas, le calcul du nombre *O* d'individus en observation est:

$$O=V-EV.$$

La méthode qui adopte cette attitude face aux exclus-vivants est parfois dite la méthode à intervalles complets.

SURVIE RELATIVE

Considérons à nouveau les 12 individus atteints du même cancer. Huit sont décédés pendant l'étude : un de ces décès est survenu, supposons, à la suite d'un accident de voiture. Les tables de survie construites antérieurement n'ont pas distingué les causes de décès. Tous les décès ont été comptés sans distinguer les causes.

Inclure les décès toute cause entraîne, c'est vraisemblable, une sous-estimation dans la probabilité cumulative de survie, prise comme mesure de la gravité de la maladie. Mais éliminer les décès par autres causes déclarées est une opération délicate. Le décès par accident de voiture est-il entièrement dû à l'accident lui-même ou ne serait-il pas de quelque façon lié au cancer considéré? C'est une distinction bien difficile à établir.

Comment alors corriger la sous-estimation? Généralement, on effectue cette correction en comparant pour un même intervalle les *probabilités cumulatives de survie observées à* des *probabilités cumulatives de survie attendues.* Les probabilités attendues sont celles calculées à partir d'un groupe similaire à celui des patients quant à l'âge, au sexe, à la période de temps considérée ou à d'autres caractéristiques pertinentes, mais un groupe non spécifiquement affecté par la maladie. Le calcul de ces probabilités attendues dépasse le cadre que nous nous sommes fixé ici.

La *survie relative* est définie comme le rapport de la probabilité cumulative de survie observée à celle attendue.

$$\text{Survie relative (à 6 mois)} = \frac{\text{Probabilité cumulative de survie observée (à 6 mois)}}{\text{Probabilité cumulative de survie attendue (à 6 mois)}}.$$

Reportons-nous à nouveau à l'exemple des 12 patients atteints du cancer dont la survie est décrite au tableau 11-4. La probabilité cumulative de survie à 6 mois (180 jours) a été estimée à 0,825. Supposons que la probabilité cumulative de survie attendue à 6 mois soit de 0,96. La survie relative à 6 mois est de 0,86 (soit 0,8250/0,96). En quelque sorte, un individu atteint de ce cancer aurait 82,5 chances sur 100 de survivre à 6 mois si l'on admet toutes les causes de décès. Par ailleurs, si l'on ne conserve comme seule cause possible de décès que le cancer considéré, la survie du patient à 6 mois serait de 86 chances sur 100.

Une survie relative égale à 1 indique que la mortalité est la même dans le groupe de patients que dans le groupe similaire ou dans la population en général. Une survie relative inférieure à 1 signifie une mortalité plus forte dans le groupe de patients. On comprend qu'une plus forte létalité se traduit par une plus faible survie relative. La survie relative est en définitive une correction pour les décès dus à d'autres causes.

RÉSUMÉ

La durée de survie à une maladie est calculée à partir d'un événement-origine, par exemple la date d'hospitalisation. On s'intéresse à la probabilité cumulative de survie ou fonction de survie. Pour un individu, celle-ci mesure, depuis l'événement-origine, la probabilité de survivre au delà d'une durée déterminée. L'estimation de la probabilité cumulative de survie doit souvent être faite à partir d'un groupe d'individus dont, pour certains, la date de décès n'est pas connue. Les données de survie forment alors une série incomplète. Les données sont censurées, rendant ainsi le calcul de la probabilité cumulative de survie plus difficile. L'estimation est faite principalement par les méthodes Kaplan-Meier ou actuarielle. La méthode Kaplan-Meier utilise une approche où les intervalles de temps sont variables; la méthode actuarielle, par contre, utilise une approche où les intervalles de temps sont fixés.

Symboles

$S(t)$, $S(t_i)$: fonction de survie au temps t, au temps t_i

$F(t)$: fonction de répartition

Prob $(T > t)$: probabilité que la durée de survie T soit supérieure à une valeur donnée t

V, V_i : vivants au début de l'intervalle, de l'intervalle i

EV, EV, : exclus-vivants dans un intervalle, dans l'intervalle i

O, O_i : individus en observation dans un intervalle, dans l'intervalle i

D, D_i : individus décédés dans un intervalle, dans l'intervalle i

q, q_i : probabilité de décès dans un intervalle, dans l'intervalle i

p, p_i: probabilité de survie sur un intervalle, sur l'intervalle i

Formules

$$S(t) = \text{Prob}\,(T > t)$$

$$F(t) = 1 - S(t) = \text{Prob}\,(T \leqslant t)$$

$$S(t_i) = p_0 \times p_1 \times \ldots \times p_{i-1}$$

$$= \left(1 - \frac{D_0}{O_0}\right)\left(1 - \frac{D_1}{O_1}\right)\left(1 - \frac{D_2}{O_2}\right)\ldots\left(1 - \frac{D_{i-1}}{O_{i-1}}\right)$$

$$\text{Survie relative (à 1 an)} = \frac{\text{probabilité cumulative de survie observée (à 1 an)}}{\text{probabilité cumulative de survie attendue (à 1 an)}}.$$

LECTURES SUGGEREES

1. JENICEK, M. et CLÉROUX, *R.* ***Épidémiologie clinique,*** Saint-Hyacinthe, Edisem, 1985, chapitre 5, section 5.4, pp. 171-182.
2. LEE, E.T. ***Statistical Methods for Survival Data Analysis,*** Belmont (USA), Lifetime Learning Publications, 1980, chapitres 2, 3 et 4, pp.9-121.
3. SCHWARTZ, D. FLAMANT, R. et LELLOUCH, J. ***L'essai thérapeutique chez l'homme,*** Paris, Flammarion, 1970, chapitre XXIII, pp.211-238.

CHAPITRE 12

Mesures de validité des tests diagnostiques (ou de dépistage)

Les mesures de validité des tests diagnostiques ou de dépistage sont des probabilités. Ce chapitre distingue la validité intrinsèque de la validité prédictive. Les mesures de la validité intrinsèque sont la sensibilité et la spécificité; celles de la validité prédictive, les valeurs prédictives positive et négative. On définit ces mesures et discute leurs interrelations. On présente également une mesure de la capacité pour un test de bien classifier les sujets. On considère deux tests appliqués en parallèle ou en série. Finalement, on énonce quelques remarques sur le choix d'un test diagnostique.

Un test doit être valide si on veut l'intégrer au processus de diagnostic ou de dépistage d'une maladie. Un test réagit-il correctement à la présence ou à l'absence de la maladie? Cette question touche la *validité intrinsèque* du test. Mais, pour un utilisateur, une autre question subsiste. Un individu positif au test est-il affecté par la maladie? En d'autres termes, dans son utilisation, un test est-il un bon indicateur de la présence ou non de la maladie? Cette deuxième question se rapporte à la *validité prédictive* du test. La validité intrinsèque d'un test est préalable à sa validité prédictive. Comment peut-on utiliser un test pour le diagnostic d'une maladie si, au départ, il ne réagit pas spécifiquement à la présence de la maladie? On approfondit plus loin le lien entre la validité intrinsèque et la validité prédictive.

VALIDITÉ INTRINSÈQUE : SENSIBILITÉ ET SPÉCIFICITÉ

La validation d'un nouveau test passe d'abord par un jugement sur son aptitude à reconnaître les malades et les non-malades. On veut mesurer sa double capacité de réagir positivement à la présence de la maladie et de ne réagir positivement qu'en sa présence.

Sensibilité et **spécificité**

On appelle *sensibilité (Sn)* du test sa capacité de donner un résultat positif quand la maladie est présente. Dans le langage des probabilités, la sensibilité mesure la probabilité conditionnelle que le test soit positif lorsque la maladie est présente. La sensibilité est estimée par la proportion de résultats positifs par suite de l'application du test à un groupe d'individus reconnus comme ayant la maladie.

On appelle *spécificité (Sp)* du test cette capacité de donner un résultat négatif quand la maladie est absente. Dans le langage des probabilités, la spécificité mesure la probabilité conditionnelle que le test soit négatif lorsque la maladie est absente. La spécificité est estimée par la proportion de résultats négatifs conséquemment à l'application du test à un groupe d'individus reconnus comme n'ayant pas la maladie.

Calcul de la sensibilité et de la spécificité

Considérons un test T proposé pour le diagnostic de la maladie *M.* Le test peut se révéler positif (T+) ou négatif (T—). Pour déterminer la sensibilité de ce test, on l'applique à *M,* sujets reconnus comme ayant la maladie *M (M +).* Pour déterminer la spécificité du test, on l'applique à M_0 sujets reconnus comme n'ayant pas la maladie *M (M —).* Les résultats du test sont alors présentés dans un tableau 2 x 2, comme au tableau 12-1.

Tableau 12-1

		M	
		+	−
T	+	a	b
	−	c	d
		M_1	M_0

On a,

$$\text{Sensibilité} = Sn = \text{Prob}\,(T+/M+) \simeq a/M_1$$
$$\text{Spécificité} = Sp = \text{Prob}\,(T-/M-) \simeq d/M_0$$

Au tableau 12-1, la cellule des a est celle dite des vrais positifs *(VP),* la cellule des *b* des faux positifs *(FP),* la cellule des *c* des faux négatifs *(FN)* et enfin la cellule des *d* celle des vrais négatifs *(VN).*

Un test T pour l'infection tuberculeuse est passé à 124 sujets tuberculeux. Parmi ceux-ci, 109 sont positifs au test T, alors que 15 sont négatifs. La sensibilité du test est estimée à 88 % :

$$Sn \simeq \frac{a}{M_1} = \frac{109}{124} = 0{,}88 \text{ ou } 88\,\%.$$

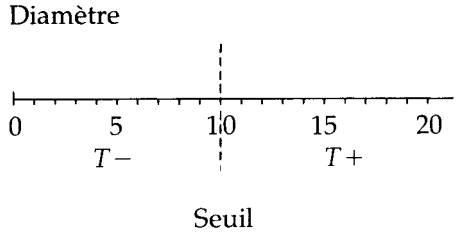
Le même test T est passé à 97 sujets sains, c'est-à-dire qui n'ont pas la maladie. Parmi ceux-ci, 76 réagissent négativement au test T, alors que 21 ont une réaction positive. La spécificité du test est estimée à 78 % :

$$Sp \simeq \frac{d}{M_0} = \frac{76}{97} = 0{,}78 \text{ ou } 78\,\%.$$

Un bon test, au sens de la validité intrinsèque, a une bonne sensibilité et une bonne spécificité.

Les définitions de sensibilité et de spécificité suggérées ici relèvent d'un cadre conceptuel simplifié. Ainsi:

- La maladie *M* en cause est une entité pathologique bien définie suivant les limites des connaissances médicales et reconnaissable par un procédé conventionnel et sûr. Ce procédé conventionnel (ou test standard) fait foi de la réalité ou vérité pathologique. La biopsie, par exemple, peut constituer un test standard pour certaines maladies graves. A la limite, l'autopsie.
- La maladie est considérée d'un strict point de vue dichotomique. La maladie est présente *(M* +) ou absente *(M —)* chez l'individu.
- De façon opérationnelle, le test passe par l'observation d'une variable indicatrice (ou critère) dont certaines valeurs ou classes de valeurs peuvent correspondre à la présence de la maladie.
- Le résultat du test s'exprime de façon dichotomique. Le résultat est positif (*T*+) ou négatif (*T*—) suivant les valeurs de la variable indicatrice. Dans le cas où la variable indicatrice est dichotomique (la reconnaissance d'un signe, d'une caractéristique, d'un organisme...), le signe du test correspond directement aux valeurs de la variable. Dans le cas où la variable indicatrice est quantitative, le résultat du test est une valeur numérique qu'il faut interpréter à partir d'un *seuil de positivité* préalablement fixé sur une échelle de la variable. A partir de ce seuil, le sujet est déclaré ou positif ou négatif suivant le résultat numérique du test, par exemple, le test de Mantoux pour la tuberculose. L'application de la tuberculine provoque une induration dont le diamètre est mesuré en millimètres. Pour juger de la présence ou non de l'infection tuberculeuse, le diamètre de l'induration observée est comparé au seuil de positivité convenu (soit 8 mm, 10 mm ou autre). Si le diamètre est déclaré égal ou supérieur au seuil, alors le résultat au test est déclaré positif. Autrement, il est négatif. On comprend ici qu'il pourrait y avoir plus de deux catégories de résultats. Par exemple, on pourrait avoir, comme c'est souvent le cas, la catégorie « douteux ».



Relation entre sensibilité et spécificité

La seule donnée de sensibilité (ou de spécificité) ne permet pas de se faire une idée correcte sur la validité intrinsèque d'un test. Par exemple, il ne suffit pas qu'un test soit sensible à 90 % pour le reconnaître valide. Si en plus, on sait que 90 % des sujets sains réagissent aussi positivement à ce test (donc spécificité de 10 %), on doit conclure que la réaction est indépendante de la présence de la maladie. Dans ce cas, malgré une bonne sensibilité, le test n'a aucune validité intrinsèque. Il ne sait pas discriminer entre sujets sains et sujets malades.

L'absence d'association entre le résultat au test et l'état pathologique se traduit par la valeur 1 du rapport $Sn/(1 - Sp)$. (En décision médicale, ce rapport est appelé rapport de vraisemblance.) Un rapport supérieur à 1 indique une association. Celle-ci est d'autant plus forte que ce rapport est plus grand. Un rapport inférieur à 1 est théoriquement possible, mais en pratique d'aucun intérêt. En effet, un tel rapport corrrespond à : $Sn + Sp < 1$, ce qui implique que la sensibilité et la spécificité ne peuvent pas être simultanément supérieures à 50 %.

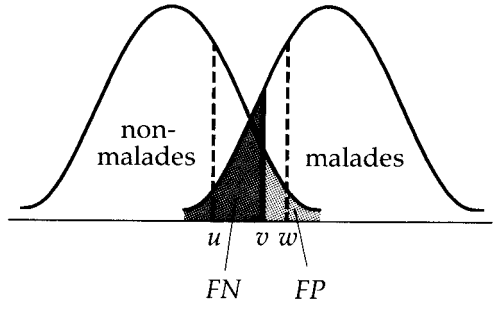
Quand la variable indicatrice d'un test est quantitative, la sensibilité et la spécificité sont toutes deux fonctions du seuil de positivité choisi et sont appelées à varier en sens inverse suivant le changement de ce seuil. Ainsi, l'amélioration de la sensibilité qu'apporte un changement du seuil de positivité s'accompagne d'une perte de spécificité et inversement. Cette relation peut être comprise en se référant à la figure 12-1.

Si le seuil de positivité passe de v à w, le nombre de faux négatifs *(FN)* va croître, ce qui correspond à une perte de sensibilité. Par contre, le nombre de faux positifs *(FP)* va diminuer, ce qui correspond à une amélioration de la spécificité. Si le seuil passe de v à u, alors l'inverse se produit.

VALIDITÉ PRÉDICTIVE

Un test ayant une bonne validité intrinsèque (bonne sensibilité et bonne spécificité) est-il pour autant un bon instrument de diagnostic ou de dépistage? En d'autres termes, le résultat positif d'un test correspond-il à une probabilité élevée d'être affecté par la maladie? Un résultat négatif correspond-il à une probabilité élevée d'être exempt de la maladie? Comme nous le verrons, une forte validité intrinsèque n'assure pas nécessairement au test une bonne validité prédictive. La validité prédictive est influencée aussi par la prévalence relative de la maladie. Les deux mesures de la validité prédictive sont la valeur prédictive positive d'un test positif et la valeur prédictive négative d'un test négatif.

Figure 12-1



Valeurs prédictives positives et négatives

La *valeur prédictive positive d'un test positif (Vp +)* mesure la probabilité conditionnelle que la maladie *(M +)* soit présente lorsque le test est positif *(T+)*. La valeur *Vp+* est estimée par la proportion de malades chez les positifs au test.

La *valeur prédictive négative d'un test négatif (Vp —)* mesure la probabilité conditionnelle que la maladie *(M —)* soit absente lorsque le test est négatif *(T—)*. La valeur *Vp—* est estimée par la proportion de sujets sains chez les négatifs au test.

Pour abréger, nous remplaçons dans la suite les expressions valeur prédictive positive d'un test positif et valeur prédictive négative d'un test négatif respectivement par *valeur prédictive positive* et *valeur prédictive négative.*

Calcul des deux valeurs prédictives

Considérons un test *T* proposé pour le diagnostic de la maladie *M.* On veut déterminer ses valeurs prédictives *Vp* + et *Vp* — . Deux approches sont alors possibles.

CALCUL EMPIRIQUE

On applique le test à un groupe d'individus. Ce groupe se subdivise en N_1 sujets positifs au test et N_0 sujets négatifs au test. Chaque individu de chaque sous-groupe est examiné par un test standard pour déterminer la présence ou non de la maladie. Les résultats d'une telle démarche peuvent être présentés dans un tableau comme au tableau 12-2.

Tableau 12-2

		M +	M −	
T	+	a	b	N_1
	−	c	d	N_0

On a

$$Vp+ = \text{Prob}\,(M+/T+) \simeq a/N_1$$

$$Vp- = \text{Prob}\,(M-/T-) \simeq d/N_0$$

CALCUL BASÉ SUR LA FORMULE DE BAYES

Soit le test *T* proposé pour le diagnostic de la maladie *M*. Supposons que l'on connaisse la prévalence relative de la maladie, la sensibilité et la spécificité du test. Alors, à partir de ces trois éléments, il est facile d'estimer les valeurs *Vp* + et *Vp* — du test. Illustrons ce fait par un exemple.

Supposons que la sensibilité et la spécificité d'un test T soient respectivement de 80 et 75 %, alors que la prévalence relative *(Pr)* de la maladie est de 10 %. On peut écrire:

$$Pr = \text{Prob}\,(M+) = 0{,}10$$

$$Sn = \text{Prob}\,(T+/M+) = 0{,}80$$

$$Sp = \text{Prob}\,(T-/M-) = 0{,}75.$$

Pour calculer les valeurs prédictives, il suffit de construire à partir de ces données un tableau 2 x 2 qui décrive un groupe type (de 1000 ou 10 000 sujets par exemple) partagé en malades (10 %) et non-malades (90 %).

Des 100 malades, le test en reconnaît 80 (Sn = 0,80) et des 900 non-malades, le test en reconnaît 675 (Sp = 0,75). En complétant par addition la marge verticale de ce tableau, on obtient le tableau 12-3.

Tableau 12-3

		M +	M −	
T	+	80	225	**305**
	—	20	675	**695**
		100	900	1000

On calcule aisément les valeurs prédictives:

$Vp+ = {}^{80}/_{305} = 0{,}26$ ou 26 %

$Vp- = {}^{675}/_{695} = 0{,}97$ ou 97 %.

Le calcul des valeurs prédictives peut être fait de façon plus formelle en adaptant à la présente situation la formule de Bayes développée au chapitre 10.

Rappelons que la valeur prédictive positive $Vp+$ est la probabilité conditionnelle d'être affecté par la maladie quand le test T est reconnu comme positif:

$$Vp+ = \text{Prob}(M+/T+).$$

Un individu positif au test peut être affecté ($M+$) ou non-affecté ($M-$) par la maladie M. Ces deux événements forment une partition. En ce cas, la formule de Bayes permet d'écrire :

$$\text{Prob}(M+/T+) = \frac{\text{Prob}(M+)\,\text{Prob}(T+/M+)}{\text{Prob}(M+)\,\text{Prob}(T+/M+) + \text{Prob}(M-)\,\text{Prob}(T+/M-)} \qquad [1]$$

c'est-à-dire,

$$Vp+ = \frac{Pr \times Sn}{Pr \times Sn + (1 - Pr) \times (1 - Sp)} \qquad [1']$$

On peut facilement transposer les expressions [1] et [1'] pour obtenir la valeur prédictive négative.

$$\text{Prob}(M-/T-) = \frac{\text{Prob}(M-)\,\text{Prob}(T-/M-)}{\text{Prob}(M-)\,\text{Prob}(T-/M-) + \text{Prob}(M+)\,\text{Prob}(T-/M+)} \qquad [2]$$

c'est-à-dire,

$$Vp- = \frac{(1 - Pr)\,Sp}{(1 - Pr)\,Sp + Pr\,(1 - Sn)} \qquad [2']$$

Si on reprend les données de l'exemple décrit au tableau 12-3, on retrouve les mêmes valeurs:

$$Vp+ = \frac{0{,}10 \times 0{,}80}{(0{,}10 \times 0{,}80) + (1 - 0{,}10)\,(1 - 0{,}75)} = \frac{0{,}08}{0{,}08 + 0{,}225} = \frac{80}{305}$$

$$Vp- = \frac{(1 - 0{,}10)\,0{,}75}{(1 - 0{,}10)\,0{,}75 + 0{,}10\,(1 - 0{,}80)} = \frac{0{,}675}{0{,}675 + 0{,}020} = \frac{675}{695}$$

Les deux expressions [1'] et [2'] illustrent bien le fait que les valeurs prédictives sont fonction non seulement de la sensibilité et de la spécificité, mais aussi de la prévalence relative de la maladie.

Des expressions [1'] et [2'], on peut facilement déduire les expressions équivalentes:

$$Vp+ = \frac{1}{1 + \frac{1 - Sp}{Sn} \cdot \frac{1 - Pr}{Pr}}$$

$$Vp- = \frac{1}{1 + \frac{1 - Sn}{Sp} \cdot \frac{Pr}{1 - Pr}}$$

Ces deux dernières expressions nous permettent de noter que, pour une validité intrinsèque déterminée, c'est-à-dire pour une sensibilité et une spécificité données, la valeur prédictive *Vp* + croît avec la prévalence relative, alors que la *Vp* — décroît avec cette mesure. On peut d'ailleurs illustrer ces faits par des exemples, ceux des tableaux 12-4, 12-5 et 12-6 pour lesquels la sensibilité et la spécificité sont fixées respectivement à 0,90 et 0,80. Un même test, appliqué dans des situations différentes (prévalences relatives différentes), conduit à des valeurs prédictives elles aussi différentes.

Tableau 12-5

Clinique: soins généraux; $Pr = 0{,}50$

		M +	M —	
T	+	90	20	110
	—	10	80	90
		100	100	200

$$Vp+ = \frac{90}{110} = 0{,}82$$

$$Vp- = \frac{80}{90} = 0{,}89$$

Tableau 12-4

Santé publique (dépistage); $Pr = 0{,}10$

		M +	M —	
T	+	1800	3 600	5 400
	—	200	14 400	14 600
		2000	18 000	20 000

$$Vp+ = \frac{1800}{5400} = 0{,}33$$

$$Vp- = \frac{14\,400}{14\,600} = 0{,}99$$

Tableau 12-6

Clinique spécialisée; $Pr = 0{,}90$

		M +	M —	
T	+	162	4	166
	—	18	16	34
		180	20	200

$$Vp+ = \frac{162}{166} = 0{,}98$$

$$Vp- = \frac{16}{34} = 0{,}47$$

Nous résumons les résultats de ces trois exemples dans le tableau 12-7. La prévalence relative est plus élevée en clinique spécialisée (0,90) qu'en santé publique (0,10). On y observe aussi une *Vp*+ plus élevée (0,98 contre 0,33); en revanche, la *Vp*– est moins élevée (0,47 contre 0,99).

De façon générale, pour un test donné, la valeur prédictive *Vp* + varie dans le même sens que la prévalence relative, alors que la valeur *Vp* — varie en sens inverse. Les deux courbes de la figure 12-2 illustrent pour cet exemple, c'est-à-dire *Sn* = 0,90 et *Sp* = 0,80, la relation qui existe entre les valeurs prédictives et la prévalence relative.

On trouve des courbes de même allure générale, c'est-à-dire croissante pour V_p+ et décroissante pour V_p-, quelle que soit la validité intrinsèque du test. Lorsque le rapport de vraisemblance $\frac{Sn}{1 - Sp}$ est plus grand que 1 ou, en des termes plus pratiques, lorsque la sensibilité et la spécificité sont conjointement supérieures à 50 %, la spécificité influence davantage la valeur prédictive positive *Vp*+ que ne le fait la sensibilité. On peut illustrer ce fait en fixant la prévalence relative disons à 0,10 ou 10 % et en faisant varier respectivement les valeurs de *Sn* et *Sp* comme au tableau 12-8.

Tableau 12-7

	Pr	*Vp* +	*Vp*—
Santé publique	0,10	0,33	0,99
Soins généraux	0,50	0,82	0,89
Clinique spécialisée	0,90	0,98	0,47

Dans les mêmes conditions, on pourrait également observer que la sensibilité influence davantage la valeur prédictive négative (*Vp*—) que ne le fait la spécificité.

CAPACITÉ D'UN TEST DE BIEN CLASSER LES SUJETS

Considérons un test T utilisé pour le diagnostic d'une maladie M. Supposons la sensibilité et la spécificité respectivement de 80 et 70 % et la prévalence relative de la maladie de 40 %.

Figure 12-2

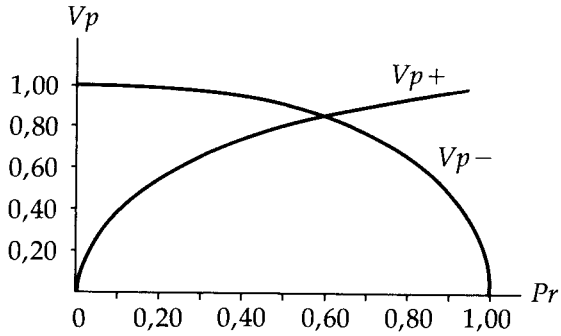


Tableau 12-8

Sp = 0,80		*Sn* = 0,80	
Sn	*Vp*+	*Sp*	*Vp*+
0,70	0,28	0,70	0,23
0,80	0,31	0,80	0,31
0,90	0,33	0,90	0,47
0,95	0,35	0,95	0,64
0,99	0,35	0,99	0,90

Pour un groupe de 500 sujets, on obtient le tableau 12-9.

Tableau *12-9*

		M +	*M* —	
T	+	**160**	*90*	
	—	*40*	**210**	
		200	*300*	*500*

Les valeurs en caractères gras 160 et 210 représentent le nombre de sujets *bien classés* respectivement chez les malades et les non-malades. Sur le nombre total de 500 sujets, la porportion de « bien classés » est égale à:

$$\frac{160 + 210}{500} = 0{,}74$$

Cette proportion qui correspond à $\frac{a + d}{M_1 + M_0}$ est appelée par certains auteurs l'« efficience » du test *T*. Elle peut s'exprimer au moyen de la prévalence relative, de la sensibilité et de la spécificité. En effet:

$$\frac{a + d}{M_1 + M_0} = \frac{M_1 \frac{a}{M_1} + M_0 \frac{d}{M_0}}{M_1 + M_0} = \frac{M_1 Sn + M_0 Sp}{M_1 + M_0} = Pr \times Sn + (1 - Pr) \times Sp$$

Cette proportion de bien classés est en définitive la somme pondérée de la sensibilité et de la spécificité, les poids étant la prévalence relative et sa valeur complémentaire.

APPLICATION DE DEUX (OU PLUSIEURS) TESTS

Il peut arriver que deux ou plusieurs tests soient utilisés pour le diagnostic d'une même maladie M. Ils peuvent être appliqués de façon concomitante (en parallèle) ou de façon séquentielle (en série). Nous allons présenter ces deux utilisations en nous restreignant d'abord au cas de deux tests, T_A et T_B.

Application en parallèle

Deux tests T_A et T_B sont appliqués en parallèle s'ils sont administrés aux patients de façon concomitante et indépendante. L'application de l'un est non conditionnelle à l'autre. Les résultats des deux tests sont alors simultanément considérés pour définir quatre classes de résultats :

$$\begin{array}{ll} (TAI, & TB_1) \\ (TA_{\mathrm{I}}, & TB_0) \\ (TA_0, & TB_1) \\ (TA_0, & TB_0) \end{array}$$

Ces quatre résultats peuvent être utilisés pour définir un autre test *T* (unique) où à des fins de généralisation, le signe du test est désigné par un indice numérique, 1 pour + et 0 pour —.

À chacune de ces quatre classes (T_{Ai}, T_{Bj}) où les indices *i* et *j* prennent les valeurs 1 ou 0, correspondent une valeur prédictive positive : Prob $(M_1 / T_{Ai} T_{Bj})$ et une valeur prédictive négative: Prob $(M_O | T_{Ai} T_{Bj})$. Les indices 1 et 0 de *M* indiquent respectivement la présence (+) et l'absence (—) de la maladie.

L'application du théorème de Bayes conduit aux relations suivantes qui permettent respectivement le calcul de valeurs prédictives positives et de valeurs prédictives négatives:

$$\text{Prob}\,(M_1/T_{Ai}T_{Bj}) = \frac{\text{Prob}\,(T_{Ai}/T_{Bj}M_1)\ \text{Prob}\,(T_{Bj}/M_1)\ \text{Prob}\,(M_1)}{\text{Prob}\,(T_{Ai}/T_{Bj}M_1)\ \text{Prob}\,(T_{Bj}/M_1)\ \text{Prob}\,(M_1) + \text{Prob}\,(T_{Ai}/T_{Bj}M_0)\ \text{Prob}\,(T_{Bj}/M_0)\ \text{Prob}\,(M_0)}$$

et

$$\text{Prob}\,(M_0/T_{Ai}T_{Bj}) = \frac{\text{Prob}\,(T_{Ai}/T_{Bj}M_0)\ \text{Prob}\,(T_{Bj}/M_0)\ \text{Prob}\,(M_0)}{\text{Prob}\,(T_{Ai}/T_{Bj}M_0)\ \text{Prob}\,(T_{Bj}/M_0)\ \text{Prob}\,(M_0) + \text{Prob}\,(T_{Ai}/T_{Bj}M_1)\ \text{Prob}\,(T_{Bj}/M_1)\ \text{Prob}\,(M_1)}$$

Si la validité intrinsèque du test T_A est indépendante de celle du test T_B, on peut montrer que la valeur prédictive positive, lorsque les deux tests sont positifs, est donnée par la formule:

$$\begin{aligned}\text{Prob}\,(M_1/T_{A1}T_{B1}) &= \frac{\text{Prob}\,(T_{A1}/M_1)\ \text{Prob}\,(T_{B1}/M_1)\ \text{Prob}\,(M_1)}{\text{Prob}\,(T_{A1}/M_1)\ \text{Prob}\,(T_{B1}/M_1)\ \text{Prob}\,(M_1) + \text{Prob}\,(T_{A1}/M_0)\ \text{Prob}\,(T_{B1}/M_0)\ \text{Prob}\,(M_0)} \\ &= \frac{(SnT_A)\,(SnT_B)\,(Pr)}{(SnT_A)\,(Sn\,T_B)\,(Pr) + (1 - SpT_A)(1 - SpT_B)(1 - Pr)} \\ &= \frac{1}{1 + \dfrac{(1 - SpT_A)}{SnT_A} \cdot \dfrac{(1 - SpT_B)}{SnT_B} \cdot \dfrac{(1 - Pr)}{Pr}}\end{aligned} \qquad [3]$$

et la valeur prédictive négative lorsque les deux tests sont négatifs par la formule:

$$\text{Prob}\,(M_0/T_{A0}T_{B0}) = \frac{1}{1 + \dfrac{(1 - SnT_A)}{SpT_A} \cdot \dfrac{(1 - SnT_B)}{SpT_B} \cdot \dfrac{Pr}{1 - Pr}} \qquad [4]$$

SnTA et SpT_A désignent respectivement la sensibilité et la spécificité du test T_A (il en va de même pour le test T_B).

Supposons que la sensibilité et la spécificité du test T_A soient respectivement de 90 % et 80 %, celles du test T_B de 70 % et 60 %. De plus, supposons que la prévalence relative de la maladie soit de 10 %. Nous admettrons que les résultats au test T_A sont indépendants de ceux du test T_B. Le calcul de la valeur prédictive lorsque les deux tests sont positifs et de la valeur prédictive négative lorsque les deux tests sont négatifs peut être fait à partir soit des formules [3] et [4], soit du tableau 12-10, construit par simulation pour un groupe de 1000 sujets répartis suivant les conditions fixées.

Tableau 12-10

				M		
				+	—	
T_A	+	T_B	+	63	72	135
			—	27	108	135
	—	T_B	+	7	288	295
			—	3	432	435
				100	900	1000

À partir des formules, on a:

$$\text{Prob}\,(M_1/T_{A1}T_{B1}) = \frac{1}{1 + \dfrac{(1 - 0{,}80)\,(1 - 0{,}60)}{(0{,}90)\,(0{,}70)} \cdot \dfrac{(1 - 0{,}10)}{0{,}10}} = 0{,}467$$

et

$$\text{Prob}\,(M_0/T_{A0}T_{B0}) = \frac{1}{1 + \dfrac{(1 - 0{,}90)\,(1 - 0{,}70)}{(0{,}80)\,(0{,}60)} \cdot \dfrac{0{,}10}{1 - 0{,}10}} = 0{,}993$$

Du tableau 12-10, on obtient:

$$\text{Prob}\,(M_1/T_{A1}T_{B1}) = \frac{63}{135} = 0{,}467$$

et

$$\text{Prob}\,(M_0/T_{A0}T_{B0}) = \frac{432}{435} = 0{,}993$$

L'application en parallèle de deux tests favorise la valeur prédictive positive lorsque les deux tests sont positifs et la valeur prédictive négative lorsque les deux tests sont négatifs.

En d'autres termes, la valeur prédictive positive totale (0,467) est généralement plus forte que les valeurs prédictives positives partielles (celles des tests T_A et T_B qui sont ici respectivement 0,333 et 0,163). De même, la valeur prédictive négative totale (0,993) est généralement plus forte que les valeurs prédictives négatives partielles (celles des tests T_A et T_B qui sont ici respectivement 0,986 et 0,947) .

Il est facile de généraliser les formules bayesiennes précédentes à plusieurs tests, de réponses multiples, appliqués pour le diagnostic

d'une maladie comprenant différents stades. Faisons-le dans le cas où il y a trois tests, T_A, T_B et Tc; pour chaque test, trois réponses i (i = 2, 1, 0) qui seront désignées par T_{Ai} pour le test T_A, etc., 2 voulant dire, par exemple, positif, 1 douteux et 0 négatif; pour la maladie M, deux stades k (k = 2, 1), k = 0 désignant l'absence de M. Dans cette situation, on a, par exemple :

$$\text{Prob}\,(M_1/T_{A_2}T_{B_2}T_{C_1}) = \frac{\text{Prob}\,(T_{A_2}/T_{B_2}T_{C_1}M_1)\ \text{Prob}\,(T_{B_2}/T_{C_1}M_1)\ \text{Prob}\,(T_{C_1}/M_1)\ \text{Prob}\,(M_1)}{\sum_{k=0}^{2} \text{Prob}\,(T_{A_2}/T_{B_2}T_{C_1}M_k)\ \text{Prob}\,(T_{B_2}/T_{C_1}M_k)\ \text{Prob}\,(T_{C_1}/M_k)\ \text{Prob}\,(M_k)}.$$

C'est la valeur prédictive de la maladie de stade 1 lorsque le test T_A est positif, le test T_B positif et le test Tc douteux. Si la validité intrinsèque des tests est indépendante l'une des autres, on peut montrer que, par exemple:

$$\text{Prob}\,(M_1/T_{A_2}T_{B_2}T_{C_1}) = \frac{\text{Prob}\,(T_{A_2}/M_1)\ \text{Prob}\,(T_{B_2}/M_1)\ \text{Prob}\,(T_{C_1}/M_1)\ \text{Prob}\,(M_1)}{\sum_{k=0}^{2} \text{Prob}\,(T_{A_2}/M_k)\ \text{Prob}\,(T_{B_2}/M_k)\ \text{Prob}\,(T_{C_1}/M_k)\ \text{Prob}\,(M_k)}.$$

Supposons que, pour chaque stade de la maladie M, la validité intrinsèque des trois tests T_A, T_B et Tc est indépendante l'une des autres et est décrite par l'ensemble des probabilités ci-contre :

Supposons de plus que la prévalence relative de la maladie par stade est donnée par: Pr_2 = 0,10, Pr_1 = 0,20 et Pro = 0,70. Alors, pour un individu positif au test T_A ($TA2$), positif au test T_B (T_{B2}) et douteux au test T_C (T_C,), la probabilité d'être affecté par la maladie au stade 1 (M,) est donnée par:

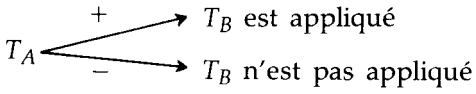
	M		
	2	1	0
T_{A2}	0,80	0,85	0,05
T_{A1}	0,15	0,10	0,05
T_{A0}	0,05	0,05	0,90
T_{B2}	0,70	0,75	0,08
T_{B1}	0,20	0,20	0,12
T_{B0}	0,10	0,05	0,80
T_{C2}	0,90	0,85	0,10
T_{C1}	0,08	0,10	0,15
T_{C0}	0,02	0,05	0,75

$$\text{Prob}\,(M_1/T_{A_2}T_{B_2}T_{C_1}) = \frac{0{,}85 \times 0{,}75 \times 0{,}10 \times 0{,}20}{0{,}80 \times 0{,}70 \times 0{,}08 \times 0{,}10 + 0{,}85 \times 0{,}75 \times 0{,}10 \times 0{,}20 + 0{,}05 \times 0{,}08 \times 0{,}15 \times 0{,}70} = 0{,}722.$$

Application en série

Deux tests diagnostiques T_A et T_B sont administrés en série pour la maladie M si l'application de l'un (disons T_B) est conditionnelle au résultat de l'autre (disons T_A). Le test T_A est d'abord appliqué, puis, le test T_B n'est appliqué qu'aux positifs du test T_A. Cette démarche est illustrée à la figure 12-3.

Figure 12-3



Nous allons nous référer à l'exemple des 1000 sujets décrits au tableau 12-10. Dans cet exemple, le test T_A a une sensibilité de 90 % et une spécificité de 80 %, alors que pour le test T_B ces valeurs sont respectivement de 70 % et 60 %. Enfin, la prévalence relative de la maladie M est de 10 %. Supposons que l'on applique aux 1000 sujets d'abord le test T_A. Suivant ce test, 270 individus (tableau 12-10) sont déclarés positifs. Remarquons que, de ce nombre, 90 sont affectés par la maladie. Appliquons maintenant le test T_B à ces 270 sujets. On obtient alors le tableau 12-11 (partie supérieure du tableau 12-10).

Tableau 12-11

		M +	M −	
T_B	+	63	72	135
	−	27	108	135
		90	180	270

La valeur prédictive positive du test T_B, conditionnelle au résultat positif du test T_A, est de 0,47, soit $^{63}/_{135}$. Remarquons que cette valeur prédictive positive conditionnelle est supérieure à celles des tests T_A et T_B qui, rappelons-le, étaient respectivement de 0,33 et 0,16.

Le test T_A a pour fonction de réduire le groupe initial à un sous-groupe où la prévalence relative de la maladie sera plus élevée. L'identification d'un groupe à prévalence relative plus forte conduit à une valeur prédictive positive $Vp+$ plus élevée, ce qui est souhaitable dans le cas où la maladie commande un traitement dispendieux ou dangereux, comportant par exemple des effets secondaires importants.

CHOIX D'UN TEST DIAGNOSTIQUE

Le choix d'un test pour le diagnostic d'une maladie repose sur un certain nombre de considérations comme la validité intrinsèque ou prédictive du test, la fréquence et la gravité de la maladie, les coûts associés au test, ... Sans entrer dans les techniques d'analyse de décision, quelques énoncés de base peuvent guider le choix d'un test.

- L'application d'un test doit d'abord être acceptable sur le plan médical et économique. On ne peut raisonnablement utiliser un test si son application comporte, en termes de coûts et de risque, plus d'inconvénients qu'il n'en peut régler ou prévenir.
- Si l'application d'un test est faite principalement pour la détection de la maladie, alors le test ayant une meilleure sensibilité offre de plus grandes chances de détecter la maladie. Dans ce contexte, le nombre de faux positifs peut poser un problème.
- Si la prévalence relative de la maladie est forte (c'est souvent le cas en milieu clinique), le nombre de faux positifs est généralement faible. Mais, même faible, ce nombre constitue un problème important lorsqu'un diagnostic faussement positif entraîne de graves conséquences pour la personne. La réduction du nombre de faux positifs passe par l'amélioration de la spécificité du test (contre souvent une diminution de la sensibilité).

Si 60 % des patients qui consultent pour la maladie M ont effectivement la maladie, alors un test sensible à 80 % et spécifique à 50 % donne une valeur prédictive positive de 71 %. Cela signifie que 29 % des positifs au test sont de faux positifs. Une amélioration de la spécificité de 50 % à 70 % réduit cette proportion de 29 % à 24 % si on suppose que la sensibilité est passée de 80 % à 75 %.

- Si la prévalence relative de la maladie est faible (c'est généralement le cas en santé publique), la détection de la maladie se fait généralement au prix d'un nombre important de faux positifs. Pour réduire ce nombre, trois moyens peuvent être envisagés :

Améliorer la spécificité du test
Comme on l'a vu, cette option se fait souvent contre la sensibilité du test.

Identifier les groupes à risque
Cette option correspond à une augmentation de la prévalence relative dans la population où le test est appliqué, donc à une amélioration de la valeur prédictive positive ($Vp+$).

Utiliser un second test
Le premier test permet de définir un groupe à risque chez qui le second test est appliqué.

RÉSUMÉ

La validité d'un test diagnostique (ou de dépistage) est intrinsèque ou prédictive. La validité intrinsèque d'un test est déterminée par sa sensibilité et sa spécificité. La sensibilité d'un test mesure sa capacité d'être positif lorsqu'il est appliqué à un sujet qui a la maladie. La spécificité d'un test mesure sa capacité d'être négatif lorsqu'il est administré à un sujet qui n'a pas la maladie. Un test peut avoir une bonne sensibilité sans avoir une bonne spécificité et inversement. Ce sont les valeurs conjointes de sensibilité et de spécificité qui nous permettent d'apprécier la validité intrinsèque d'un test. La validité prédictive est déterminée par les valeurs prédictives positive et négative, qui dépendent non seulement de la sensibilité et de la spécificité du test, mais aussi de la prévalence relative de la maladie. De façon générale, pour un test donné, la valeur prédictive positive varie dans le même sens que la prévalence relative, alors que la valeur prédictive négative varie en sens inverse. Deux tests peuvent être appliqués de façon concomitante et indépendante (application en parallèle) ou de façon séquentielle (application en série). Le choix d'un test pour le diagnostic d'une maladie, outre qu'il s'appuie sur des considérations de validité intrinsèque et prédictive, repose aussi sur la gravité de la maladie, les coûts associés, etc.

Symboles

		M		
		+	—	
T	+	a	b	N_1
	—	c	d	N_0
		M_1	M_0	

M, T: maladie, test

T+, *T* – *:* test positif, test négatif

M+, *M*– *:* malades, non-malades

Sn, Sp : sensibilité, spécificité

Vp + , *Vp* – *:* valeur prédictive positive, négative

Pr: prévalence relative

Formules

$$Sn = \text{Prob}\,(T+/M+) \simeq \frac{a}{M_1}$$

$$Sp = \text{Prob}\,(T-/M-) \simeq \frac{d}{M_0}$$

$$Vp+ = \text{Prob}\,(M+/T+) \simeq \frac{a}{N_1} = \frac{Pr \times Sn}{Pr \times Sn + (1 - Pr) \times (1 - Sp)} = \frac{1}{1 + \dfrac{1 - Sp}{Sn} \cdot \dfrac{1 - Pr}{Pr}}$$

$$Vp- = \text{Prob}\,(M-/T-) \simeq \frac{d}{N_0} = \frac{(1 - Pr)\,Sp}{(1 - Pr)\,Sp + Pr\,(1 - Sn)} = \frac{1}{1 + \dfrac{1 - Sn}{Sp} \cdot \dfrac{Pr}{1 - Pr}}$$

■ proportion de « bien classés »:

$$\frac{a + d}{M_1 + M_0} = Pr \times Sn + (1 - Pr) \times Sp$$

■ Application en parallèle de deux tests T_A et T_B:

$$\text{Prob}\,(M_1/T_{A1}T_{B1}) = \frac{1}{1 + \dfrac{(1 - SpT_A)}{SnT_A} \cdot \dfrac{(1 - SpT_B)}{SnT_B} \cdot \dfrac{(1 - Pr)}{Pr}}$$

$$\text{Prob}\,(M_0/T_{A0}T_{B0}) = \frac{1}{1 + \dfrac{(1 - SnT_A)}{SpT_A} \cdot \dfrac{(1 - SnT_B)}{SpT_B} \cdot \dfrac{Pr}{(1 - Pr)}}$$

LECTURES SUGGÉRÉES

1. FLETCHER, R.H., FLETCHER, S.W. et WAGNER, E.H. ***Clinical Epidemiology,*** Baltimore, Williams and Wilkins, 1982, chapitres 3 et 4, pp.41-74.
2. GALLEN, R.S. et GAMBINO, S.R. ***Beyond Normality,*** New York, John Wiley and Sons, 1975, chapitres 2, 3, 4, 5 et 6, pp. 9-48.
3. PHILIPPE, P. ***Épidémiologie pratique,*** Montréal, Presses de l'Université de Montréal, 1985, appendice, pp.131-138.

CHAPITRE 13

Valeur-*p* ou degré de signification

Ce chapitre présente la valeur-*p* ou valeur de *p* d'abord comme une mesure de compatibilité entre un résultat et une hypothèse. La valeur-*p* se révèle assez naturellement ensuite comme une mesure de la vraisemblance d'une hypothèse et se distingue des tests statistiques d'hypothèses bien qu'elle soit un élément essentiel à leur pratique. Nous verrons aussi la distinction entre la valeur-*p* et le seuil de signification a d'un test.

Sans qu'elle soit spécifique à l'épidémiologie, la valeur-*p*, ou degré de signification, est une mesure de probabilité largement répandue dans les articles scientifiques de cette discipline. En raison de sa grande diffusion, il nous semble intéressant de la présenter.

Nous insisterons surtout sur sa nature; son calcul est une question plutôt technique et, pour être effectué, il exige une connaissance des modèles de distribution de probabilité dont la présentation n'entre pas dans la composition de ce manuel (exception faite de celle sur l'importante distribution normale). Aussi, pensons-nous que le calcul d'une valeur-*p* est superflu à sa compréhension.

La valeur-*p*, dans son utilisation, peut se comprendre comme une mesure de la vraisemblance d'hypothèses statistiques. A partir de données qu'il observe, un chercheur s'interroge sur la valeur réelle d'une mesure de fréquence dans une population ou sur la moyenne réelle de la tension artérielle systolique. Un autre s'interroge sur l'existence d'une association entre un facteur d'exposition et une maladie. Parfois, c'est l'efficacité d'un traitement qui intéresse l'investigateur. Dans chacun de ces cas, et dans bien d'autres, le chercheur est appelé à mesurer la vraisemblance d'une hypothèse.

Avant de faire apparaître la valeur-*p* comme une mesure de vraisemblance, nous allons la présenter d'abord comme mesure de la compatibilité d'un résultat avec une hypothèse.

VALEUR-*p* COMME MESURE DE COMPATIBILITÉ D'UN RÉSULTAT AVEC UNE HYPOTHÈSE

Pour bien concrétiser les notions de compatibilité et de valeur-*p*, nous allons utiliser l'exemple classique du lancer d'une pièce de monnaie. Plus loin, d'autres exemples, plus épidémiologiques ceux-là, seront apportés pour renforcer la compréhension.

Compatibilité d'un résultat avec une hypothèse

Considérons l'expérience aléatoire de lancer 100 fois de suite une pièce de monnaie. Après une telle expérience, on observe 45 faces. S'agit-il d'un résultat prévisible, attendu ou plutôt d'un résultat inattendu, insolite?

Remarquons d'abord que la probabilité d'observer 45 faces dépend de celle d'observer face à chaque lancer. Ainsi, pour juger de la prévisibilité d'un résultat tel 45 faces, il est nécessaire de se référer à la valeur théorique de la probabilité liée à l'expérience aléatoire de base, « lancer la pièce de monnaie ». Quelle est la probabilité d'obtenir face au lancer de la pièce de monnaie. L'énoncé d'une telle valeur de probabilité, déduite ou postulée, est ce que l'on peut appeler une *hypothèse statistique.*

Une question intéressante se pose alors: celle de la compatibilité du résultat observé avec l'hypothèse retenue. Si l'on admet, pour la pièce de monnaie, l'hypothèse qu'elle est bien équilibrée, c'est-à-dire que la probabilité л d'observer face à chaque lancer est de $^1/_2$ (л = ½), peut-on dire que le résultat 45 faces lui est compatible?

Événement ponctuel et événement-intervalle

Pour apprécier le degré de compatibilité d'un résultat avec une hypothèse, il est naturel de se référer à la probabilité de l'événement qui correspond à ce résultat. Par exemple, au résultat 45 faces est associée spontanément la probabilité de l'événement 45 faces. En termes équivalents, le degré de compatibilité du résultat 45 faces avec l'hypothèse $л = {}^{1}/_{2}$ pourrait être la probabilité d'observer dans cette hypothèse l'événement 45 faces. Nous allons voir que cette probabilité ne fera pas l'affaire.

Sans entrer dans l'intimité des calculs (ce qui nécessiterait ici l'utilisation du modèle de distribution binomiale), on peut établir que la probabilité d'observer 45 faces après 100 lancers d'une pièce de monnaie bien équilibrée est de 5 % environ (Prob {f = 45} = 0,04847). C'est une probabilité plutôt faible pour un événement qui à priori apparaît assez plausible. Qu'en serait-il de l'événement 50 faces? Sa probabilité est d'environ 8 %. Elle demeure aussi étonnamment faible, beaucoup plus que le 50 % suggéré par l'intuition. Nous nous retrouvons face à un certain paradoxe. D'un côté, nous sommes en présence de résultats, 45 ou 50 faces, jugés à priori assez ou très compatibles avec l'hypothèse d'une pièce de monnaie bien équilibrée; d'un autre côté, les probabilités de ces *événements ponctuels,* 45 et 50 faces, sont relativement faibles.

Nous constatons ainsi que le jugement de compatibilité d'un résultat avec une hypothèse est difficile s'il est fait en référence aux événements ponctuels. Dans le cadre de notre exemple, pour les événements {0}, {30}, {40}, {50}, {60}, {70}, {100 faces}, on a respectivement les probabilités: 0,00000 (presque 0), 0,00002, 0,01084, 0,07959, 0,01084, 0,00002, 0,00000. Tous les résultats possibles pour cette expérience aléatoire, même les plus probables, réfèrent à des événements de probabilité faible. On peut imaginer la situation pour une expérience aléatoire qui consiste à lancer 1000 fois la pièce de monnaie. La probabilité de l'événement le plus probable, « observer 500 faces », est presque nulle. Que pourrait-on imaginer comme probabilité de l'événement « observer 500 000 faces » dans une expérience aléatoire comportant 1000 000 de lancers de la pièce de monnaie? N'est-elle pas pratiquement 0?

Pour pallier cette difficulté d'interprétation de la compatibilité, le résultat observé sera relié non pas à un événement ponctuel mais à un *événement-intervalle.* Voici comment ce dernier peut être défini.

Dans l'hypothèse d'une pièce de monnaie bien équilibrée, la valeur 50 faces (sur 100 lancers) correspond à une sorte de valeur centrale. Le résultat observé 45 faces contraste cette valeur centrale par une différence de 5. D'autres résultats la contrastent autant ou davantage. Ce sont les résultats aussi ou plus extrêmes que ne l'est 45 par rapport à 50. Ces résultats sont:

0 face, 1, ..., 44, 45, 55, 56, ..., 99, 100

Réunis, ils définissent l'ensemble des résultats possibles d'un événement que l'on note E_{45}. C'est ici l'événement-intervalle recherché.

Le degré de compatibilité du résultat 45 avec l'hypothèse de la pièce de monnaie bien équilibrée est apprécié en se référant à la probabilité cette fois de l'événement-

intervalle E_{45} plutôt qu'à la probabilité de l'événement ponctuel $\{f = 45\}$. Nous allons dénoter par Prob (E_{45}) la probabilité de l'événement-intervalle E_{45}. On peut établir pour la pièce de monnaie bien équilibrée que :

$$\text{Prob } (E_{45}) = 0{,}3682$$

Pour caractériser la compatibilité, cette dernière probabilité est plus intéressante que la probabilité de l'événement ponctuel (qui n'était que de 0,0485). Elle traduit mieux ce que suggère l'intuition quant au degré de compatibilité du résultat 45 faces avec l'hypothèse d'une pièce de monnaie bien équilibrée. Nous constatons ainsi que le jugement de compatibilité d'un résultat avec une hypothèse est possible s'il est fait en référence aux événements-intervalles.

Événement-intervalle et valeur-*p*

Les événements-intervalles ont l'avantage d'avoir des probabilités faibles, moyennes ou fortes, aussi fortes que 1, comme le montrent les valeurs suivantes (pour les 100 lancers d'une pièce de monnaie bien équilibrée).

$$\begin{aligned}
\text{Prob } (E_0) &= 0{,}0000;\\
\text{Prob } (E1) &= 0{,}0000;\\
\text{Prob } (E_{40}) &= 0{,}0569;\\
\text{Prob } (E_{45}) &= 0{,}3682;\\
\text{Prob } (E_{49}) &= 0{,}9204;\\
\text{Prob } (E_{50}) &= 1.
\end{aligned}$$

La probabilité est forte pour l'événement-intervalle qui regroupe des résultats dans le voisinage de la valeur centrale 50; à l'inverse, elle est faible pour celui qui regroupe uniquement des valeurs extrêmes, c'est-à-dire éloignées de la valeur centrale 50. Les probabilités des événements-intervalles se distinguent entre elles plus facilement que celles des événements ponctuels. En termes probabilistes donc, les événements-intervalles sont plus intéressants que les événements ponctuels pour juger de la compatibilité d'un résultat avec une hypothèse.

Ainsi, au résultat obtenu, nous associons la probabilité de l'événement-intervalle correspondant. Par exemple, au résultat 45 faces, nous associons la probabilité de l'événement-intervalle E_{45}, etc. Cette règle de correspondance entre le résultat obtenu et la probabilité d'un événement-intervalle bien spécifique définit ce que l'on appelle la *valeur-p* du résultat. La valeur-*p* est la mesure de compatibilité recherchée. Dans notre exemple,

$$\text{valeur-}p = \text{Prob } (^{E}45)$$

De façon générale,

$$\text{valeur-}p = \text{Prob } (E_k)$$

où E_k est l'événement-intervalle qui correspond au résultat k.

Pour une expérience déterminée, la valeur-*p* est donc la probabilité d'un événement (événement-intervalle) dont la composition est liée d'une part au résultat observé et d'autre part à l'hypothèse proposée. Quand on observe un résultat consécutif à une expérience aléatoire, la *valeur-p* mesure, pour une hypothèse proposée, la probabilité d'obtenir, dans une répétition de l'expérience, un résultat aussi ou plus extrême que celui observé. Il est important de répéter, pour mieux souligner, que la valeur-*p* est non seulement associée au résultat observé mais qu'elle est intimement liée à une hypothèse comme nous allons maintenant le voir.

Valeur-*p* et hypothèses statistiques

Dans notre exemple de la pièce de monnaie, la valeur-*p* a été calculée pour une hypothèse bilatérale (pièce de monnaie bien équilibrée). Les hypothèses peuvent être unilatérales ou bilatérales, simples ou composites. C'est ce que nous allons expliquer maintenant. Toujours à l'aide de l'expérience de 100 lancers d'une pièce de monnaie et le résultat observé 45 faces, nous donnerons la valeur-*p* pour différents types d'hypothèses. Par ces exemples, nous allons nous rendre compte que la valeur-*p* change avec l'hypothèse que l'on avance.

Hypothèse (H_a): la pièce de monnaie est bien équilibrée, c'est-à-dire que la probabilité л d'observer face à chaque lancer est $^1/_2$ (л = $^1/_2$).

Cette hypothèse est *simple* dans le sens qu'elle stipule une valeur unique pour л. Elle est *bilatérale* puisque les résultats qui lui sont compatibles se trouvent dans le voisinage de la valeur centrale 50, aussi bien à gauche qu'à droite. Les résultats qui s'en écartent (valeurs extrêmes) par valeurs supérieures comme 70 et 80, ou par valeurs inférieures comme 20 et 10, sont jugés peu ou très peu compatibles. Selon cette hypothèse, la valeur-*p* du résultat 45 faces est égale, comme on l'a vu, à 0,3682.

$$\text{valeur-}p = \text{Prob}\,(E_{45}) = 0{,}3682$$

Hypothèse (H_b): la pièce de monnaie n'est pas bien équilibrée; la probabilité л d'observer face à chaque lancer est égale à 0,60 (л = 0,60).

Toujours pour l'expérience des 100 lancers, la valeur centrale est ici 60 faces. Toute valeur se situant dans le voisinage de celle ci (à gauche ou à droite) correspond à un résultat compatible. Par contre, les valeurs aussi extrêmes que 20 ou 90 peuvent être considérées comme non-compatibles avec l'hypothèse H_b. C'est aussi une hypothèse simple et bilatérale. Le résultat 45 faces contraste la valeur centrale 60 par une différence de 15. L'événement-intervalle E_{45} se compose des résultats suivants:

0 face, 1, ..., 44, 45, 75, 76 ... 100

Selon l'hypothèse H_b, la valeur-*p* du résultat 45 faces est égale à 0,0046.

$$\text{valeur-}p = \text{Prob}\,(E_{45}) = 0{,}0046$$

Évidemment, le résultat 45 faces est moins compatible avec l'hypothèse H_b qu'avec l'hypothèse H_a.

Hypothèse (H_C): la probabilité d'observer face à chaque lancer est égale ou supérieure à 0,60 (л ≥ 0,60).

Le résultat 45 faces est peu compatible avec cette hypothèse; les valeurs extrêmes 20 et 10 le sont encore moins. Par contre, on comprend aisément que les résultats 80 et 90 sont compatibles avec l'hypothèse H_C.

Examinons de plus près, dans ses implications, l'hypothèse H_C. Quelle est la valeur centrale? Pour pouvoir l'identifier, il faut reconnaître la probabilité л d'observer face à chaque lancer. Quelle est cette probabilité? Est-ce 0,60, 0,72, 0,82, 0,85, ... ? Ce qu'établit l'énoncé de H_C, c'est que cette probabilité est supérieure ou égale à 0,60. Cet énoncé admet donc non pas une seule mais une multitude de valeurs possibles pour л, toute valeur numérique comprise entre 0,60 et 1, y comprises ces deux limites. En conséquence,

on peut comprendre H, comme une *hypothèse composite.* La probabilité 0,60 est une sorte de valeur limite inférieure acceptable pour le paramètre л. Aussi peut-on dire que la valeur 60 faces pour l'expérience des 100 lancers correspond à la valeur centrale (limite) pour H_C. Tout résultat dont la valeur se situe dans le voisinage de 60 ou lui est supérieure est compatible avec l'hypothèse; par contre, tout résultat qui s'écarte du voisinage de la valeur centrale (limite) 60, par valeurs inférieures seulement, lui est peu compatible. L'hypothèse HC est du type *unilatéral.* Le résultat 45 faces contraste la valeur centrale (limite) 60 par une différence de 15. L'événement-intervalle E_{45} se compose ici des résultats:

0 face, 1, ..., 44, 45

Ce sont les résultats qui contrastent autant ou davantage la valeur centrale 60, mais uniquement ceux qui lui sont inférieurs. Selon cette hypothèse, la valeur-*p* du résultat 45 faces est égale à 0,0017

Valeur-*p* – Prob (E_{45}) = 0,0017

Hypothèse (H_d)*:* la probabilité d'observer face à chaque lancer est égale ou inférieure à 0,40 ($л \leq 40$). 0,40).

Cette hypothèse, on le sait maintenant, est unilatérale et composite. Les résultats 20 et 10 sont compatibles avec l'hypothèse H_d, alors que les résultats 80 et 90 ne le sont pas. Le résultat observé 45 contraste la valeur centrale (limite) 40 par une différence de 5. L'événement-intervalle E_{45} se compose des résultats:

45 faces, 46, ..., 99, 100

Ce sont les résultats qui contrastent autant ou davantage la valeur centrale 40, mais uniquement ceux qui lui sont supérieurs. Selon cette hypothèse, la valeur-*p* du résultat 45 faces est égale à 0,1789

valeur-*p* = Prob (E_{45}) = 0,1789

Regardons maintenant l'interprétation de la valeur-*p* dans le cadre de certaines expériences aléatoires plus pertinentes à l'épidémiologie.

Prévalence relative de la maladie M

On suppose que la prévalence relative de la maladie *M* dans une certaine population est 10 %. (C'est l'énoncé d'une hypothèse bilatérale : *Pr* = 0,10). Pour avoir une idée plus juste de cette prévalence, on prélève de la population (c'est l'expérience aléatoire) un échantillon de 250 sujets qui sont examinés pour la maladie *M.* Sur cet échantillon, on dénombre 30 cas de la maladie. Quelle est la compatibilité de ce résultat avec l'hypothèse d'une prévalence relative de 10 % ? Le résultat observé 30 contraste la valeur centrale 25 par une différence de 5. Pour juger de la compatibilité, nous référons à l'événement-intervalle E_{30} qui comprend ici les valeurs :

0 cas, 1, ..., 19, 20, 30, 31, ..., 249, 250.

Selon cette hypothèse, la valeur-*p* de notre résultat est égale à 0,3423. Le degré de compatibilité est assez bon.

Moyenne de la tension artérielle dans une population

On suppose que la moyenne de la tension artérielle systolique dans une population est de 120 mmHg avec un écart-type de 10 mmHg. (C'est l'énoncé d'une hypothèse bilatérale: moyenne = 120 mmHg.) Pour avoir une idée plus juste de cette moyenne, on prélève de cette population

un échantillon aléatoire de 100 sujets qui sont examinés pour la tension artérielle. Dans cet échantillon, on trouve une tension moyenne de 122 mmHg. Quelle est la compatibilité de cette observation avec l'hypothèse d'une moyenne égale à 120 mmHg? Le résultat observé 122 contraste la valeur centrale 120 par une différence de 2. L'événement-intervalle E_{122} correspond alors à l'ensemble :

{tension ≤ 118 ou tension ≥ 122}

Selon cette hypothèse, la valeur-*p* du résultat 122 mmHg est 0,0456. Le degré de compatibilité est plutôt faible. Il est intéressant de noter ici que la probabilité d'un événement ponctuel, par exemple « tension moyenne égale exactement à 122 mmHg », est nulle. C'est le cas pour la probabilité de tout événement ponctuel défini dans le cadre d'une expérience aléatoire qui met en cause l'observation d'une variable continue. Cette situation montre bien l'utilité des événements-intervalles. Si la probabilité d'un événement ponctuel quelconque est nulle, il n'en est pas de même pour les événements-intervalles.

Association entre un facteur E et une maladie M
On planifie une étude cas-témoins pour déterminer l'existence d'une association entre le facteur E et la maladie *M*. Cinq cents (500) cas sont comparés à 500 témoins. On observe 450 cas exposés au facteur *E* contre 420 chez les témoins. Les proportions d'exposés chez les cas et chez les témoins sont respectivement de 0,90 ($^{450}/_{500}$) et 0,84 ($^{420}/_{500}$). Quelle est la compatibilité de ce résultat avec l'hypothèse qu'il n'y a pas d'association positive, que la différence entre les deux proportions est théoriquement au plus égale à 0? (Cette hypothèse est unilatérale: $л, - л \leq 0$.) La différence observée entre les deux proportions est 0,06. Elle contraste la valeur centrale (limite) 0 par 0,06. L'événement-intervalle $E_{0,06}$ correspond alors à l'ensemble :

{différence de proportion 0,06}

D'après cette hypothèse, la valeur-p du résultat « différence = 0,06 » est 0,0012. Le degré de compatibilité est très faible.

VALEUR-*p* COMME MESURE DE VRAISEMBLANCE D'UNE HYPOTHÈSE

Nous avons développé l'idée d'une valeur-*p* comme une mesure de la compatibilité, de la conformité d'une observation avec une hypothèse. Cette idée est proche de celle d'une valeur-*p* comprise comme mesure de vraisemblance d'une hypothèse. Par suite d'une expérience aléatoire, nous nous trouvons généralement en présence d'une part d'un résultat connu, observé, et d'autre part, d'une hypothèse dont la véracité est inconnue. Si la valeur-*p* du résultat observé est faible, ce qui traduit une faible compatibilité entre le résultat et l'hypothèse, nous pourrons mettre en doute la véracité de cette hypothèse; celle-ci nous apparaîtra peu vraisemblable. Au contraire, si la valeur-*p* est moyenne ou forte, nous pourrons plus difficilement mettre en doute la véracité de l'hypothèse; celle-ci nous apparaîtra alors vraisemblable.

Concrétisons cette idée de la valeur-*p* comme mesure de vraisemblance d'une hypothèse en reprenant certains des exemples évoqués précédemment.

Lancer d'une pièce de monnaie
En supposant que la pièce de monnaie est bien équilibrée, pour un résultat de 45 faces sur 100 lancers, on trouve une valeur-*p*

égale à 0,3682. Nous ne savons pas si cette hypothèse est correcte. Avec une telle valeur-*p*, nous sommes toutefois portés à considérer l'hypothèse de la pièce de monnaie bien équilibrée comme plutôt vraisemblable.

Prévalence relative de la maladie M
En tenant pour acquis que la prévalence relative est égale à 10 %, pour une observation de 30 cas sur 250 sujets, la valeur-*p* est égale à 0,3423. Avec cette valeur-*p*, on peut juger plutôt vraisemblable l'hypothèse d'une prévalence relative égale à 10 %.

Moyenne de tension artérielle
En supposant que la moyenne de tension artérielle dans une population est égale à 120 mmHg, pour une moyenne observée de 122 mmHg dans un échantillon de 100 sujets, la valeur-p est égale à 0,0456. Avec une telle valeur-*p*, on peut douter de la véracité de l'hypothèse d'une tension artérielle moyenne de 120 mmHg dans cette population. L'hypothèse apparaît peu vraisemblable.

Association entre un facteur E et une maladie M
On observe une différence de 0,06 entre les proportions de sujets exposés chez les cas et de sujets exposés chez les témoins. D'après l'hypothèse qu'il n'y a pas d'association positive entre le facteur *E* et la maladie *M,* la valeur-*p* a été estimée à 0,0012. Avec une telle valeur-*p*, on peut douter fortement de la véracité de cette hypothèse.

Dans un cadre non-décisionnel auquel appartient souvent la recherche épidémiologique, la valeur-*p* est prise en soi comme une mesure de la vraisemblance d'une hypothèse. Il revient au chercheur ou au lecteur d'interpréter cette valeur au vu des tailles d'échantillons, de la nouveauté des résultats ou de l'intérêt qu'ils suscitent.

VALEUR-*p* ET TEST STATISTIQUE

Nous n'avons pas ici comme dessein d'étudier les tests statistiques. Nous voulons seulement situer le calcul d'une valeur-*p* comme l'une des étapes de la pratique de tout test statistique. Pour l'essentiel, un *test statistique* est une procédure qui permet à un investigateur de choisir, non sans risque de se tromper, entre deux hypothèses.

Valeur-*p* comme étape d'un test statistique

Dans les exemples cités, le calcul de la valeur-p s'est toujours fait en supposant correcte une hypothèse : la pièce de monnaie est bien équilibrée; la prévalence relative de la maladie est égale à 10 %; la tension artérielle moyenne est égale à 120 mmHg; il n'y a pas d'association positive entre E et M. Dans chaque cas, c'est l'hypothèse dite *nulle,* désignée par H_o.

La valeur-*p* mesure donc la vraisemblance de l'hypothèse nulle (H_o). En présence d'une petite valeur-*p*, donc d'un manque important de compatibilité entre le résultat observé et H_o, nous pourrons mettre en doute l'hypothèse nulle et lui préférer sa négation ou *contre-hypothèse.* Cette dernière, souvent appelée (à tort en français) hypothèse alternative, est désignée par H_1. Pour chacun des quatre exemples, la contre-hypothèse (H_1) est:

— la pièce de monnaie n'est pas bien équilibrée;

— la prévalence dans la population n'est pas 10%;

— la tension artérielle moyenne n'est pas 120 mmHg;

— il y a une association positive entre *E* et *M.*

La valeur-*p* joue donc un rôle dans un test statistique. Elle permet de décider entre deux hypothèses: l'hypothèse nulle (H_o) et la contre-hypothèse (H_1). Le manque de compatibilité entre un résultat observé et l'hypothèse nulle force le choix de H_1. Par contre, une bonne conformité de ce résultat avec H_o ne constitue pas une preuve de la véracité de H_o, mais invite à son non-rejet.

Valeur-*p* et seuil de signification α

Il arrive que l'investigateur soit appelé, à la suite des résultats d'une étude, à prendre une décision en faveur d'une hypothèse. Pour quelle grandeur de la valeur-*p* un investigateur est-il prêt à rejeter l'hypothèse nulle pour accepter la contre-hypothèse? Parfois la vraisemblance est tellement faible (comme dans l'exemple de l'association entre *E* et *M),* qu'il serait ridicule de ne pas rejeter l'hypothèse nulle. Ce n'est cependant pas toujours aussi clair. Dans le cas où la valeur-*p* est 0,147, est-on prêt à déclarer invraisemblable l'hypothèse nulle? Est-on prêt à déclarer, sur le plan statistique, que la prévalence relative de la maladie dans une population est de l'ordre de 10 %? La réponse dépend de l'importance du risque consenti à rejeter l'hypothèse nulle lorsqu'elle est vraie. En d'autres termes, si l'hypothèse nulle est correcte (ce dont on n'est jamais sûr), on ne voudrait pas que le risque ou la probabilité de la rejeter soit plus grand qu'un certain seuil (appelé *seuil de signification ou de décision α).* Le risque consenti est généralement petit (on le comprend bien); il ne peut cependant pas être nul. Ce serait alors ne prendre aucun risque, ne jamais rejeter l'hypothèse nulle de peur de commettre une erreur. Classiquement, le seuil α est fixé à 0,05 (ou 5 %). Il pourrait être plus sévère (disons de 0,01) ou plus large (soit 0,10). Ce qui importe, dans le cadre de la prise de décision, c'est que le seuil soit fixé à priori, au moment de la planification de l'étude, et non pas à posteriori au vu des résultats. Une valeur-*p* inférieure à α est considérée comme petite. Elle entraîne le rejet de H_0 et, par voie de conséquence, l'acceptation de H_l.

FAUSSES INTERPRÉTATIONS DE LA VALEUR-*p*

La valeur-*p* n'est pas la probabilité d'obtenir par hasard un résultat aussi extrême ou plus extrême que celui qui a été observé. Il manque à cette phrase la référence à l'hypothèse donc au modèle de distribution à partir duquel la valeur-*p* a été calculée. Si l'hypothèse, donc la distribution, n'est pas connue, la valeur-*p* n'est pas calculable.

La valeur-*p* n'est pas non plus la probabilité d'obtenir par hasard le résultat observé étant donné une hypothèse. Cette dernière probabilité est, dans le cas de distribution continue on s'en souvient, toujours égale à zéro quel que soit le résultat observé. Il faut ajouter au dernier énoncé: ou un résultat plus extrême.

Enfin, la valeur-*p* n'est pas la probabilité que l'hypothèse nulle soit vraie. De façon complémentaire, elle n'est pas non plus la

probabilité que la contre-hypothèse soit vraie. La valeur-*p* mesure la vraisemblance de l'hypothèse nulle. À la vue d'un résultat, une hypothèse peut paraître vraisemblable sans être vraie pour autant.

RÉSUMÉ

La valeur-*p* est une probabilité. Elle mesure le degré de compatibilité entre un résultat observé et une hypothèse. Cette mesure est obtenue en calculant la probabilité d'un événement-intervalle. Pour une hypothèse proposée, l'événement-intervalle rassemble tous les résultats aussi ou plus extrêmes que le résultat observé. Quand on observe un résultat par suite d'une expérience aléatoire, la valeur-*p* mesure, pour une hypothèse proposée, la probabilité d'obtenir, dans une répétition de l'expérience, un résultat aussi ou plus extrême que celui observé. La valeur-*p* peut être comprise comme mesure de vraisemblance d'une hypothèse. Plus la valeur-*p* du résultat observé est faible, plus nous pouvons mettre en doute la véracité de l'hypothèse. La valeur-*p* est une étape d'un test statistique puisque celui-ci est une procédure qui, pour l'essentiel, permet de choisir entre deux hypothèses: l'hypothèse nulle et la contre-hypothèse. Dans un contexte décisionnel, le rejet ou le non-rejet de l'hypothèse nulle est lié au choix d'un seuil de signification. Si la valeur-*p* est inférieure à ce seuil, l'hypothèse nulle est rejetée.

Symboles

E_k: événement-intervalle associé au résultat k

Prob (E_k): probabilité de l'événement-intervalle

Formule

Valeur-*p* = Prob (E_k)

LECTURES SUGGÉRÉES

1. FLETCHER, R.H., FLETCHER, S.W. et WAGNER, E.W. ***Clinical Epidemiology,*** Baltimore, Williams and Wilkins, 1982, chapitre 9, pp. 153-160.
2. GOLDBERG, M. ***L'Épidémiologie sans peine,*** Paris, Éditions médicales Roland Bettex, 1986, pp. 69-72.
3. ROTHMAN, K.J. ***Modern Epidemiology,*** Boston, Little, Brown, 1986, chapitre 5, pp. 115-129.

PARTIE V

Validité et précision

CHAPITRE 14

Notion de justesse

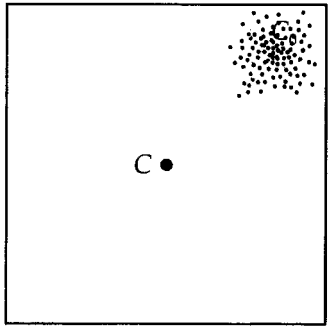
Ce chapitre différencie les deux concepts de précision et de validité d'une mesure et distingue entre validité externe et validité interne. La validité externe touche la question de la généralisation des résultats considérée du point de vue statistique ou scientifique; la validité interne réfère aux défauts de structure de l'étude.

Toute mesure faite en épidémiologie, qu'elle soit une mesure de fréquences, d'association ou autre, est sujette à l'erreur. D'un point de vue méthodologique, on doit viser à minimiser l'erreur, c'est-à-dire à tendre vers la plus grande justesse possible. La *justesse* d'une mesure peut être affectée par deux types d'erreur: l'erreur *aléatoire* et l'erreur *systématique.* La présence d'erreurs aléatoires conduit à des mesures moins précises, alors que celle d'erreurs systématiques (ou biais) mène à des mesures moins valides. On peut comprendre la *précision* comme l'absence relative d'erreurs aléatoires et la *validité* comme l'absence relative d'erreurs systématiques.

L'exemple du tireur permet de mieux comprendre la différence entre précision et validité, ou entre erreur aléatoire et erreur systématique. Le tireur vise à plusieurs reprises un point-cible *C*. Les points de touche pourraient se situer systématiquement à droite du point-cible *C*, comme à la figure 13-1.

Ce tir faussé, en ce sens que les tirs ont été exécutés comme s'il s'agissait d'une cible (C_0) totalement différente de *C*, caractérise un manque de validité ou l'erreur systématique.

Figure 13-1



Par ailleurs, sans référence aucune au point-cible *C*, on observe une variation dans les positions des points d'impact les uns par rapport aux autres. Cette situation caractérise le manque de précision ou l'erreur aléatoire. L'erreur totale peut s'expliquer par l'erreur systématique (ou biais) et par l'erreur aléatoire. En d'autres termes, la validité et la précision sont les deux composantes de la justesse.

PRÉCISION (ABSENCE D'ERREUR ALÉATOIRE)

La précision d'une mesure dépend de deux facteurs : la taille des échantillons et la variabilité du caractère étudié (écart-type). Plus la taille des échantillons est importante, plus grande est la précision des estimations. A l'inverse, plus grande est la variabilité du caractère étudié, moins bonne est la précision de l'estimation.

On comprend très bien qu'une étude visant à estimer la prévalence d'une certaine maladie dans une population donnera une estimation de meilleure précision avec un échantillon de 100 sujets qu'avec un échantillon de 10. Par ailleurs, pour une même taille d'échantillon (disons 100 sujets), l'estimation du poids moyen à la naissance aura une meilleure précision si l'écart-type est de 300 plutôt que 500 g.

Le problème de la précision des estimations relève de la statistique. D'ailleurs, cette discipline a développé un ensemble d'outils qui permettent de juger de la précision atteinte dans les estimations. Nous aborderons cette question de la précision dans le chapitre 17 qui se rapporte à l'estimation par intervalle de confiance.

VALIDITÉ (ABSENCE D'ERREUR SYSTÉMATIQUE)

La validité d'une mesure, on le rappelle, réfère à l'absence relative de biais ou d'erreur systématique dans les mesures. On peut distinguer deux sortes de validité: la validité externe et la validité interne. La première réfère à l'idée de généralisation des résultats obtenus, la deuxième aux caractéristiques qui touchent la structure même de l'étude.

Validité externe

La *validité externe* réfère au fait que les résultats d'une étude sont généralisables; il s'agit d'un concept relatif. La généralisation des résultats peut être faite à une population déterminée, alors qu'elle ne peut pas l'être à une autre population. Les résultats d'une étude qui a permis d'estimer la force d'association entre un facteur et une maladie chez l'homme peuvent être généralisés à une certaine population masculine où a été conduite l'étude. Mais le sont-ils à la population totale, comprenant hommes et femmes? Une mesure généralisable à une certaine population est, eu égard à cette population, valide au plan externe.

La *généralisation* peut être envisagée suivant deux points de vue : statistique ou scientifique.

GÉNÉRALISATION D'UN POINT DE VUE STATISTIQUE

D'un point de vue statistique ou d'échantillonnage, la généralisation est possible en autant que les résultats ont été obtenus à partir d'échantillons représentatifs de la population visée par l'étude. On comprend intuitivement qu'un échantillon est représentatif s'il reproduit les caractéristiques essentielles de la population de laquelle il est prélevé. Pour estimer la taille moyenne de la population étudiante, le groupe d'étudiants qui s'adonnent à la pratique du ballon panier constitue un mauvais échantillon. Si l'on veut estimer la prévalence de la cirrhose du foie dans la population générale, l'ensemble des patients d'une clinique de désintoxication est un mauvais échantillon. Pour se protéger contre le manque de représentativité d'un échantillon, on peut recourir à la sélection des sujets par tirage au sort.

Le tirage au sort évite les choix arbitraires, les préférences accordées à tel ou tel groupe d'individus de la population. Ce tirage permet de réduire les biais dans la sélection d'un échantillon. L'échantillonnage le plus susceptible de fournir un échantillon représentatif d'une population obéit au hasard. Ce paradoxe n'est toutefois qu'apparent, car hasard ne signifie ni désordre ou fouillis, ni anarchie ou fantaisie, contrairement à la conception populaire. Le prélèvement d'un échantillon, au moyen du tirage au sort, n'est pas un geste « à la bonne franquette ». Le hasard est soumis à des régularités, ce que nous apprend d'ailleurs la loi des grands nombres. Une des conséquences pratiques de cette loi est qu'à mesure qu'on augmente le nombre d'individus d'un échantillon aléatoire, la probabilité devient plus forte que cet échantillon soit une image fidèle de la population.

Ce concept de généralisation d'un point de vue statistique est en un sens limitatif. Dans les études cas-témoins, les cas ne sont pas forcément, au sens statistique, représentatifs de l'ensemble

des cas de la population, et il en va de même pour les témoins. Il serait très difficile de mettre en évidence certaines relations, de dégager certaines associations, particulièrement dans le cas de l'étude d'une maladie rare ou d'une exposition rare, s'il fallait satisfaire à l'exigence de la représentativité statistique. Dans ce genre d'études, les sujets disponibles sont ceux qui s'imposent d'emblée. Le problème de la représentativité statistique a ici moins d'importance. Dans les études épidémiologiques, souvent la question de la généralisation est posée d'un point de vue scientifique.

GÉNÉRALISATION D'UN POINT DE VUE SCIENTIFIQUE

D'un point de vue scientifique, la généralisation repose sur les connaissances que l'on a du phénomène étudié. Si on reconnaît aux caractéristiques considérées une sorte de permanence biologique, physiologique, physique entre les individus, il est possible de généraliser les résultats d'un groupe à un autre. Une étude a permis d'établir une association entre le fait de fumer et le cancer du poumon dans une population masculine. D'un point de vue scientifique, l'association est généralisable à une population féminine si on comprend que la physiologie du poumon est la même chez l'homme et chez la femme et que les mécanismes cancérigènes en cause sont indépendants du sexe. Par contre, pour une maladie associée à des facteurs hormonaux, il serait plus difficile de généraliser à une population féminine les résultats d'une étude menée auprès d'une population masculine.

Validité interne

Si l'organisation interne d'une étude entraîne une distorsion dans la mesure d'estimation, alors cette mesure manque de *validité interne.* C'est ce qui se produit, par exemple, lorsqu'on évalue le poids des personnes avec une balance défectueuse. Comme la balance bien calibrée, une étude à visée étiologique doit être exempte de défauts internes qui entraînent une distorsion de la mesure. Ces défauts peuvent se rencontrer à trois étapes du déroulement de l'étude: à l'étape de la sélection des sujets, à celle de la collecte de l'information et à celle de l'analyse. Suivant cette classification, on distingue trois grandes familles de biais (erreurs systématiques) : les biais de sélection, les biais d'information et les biais de comparaison (ou plus particulièrement les biais de confusion). La description de ces trois types de biais sur les mesures d'association fait l'objet du chapitre suivant.

RÉSUMÉ

La justesse d'une mesure peut être affectée par deux types d'erreur: l'erreur aléatoire et l'erreur systématique. La présence d'erreurs aléatoires traduit un manque de précision, celle d'erreurs systématiques reflète un manque de validité. La précision dépend de la taille des échantillons et de la variabilité du caractère étudié. La validité est externe ou interne. La validité externe concerne la généralisation des résultats et peut-être envisagée d'un point de vue statistique ou scientifique. Au plan statistique, la généralisation des résultats est possible si l'étude porte sur des échantillons obtenus par tirage au sort. Au plan scientifique, la généralisation d'un groupe à un autre

possible si l'on reconnaît, entre les individus de ces deux groupes, une sorte de permanence biologique, physiologique, physique des caractéristiques considérées. Le manque de validité interne résulte des défauts au plan de la sélection des sujets, de la collecte de l'information et de la comparabilité des groupes. Ces deux composantes de la justesse, validité et précision, conduisent aux trois derniers chapitres. Les chapitres 15 et 16 concernent la validité, le chapitre 17, la précision statistique.

LECTURES SUGGÉRÉES

1. KLEINBAUM, D., KUPPER, L.L. et MORGENSTEIN, H. ***Epidemiologic Research,*** Belmont (USA), Lifetime Learning Publications, 1982, chapitre 10, pp. 183-193.

2. RUMEAU-ROUQUETTE, C., BRÉART, G. et PADIEU, R. ***Méthodes en épidémiologie,*** Paris, Flammarion, 1985, chapitre IV, pp.42-43.

3. ROTHMAN, K.J. ***Modern Epidemiology,*** Boston, Little, Brown, 1986, chapitres 7 et 8, pp. 77-113.

CHAPITRE 15

Biais dans les mesures d'association *RR* et *RC*

Dans ce chapitre sont discutés les biais de sélection, d'information et de confusion, spécialement en rapport avec les mesures d'association *RR* et *RC*. Les biais sont des distorsions qui affectent les estimations des mesures. On tente d'identifier les sources responsables de ces biais, les moyens de les détecter et de les contrôler.

Dans ce chapitre, nous discutons des trois grands types de biais qui peuvent affecter les mesures *RR* et *RC*. Nous limitons la discussion à ces deux mesures pour deux raisons: d'abord parce qu'elles occupent une place importante dans la recherche étiologique et, en conséquence, sont largement diffusées dans la littérature, ensuite parce que leur utilisation dans la présentation permet une discussion assez étendue de la notion de biais.

Nous présentons donc les biais de sélection, d'information et de confusion et discutons de l'influence qu'ils peuvent avoir sur le *RR* dans les études de cohorte et sur le *RC* dans les études cas-témoins.

BIAIS DE SÉLECTION

Le biais de sélection dont il est question ici concerne la distorsion des résultats induite par une sélection préférentielle des sujets à comparer.

Définition du biais de sélection

Le *biais de sélection* est une distorsion dans l'estimation d'une mesure causée par les défauts de l'étude au niveau de la sélection des sujets. Il y a biais de sélection dans les études de cohorte si le recrutement des sujets exposés (des sujets non-exposés) est lié à la présence de la maladie, et dans les études cas-témoins, si le recrutement des cas (des témoins) est liée à la présence de l'exposition au facteur étudié. Voici quelques exemples

- Dans une étude de cohorte, on veut comparer la morbidité des travailleurs à celle de la population générale. Une telle comparaison souffre généralement d'un biais de sélection. En effet, la sélection des individus pour le marché du travail comprend comme critère important la « bonne santé ». L'absence de la maladie chez l'individu augmente ses chances d'accès au travail. Le groupe des travailleurs constitue un groupe d'individus d'abord sélectionnés pour des raisons de santé. L'absence de maladie a modifié chez l'individu ses chances de devenir travailleur et donc de faire partie de l'étude. On doit retenir qu'en général la santé des travailleurs est meilleure que celle de la population entière. Ce biais de sélection est connu dans la littérature anglophone sous l'appellation de « healthy worker effect », qu'on pourrait traduire par « effet de bonne santé ».
- Dans une étude de cohorte, il est possible que les exposés malades présentent moins de risque d'être perdus au suivi que ne le sont les exposés non-malades, les non-exposés malades et les non-exposés non-malades.
- Dans une étude cas-témoins, on veut comparer l'histoire des traumatismes accidentels chez des cas de spondylites arthritiques ankylosantes à un groupe témoin tiré de la population générale. En raison d'une surveillance médicale qu'entraîne le traumatisme, une spondylite arthritique peut avoir une plus forte probabilité d'être détectée chez un individu qui a subi un traumatisme. En conséquence, même sans
association entre le facteur et la maladie, on pourrait observer un plus grand nombre de cas que de témoins, ayant été affectés par le traumatisme. Une telle étude présenterait alors un biais de sélection.

Nous proposons dans l'annexe A de ce chapitre quelques expressions formelles qui caractérisent le biais de sélection, tant pour le rapport des cotes *RC* que pour le risque

relatif *RR*. Nous appuyons cette présentation d'exemples numériques.

Sources des biais de sélection

ÉTUDE DE COHORTE

Souvent l'état de santé ou la présence de problèmes particuliers influence la participation des sujets aux études. Les non-répondants peuvent comprendre des individus qui sont le moins soucieux de leur état de santé et des individus qui ne peuvent pas collaborer pour des raisons de santé. Les perdus-au-suivi sont souvent des sujets plus mobiles, plus jeunes, risquant moins d'être affectés par la maladie; d'autres ont pu déménager pour diverses raisons liées à leur santé. Le refus de participer est une autre source potentielle de biais de sélection.

ÉTUDE CAS-TÉMOINS

Dans les études cas-témoins, on peut identifier trois grandes sources de biais de sélection : le biais de survie sélective, le biais de détection et le biais d'admission.

- Le *biais de survie sélective* est celui qui, dans l'estimation du *RTi,* peut être induit par l'étude de cas prévalents. Si l'exposition modifie la durée de survie des cas ou des témoins, l'observation de données de prévalence conduira à une mesure biaisée du rapport des taux d'incidence *RTi.* Supposons, par exemple, que les exposés décèdent plus rapidement que les non-exposés, alors un groupe de cas prévalents comptera un plus grand nombre de non-exposés. En d'autres termes, il y aura sur-représentation des cas les plus longs, c'est-à-dire des non-exposés.
- Il y a *biais de détection si* le facteur, outre le lien qu'il puisse avoir avec la maladie, influence directement la détection de la maladie. C'est le cas, par exemple, du traumatisme accidentel (le facteur) qui entraîne une surveillance médicale assidue et conduit alors à une plus forte probabilité de détection de la spondylite arthritique.
- Une étude cas-témoins qui se déroule en milieu hospitalier peut présenter un *biais d'admission.* Ce biais, connu sous le nom de biais de Berkson, résulte de probabilités différentes d'être admis à l'hôpital pour les cas exposés et non-exposés et pour les témoins exposés et non-exposés. Les cas exposés sont peut-être plus facilement admis à l'hôpital ou encore les témoins sont peut-être hospitalisés à cause de la présence du facteur.

Détection des biais de sélection

On peut définir certaines conditions qui permettent de mieux détecter le biais de sélection. Par exemple, l'utilisation de plusieurs groupes témoins peut permettre de porter un jugement sur l'existence possible d'un tel biais dans les résultats. Si la mesure d'association est la même quel que soit le groupe témoin utilisé, la présence d'un biais de sélection est peu vraisemblable. Si, par ailleurs, cette mesure varie avec le groupe témoin utilisé pour la comparaison, on peut suspecter la présence de biais de sélection.

Contrôle des biais de sélection

Au plan théorique, le biais de sélection peut être contrôlé tant au niveau de l'échantillonnage qu'à celui de l'analyse.

AU NIVEAU DE L'ÉCHANTILLONNAGE

Le biais de sélection peut être contrôlé au niveau de l'échantillonnage à condition que l'investigateur soit averti des sources potentielles d'un tel biais. Dans les études de cohorte, l'investigateur s'informe de la qualité des listes qu'il désire utiliser comme cadre d'échantillonnage; il cherche à réduire à néant le nombre de perdus-au-suivi, de non-réponses ou de refus. Dans les études cas-témoins, le chercheur s'enquerra, par exemple, des processus diagnostiques des cas pour reconnaître une action possible du facteur dans ces processus.

AU NIVEAU DE L'ANALYSE

Le contrôle du biais de sélection au niveau de l'analyse est pratiquement impossible. Un investigateur peut difficilement, par exemple, connaître à priori les chances qu'a un cas exposé d'être sélectionné par rapport à celles d'un cas non-exposé; celles d'un témoin exposé par rapport à celles d'un témoin non-exposé. Cette difficulté s'explique d'autant plus que, souvent, on n'est pas en mesure d'identifier clairement la population visée par l'étude.

En conclusion, disons que le biais induit par des défauts de sélection affecte davantage les études cas-témoins et constitue pour la validité de celles-ci un des problèmes importants. D'une part, le recrutement des cas et des témoins peut être influencé par le facteur d'exposition puisque celui-ci est apparu avant que les sujets ne soient sélectionnés. D'autre part, le choix des témoins est toujours une source potentielle de biais de sélection: les témoins hospitalisés sont sujets au biais d'admission, les témoins tirés de la population sont sujets à l'auto-sélection.

Qu'il s'agisse d'études cas-témoins ou d'études de cohorte, il demeure très important d'essayer de reconnaître l'influence de ce biais sur la mesure. Le chercheur doit tenter de savoir s'il conduit à une surestimation ou à une sous-estimation de la mesure.

BIAIS D'INFORMATION

Le *biais d'information* relève d'erreurs de classement des sujets qui peuvent affecter aussi bien l'exposition que la maladie (ou d'autres variables). Un sujet malade peut être classifié comme non-malade, un sujet exposé comme non-exposé et vice versa. Une *erreur de classement* est généralement le résultat d'un instrument d'observation défectueux. Elle peut aussi résulter, au plan plus général, d'un cadre d'observation inadéquat. Avant d'aborder le problème du biais d'information pour les mesures d'association *RR* et *RC,* nous allons examiner comment les erreurs de classement peuvent influencer l'estimation d'une proportion.

Erreur de classement et estimation d'une mesure de fréquence

Nous devons rappeler que la validité d'un processus *Q* (questionnaire, appareil, test diagnostique ou autre procédé), destiné à classer les sujets pour une caractéristique, est généralement déterminée par les valeurs de sensibilité, de spécificité et de la fréquence de la caractéristique (exposition ou maladie). L'erreur de classement est donc directement tributaire de ces valeurs. Considérons d'abord le tableau 15-1,

qui présente les résultats (fictifs) d'un test diagnostique *T* administré à un groupe de 1000 sujets où la prévalence relative de la maladie est de 0,10 (10 %).

II y a ici absence totale d'erreur de classe- ment. Le test T a une sensibilité de 100 % et une spécificité de 100 % *(Sn = 1* et *Sp* = 1). Conformément aux résultats de la dernière ligne du tableau, la prévalence relative vraie est de 0,10 ($^{100}/_{1000}$). La prévalence relative estimée par le test, dont le résultat figure à la dernière colonne du tableau, est aussi de 0,10 ($^{10}/_{1000}$). Ces deux valeurs correspondent et l'estimation de la prévalence relative est non-biaisée.

Considérons maintenant le tableau 15-2 pour lequel il y a présence d'erreur de classement. Le test *T* a une sensibilité de 90 % et une spécificité de 70 % ($Sn = 0,90$ et $Sp = 0,70$). La prévalence relative estimée $\hat{P}r$ est cette fois de 0,36 ($^{360}/_{1000}$), alors que la prévalence relative vraie Pr est toujours de 0,10. La valeur estimée $\hat{P}r$ ne correspond plus à la valeur vraie Pr. L'estimation de la prévalence relative est biaisée.

Il peut y avoir erreur de classement, c'est-à-dire sensibilité ou spécificité différente de 100 %, sans que nécessairement l'estimation s'en trouve biaisée. Il suffit de considérer l'exemple suivant. Si la prévalence relative vraie est de 75 % *(Pr* = 0,75), la sensibilité de 90 % et la spécificité de 70 %, la prévalence relative estimée sera aussi de 75 % comme on peut le voir au tableau 15-3.

Pour qu'une erreur de classement conduise à une estimation biaisée, la condition d'une sensibilité et/ou d'une spécificité inférieure à 1 est nécessaire mais non-suffisante.

Formellement, l'estimation $\hat{P}r$ peut s'exprimer comme:

$$\hat{P}r = Pr \cdot Sn + (1 - Pr)(1 - Sp)$$

Il est facile de vérifier cette relation à l'aide des données du tableau 15-2 ou du tableau 15-3.

$$\hat{P}r = 0,10 \times 0,90 + (1 - 0,10)(1 - 0,70) = 0,36 \quad \text{(tableau 15-2)}$$

$$\hat{P}r = 0,75 \times 0,90 + (1 - 0,75)(1 - 0,70) = 0,75 \quad \text{(tableau 15-3)}$$

Tableau 15-1

		M +	M —	
T	+	100	0	100
	—	0	900	900
		100	900	1000

Tableau 15-2

		M +	M —	
T	+	90	270	360
	−	10	630	640
		100	900	1000

Tableau 15-3

		M +	M —	
T	+	675	75	750
	−	75	175	250
		750	250	1000

De la relation formelle précédente, on déduit assez facilement que l'erreur de classement conduit à une estimation biaisée si et seulement si:

$$\frac{1 - Sp}{1 - Sn} \neq \frac{Pr}{1 - Pr}.$$

Biais d'information pour les mesures d'association *RR* et *RC*

Le *biais d'information* pour les mesures d'association *RR* et *RC* est une distorsion de leur estimation causée par les erreurs de classement. Celles-ci peuvent être différentielles ou non-différentielles pour les groupes à comparer. Une erreur de classement est dite *non-différentielle si* le processus de classement a la même sensibilité et la même spécificité sur les deux groupes à comparer. Autrement l'erreur de classement est dite *différentielle.*

ERREUR DE CLASSEMENT NON-DIFFÉRENTIELLE

Il y a erreur de classement non-différentielle par rapport à la maladie lorsque, par exemple dans une étude de cohorte un test diagnostique pour la maladie considérée est administré avec les mêmes erreurs diagnostiques (même sensibilité, même spécificité), que les individus soient exposés ou non. Il y a erreur de classement non-différentielle par rapport à l'exposition lorsque, par exemple dans une étude cas-témoins, la sensibilité et la spécificité du questionnaire destiné à classer les individus selon qu'ils sont exposés ou non-exposés, sont les mêmes que les individus soient des cas ou des témoins. De façon générale, on peut démontrer que l'effet de l'erreur de classement non-différentielle est de forcer la mesure d'association, *RR ou RC,* vers la valeur de non-association, en l'occurrence la valeur 1.

ERREUR DE CLASSEMENT DIFFÉRENTIELLE

Dans une étude de cohorte, il peut y avoir erreur de classement différentielle par rapport à la maladie lorsque, par exemple, le test diagnostique administré aux sujets exposés est différent de celui administré aux sujets non-exposés. Il y a erreur de classement différentielle lorsque, par exemple, dans une étude cas-témoins, les cas, contrairement aux témoins, sont soigneusement investigués (avec peut-être même une certaine insistance) pour la présence du facteur. Une telle situation signifierait que le processus de classement aurait une meilleure sensibilité chez les cas que chez les témoins, sans qu'il y ait de différence pour la spécificité.

Il est possible d'imaginer des situations (voir l'annexe B à la fin du chapitre) qui montrent que le biais dû à une erreur de classement différentielle ne va pas toujours dans le même sens, qu'il y a tantôt surestimation, tantôt sous-estimation. Il est difficile d'établir des règles simples pour juger du sens que peut prendre le biais. Tout au plus pouvons-nous dire, pour les études cas-témoins, que si la sensibilité est plus forte mais la spécificité moins forte chez les cas que chez les témoins, alors le biais force l'estimation vers la valeur 1, c'est-à-dire vers la valeur de non-association. On peut faire le même énoncé pour les études de cohorte en changeant les termes « cas » et « témoins » respectivement pour « exposés » et « non-exposés ».

Sources des biais d'information

Les sources des biais d'information, nous le rappelons, sont les erreurs de classement induites par les processus ou instruments d'observation. Il peut s'agir, au plan général, d'un cadre d'observation défectueux. L'investigateur est influencé dans son observation par la connaissance qu'il a du sujet. La reconnaissance d'un cas fausse le jugement de l'observateur sur l'importance de l'exposition; la reconnaissance d'un sujet exposé influence la lecture d'une radiographie. Les cas sont plus soigneusement investigués que les témoins et, les sujets exposés davantage que les sujets non-exposés. Ce peut être encore un questionnaire mal construit, un instrument ou appareil mal calibré.

Contrôle des biais d'information

On peut contrôler le biais d'information au niveau de la planification de l'étude et au niveau de l'analyse.

AU NIVEAU DE LA PLANIFICATION

Pour contrôler le biais d'information au niveau de la planification, l'investigateur essaie de choisir des instruments d'observation ayant la plus forte validité intrinsèque possible (bonne sensibilité, bonne spécificité). Il raffine ses propres instruments (questionnaire, interview), définit un cadre rigoureux d'observation, utilise des techniques à simple ou double insu, quand c'est possible, pour ménager la susceptibilité de l'observateur et de l'observé.

Notons qu'un investigateur qui planifierait son étude en essayant d'égaliser les erreurs de classement entre les deux groupes à comparer n'est pas à l'abri du bais d'information. Il faut se rappeler que l'erreur de classement non-différentielle (ce dont il s'agit ici) conduit quand même à un biais d'information.

AU NIVEAU DE L'ANALYSE

La correction du biais d'information est rarement possible au niveau de l'analyse faute de connaître les valeurs correctes de sensibilité ou de spécificité des instruments d'observation. Par contre, même si on ne connaît pas l'ampleur du biais, il est toujours souhaitable pour le chercheur de savoir s'il s'agit d'un biais qui entraîne une sous-estimation ou une surestimation de la mesure.

BIAIS OU EFFET DE CONFUSION

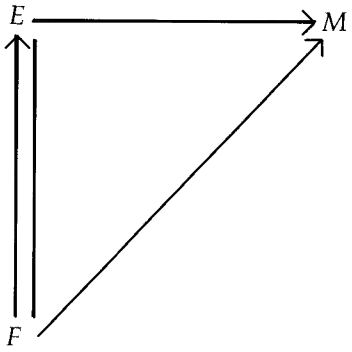
Le biais de confusion (ou simplement confusion) met en cause la présence d'un tiers facteur qui perturbe l'association entre un facteur et une maladie. Ce biais fait l'objet d'une préoccupation constante de la part de l'épidémiologiste. Après avoir défini ce biais, nous distinguerons les notions d'effet confondant et d'effet modifiant d'un facteur. Nous présenterons ensuite les deux sources principales de ce biais. Et enfin, nous discuterons de différents modes de détection et de contrôle de ce biais.

Définition du biais de confusion

Le *biais de confusion,* dans l'estimation d'une association entre un facteur E et une maladie *M*, peut-être défini comme la distorsion attribuable à un tiers facteur *F*, appelé alors

facteur confondant ou de confusion. Pour produire un biais, ce facteur *F* doit être associé, dans les données, au facteur *E* et, de façon indépendante, doit être associé *à M* comme facteur de risque. La figure 15-1 montre le type d'association qui doit lier le facteur *F* respectivement *à E* et *à M* pour qu'il soit confondant. La flèche qui va de *F à M* indique que *F* est un facteur de risque pour *M,* celle qui va de *F à E* indique que *F* est prédictif (ou concomitant) pour *E.* En résumé, pour que *F* soit facteur confondant, il faut qu'il soit facteur de risque pour *M* et qu'il soit facteur prédictif ou concomitant pour *E.*

Figure 15-1



L'âge est peut-être le facteur confondant le plus présent dans les études. Il est généralement associé à la maladie comme facteur de risque et se trouve associé, à tout le moins de façon concomitante, à une multitude de facteurs d'exposition (habitude de vie, occupation...). Vu son importance, nous utiliserons l'âge comme facteur confondant dans la plupart de nos exemples. Enfin, pour simplifier la présentation, nous demeurons à l'intérieur des limites suivantes:

- Il y a un seul facteur *E* en association avec une seule maladie *M.*
- Il n'y a qu'un seul tiers facteur *F.*
- Il n'y a pas d'autres biais que celui de confusion qui puisse affecter les résultats.
- Les facteurs *E et F* et la maladie *M* sont des variables dichotomiques. Rappelons que les catégories de *E* et *M* sont désignées conventionnellement par + et – , celles de *F* par F_1 et F_2. On peut parler de la strate F_1 et de la strate F_2. Les sujets de la catégorie *(E+)* sont les sujets exposés; ceux de la catégorie *(E —)* sont les sujets non-exposés.

Distinction entre confusion et modification

Une étude sur l'association entre le facteur *E* et la maladie *M* conduit aux résultats suivants. Pour les deux groupes ou strates d'âge considérés, 1 (40-49 ans) et 2 (50-59 ans), le risque relatif est 2: $RR_1 = 2$ et $RR_2 = 2$. Par contre, pris globalement, c'est-à-dire sans distinguer les groupes d'âge, le *RR* est 2,36. Cette situation est-elle possible? Pour s'en convaincre, on peut se reporter au tableau 15-7. Le risque relatif global brut 2,36 *(RR_b* = 2,36) n'est pas conforme à ce qui est décrit sur les groupes d'âge. Cette estimation, influencée par l'action de *F* dans les données, est ici une mesure biaisée. *F* est un facteur de confusion.

Considérons maintenant une autre situation, comme celle décrite au tableau 15-8. L'étude conduit cette fois aux résultats suivants :

$$RR_1 = 2,\ RR_2 = 3,\ RR_b = 3{,}5.$$

On observe deux faits distincts :

1) le risque relatif varie avec le groupe d'âge:

$$RR_1 = 2,\ RR_2 = 3.$$

2) le risque relatif brut est supérieur aux deux risques relatifs spécifiques:

$$3{,}5 > 3 > 2.$$

Le premier fait est l'indication que l'effet change avec le facteur *F*. Cela revient à dire que le risque relatif est ici plus élevé chez les personnes plus âgées. L'âge a modifié l'effet de *E* sur *M*. Le facteur *F* est ici un *facteur modifiant.* La modification se révèle dans l'hétérogénéité des mesures spécifiques.

Le deuxième fait est l'indication que le RR_b est une mesure biaisée par l'action du facteur *F* dans les données. Le facteur *F* est aussi, dans la situation présente, un facteur de confusion. Dans cet exemple, où les mesures spécifiques sont *2* et 3 et la mesure brute est 3,5, le facteur *F* est à la fois modifiant et confondant. Il est possible toutefois de rencontrer des situations où le facteur *F* est modifiant sans être confondant ou confondant sans être modifiant. S'il y a modification de l'effet, il ne s'ensuit pas obligatoirement qu'il y a confusion de l'effet et inversement. On trouve à l'annexe C de ce chapitre des exemples numériques qui permettent d'approfondir la distinction entre les effets de confusion et de modification.

La présence d'un facteur confondant entraîne un biais dans l'estimation d'une mesure globale brute d'association. La présence d'un facteur modifiant change les valeurs des mesures spécifiques d'association et reflète une interaction possible entre les facteurs *E* et *F*. Le problème de la confusion est moins pertinent s'il y a un effet de modification. Dans ce cas, pour mieux faire ressortir l'effet de modification, il est préférable de présenter les mesures spécifiques plutôt que la mesure globale brute. S'il y a une faible confusion, la mesure brute se situera dans l'intervalle déterminé par les mesures spécifiques. L'effet de confusion sur la mesure brute est alors difficilement perceptible, comme on peut le voir à l'annexe C de ce chapitre (tableaux C15-4 et C15-8). Par contre, s'il y a une forte confusion, la mesure brute se situera à l'extérieur de l'intervalle. L'effet de confusion est manifeste, comme on peut le voir à l'annexe C (tableau C15-3 et C15-7). Si, en dépit de la modification, on veut présenter la mesure brute d'association, il faudra alors corriger le biais de confusion qui l'affecte.

Sources de confusion

La confusion dans les données a deux sources principales: la confusion potentielle dans la population et la confusion induite au niveau de la sélection.

CONFUSION POTENTIELLE

Si, dans la structure même de la population où est conduite l'étude, le facteur *F* est confondant (mais non-modifiant), alors la confusion se manifestera en toute bonne probabilité dans les données de l'étude, à moins qu'elle ne soit contrôlée. Considérons, par exemple, la population décrite au tableau 15-4 où le facteur *F* est confondant pour l'association entre le facteur *E* et la maladie *M*. Un investigateur décide de

conduire une étude dans cette population à partir d'échantillons aléatoires de 3000 sujets exposés et de 3000 sujets non-exposés. Il pourrait obtenir en toute probabilité le tableau 15-5. L'effet de confusion du facteur *F* sur la mesure d'association s'est transmis de la population à l'étude. Dans l'étude, la valeur 2,4 du *RR* est une valeur biaisée.

Tableau 15-5: Étude

M	E +	E —
+	108	45
	3000	3000

$RR = 2{,}4$

Tableau 15-4: Population

Strate 1 *(F_1)*

M	*E* +	*E* —
+	400	500
	20 000	50 000

$RR_1 = 2$

Strate 2 *(F_2)*

M	*E* +	*E* —
+	3200	1000
	80 000	50 000

$RR_2 = 2$

Global

M	*E* +	*E* —
+	3600	1500
	100 000	100 000

$RR_b = 2{,}4$

Tableau 15-6: Population

Strate 1 *(F_1)*

M	*E* +	*E* —
+	400	800
	20 000	80 000

$RR_1 = 2$

Strate 2 *(F_2)*

M	*E* +	*E* —
+	640	1280
	16 000	64 000

$RR_2 = 2$

Global

M	*E* +	*E* —
+	1040	2080
	36 000	144 000

$RR_b = 2$

CONFUSION INDUITE

S'il existe une association réelle entre *F* et *M (F → M)* et si des défauts de sélection induisent artificiellement une association entre *E* et *F,* alors il en résultera dans les données une situation de confusion. Cette confusion est dite induite par la sélection. Considérons, par exemple, la population décrite au tableau 15-6. On remarque que le facteur *F* est non-confondant et que la distribution du facteur *F* est la même chez les sujets exposés et les sujets non-exposés ($^{20\ 000}/_{36000} = {}^{80\ 000}/_{144\ 000}$). Une étude est conduite dans cette population à partir d'échantillons de 1800 sujets exposés et 1800 sujets non-exposés. Les résultats de la sélection sont décrits au tableau 15-7.

La sélection a conduit à des distributions différentes du facteur *F* chez les sujets exposés et les sujets non-exposés ($^{1000}/_{1800} \neq {}^{1400}/_{1800}$). Elle a induit dans les données une association entre *E* et *F.* Et, puisque le facteur *F* est un facteur de risque pour la maladie (le risque change avec la valeur de *F:* $^{400}/_{20\ 000} \neq {}^{640}/_{16\ 000,}\ {}^{800}/_{80\ 000} \neq {}^{1280}/_{64\ 000}$), alors *F* est facteur confondant dans les données. La valeur 2,36 est une estimation biaisée du *RR.*

Détection d'un effet de confusion

La détection de F comme facteur confondant dans les données peut se faire de différentes façons. La plus immédiate, naturelle, est de comparer la mesure brute aux mesures spécifiques. Une autre façon consiste à comparer la mesure brute à une mesure ajustée, qui est en quelque sorte une somme pondérée des mesures spécifiques (notion approfondie au chapitre suivant). Enfin, on peut

Tableau 15-7: Étude

Strate 1 *(F_1)*

		E +	*E* —
M	+	20	14
	—	1000	1400

RR_1=2

Strate 2 *(F_2)*

		E +	*E* —
M	+	32	8
	—	800	400

RR_2=2

Global

		E +	*E* —
M	+	52	22
	—	1800	1800

RR_b = 2,36

détecter la confusion en vérifiant formellement les associations entre *F et E* et entre *F* et *M.* Nous allons examiner ces différents modes de détection en supposant que le facteur *F* est peu ou pas modifiant. Cette hypothèse permet de simplifier la discussion. De plus, comme on l'a souligné, le problème de la confusion perd de sa pertinence lorsque le facteur est fortement modifiant.

COMPARAISON DE LA MESURE BRUTE AUX MESURES SPÉCIFIQUES

Si la mesure brute ne se situe pas entre la plus petite et la plus grande des mesures spécifiques, alors il y a confusion. La mesure brute est une estimation biaisée. Si la mesure se situe entre la plus petite et la plus grande des mesures spécifiques, la confusion n'est pas évidente. En rapport avec ce critère d'une comparaison, nous suggérons deux exemples où il y a confusion, l'un d'une étude de cohorte, l'autre d'une étude cas-témoins.

Considérons l'étude de cohorte décrite au tableau 15-8. Le risque relatif brut RR_b (= 3,5) se situe à l'extérieur de l'intervalle $[RR_1, RR_2]$ ou [2, 3]. C'est donc l'indication de confusion dans les données. Le RR_b est une estimation biaisée.

Considérons l'étude cas-témoins décrite au tableau 15-9. Le rapport brut des cotes RC_b (= 3,45) se situe à l'extérieur de l'intervalle $[RC_1, RC_2]$ ou [2, 3]. C'est donc l'indication de

Tableau 15-8: Étude de cohorte

Strate 1 *(F_1)*

		E +	*E* —
M	+	20	20
	—	500	1000

$RR_1 = 2$

Strate 2 *(F_2)*

		E +	*E* —
M	+	120	20
	—	1000	500

$RR_2 = 3$

Global

		E +	*E* —
M	+	140	40
	—	1500	1500

$RR_b = 3,5$

Tableau 15-9: Étude cas-témoins

Strate 1 *(F1)*

	+	—	
Cas	40	60	100
Témoins	50	150	200

$RC_1 = 2$

Strate 2 *(F_2)*

	+	—	
Cas	150	50	200
Témoins	50	50	100

$RC_2 = 3$

Global

	+	—	
Cas	190	110	300
Témoins	100	200	300

$RC_b = 3,45$

confusion dans les données. Le RC_b est une estimation biaisée.

COMPARAISON DE LA MESURE BRUTE À UNE MESURE AJUSTÉE

Une mesure ajustée est une pondération des mesures spécifiques. Par conséquent, elle se situe entre la plus petite et la plus grande des mesures spécifiques. En l'absence de modification, la mesure ajustée ne diffère pas des mesures spécifiques, quel que soit le système de poids choisi. Ainsi, s'il n'y a pas de modification, une différence entre une mesure ajustée et une mesure globale brute indique un effet de confusion. Par ailleurs, s'il y a modification, le jugement sur la confusion est plus délicat. Il est relatif au type d'ajustement utilisé.

VÉRIFICATION FORMELLE DES ASSOCIATIONS

Rappelons que le facteur *F* est confondant s'il est *à* la fois associé *à E (F → E ou F — E)* et *à M (F → M).* On peut donc détecter la confusion en vérifiant formellement les associations entre *F* et *E* et entre *F et M,* dans les données. Il est possible de mesurer ces associations *à* l'aide du *RC.* Le *RC* est appliqué d'abord à l'association entre *F* et *E, puis à* l'association entre *F* et *M.* Si ces *RC* sont simultanément différents de 1, alors c'est l'indication que *F* est facteur de confusion. Le lecteur intéressé trouvera à l'annexe D de ce chapitre une présentation de la détection de la confusion par une vérification formelle des associations.

Notons enfin que les tests statistiques ne peuvent pas être utilisés pour détecter la confusion. Comparer une mesure brute à une mesure ajustée à l'aide d'un test statistique ne permet pas de reconnaître si un facteur est confondant. La confusion n'est pas affaire de statistique. Elle relève de la validité. Un risque relatif brut RR_b égal à 3 et un risque relatif ajusté RR_a égal à 2 peuvent être l'indication de confusion quelle que soit la taille des échantillons.

Contrôle des effets de confusion

Le contrôle de la confusion doit viser à éliminer la distorsion induite par le facteur *F* dans l'estimation de l'association. Formellement, il y a contrôle du facteur confondant si l'une au moins des deux associations, *(F → E)* et *(F→M),* est neutralisée. Ce contrôle peut être exercé à deux moments différents de l'étude: lors de l'échantillonnage ou lors de l'analyse. Examinons sommairement les moyens de contrôle, en distinguant les études de cohorte et les études cas-témoins.

Contrôle dans les études de cohorte (mesure *RR)*

Le contrôle de la confusion dans les études de cohorte (pour la mesure *RR)* peut être fait soit au niveau de l'échantillonnage, soit au niveau de l'analyse.

AU NIVEAU DE L'ÉCHANTILLONNAGE

Le contrôle au niveau de l'échantillonnage peut être fait par restriction ou par assortiment.

Par restriction
La variable confondante est restreinte à certaines catégories. Par exemple, pour contrôler la variable sexe, on restreint l'étude, disons, aux

femmes. Pour contrôler l'âge, on restreint l'étude à un groupe d'âge, disons 40-49 ans.

Par assortiment
L'assortiment vise à assigner au groupe des sujets exposés un groupe de sujets non-exposés ayant la même distribution du facteur *F* à contrôler. Par exemple, si 10 % des sujets exposés se trouvent dans le groupe d'âge 40-49 ans, on s'assurera que parmi les sujets non-exposés il s'en trouve aussi 10 % dans le groupe d'âge 40-49 ans. Une façon de faire est, par exemple, d'assigner à chaque sujet exposé un sujet non-exposé assorti pour la variable âge, en l'occurrence pour la variable à contrôler *F*. Ce type d'assortiment par paire (exposé—non-exposé) est un *appariement.* Pour l'âge, chaque sujet exposé est assorti (apparié) à un sujet non-exposé du même groupe d'âge. Le procédé d'assortiment garantit une même distribution de la variable confondante pour les deux groupes à comparer. Ce procédé permet de briser l'association entre les facteurs *F* et *E*.

Considérons un exemple qui illustre l'effet de l'assortiment dans le contrôle de la confusion. Le tableau 15-10 décrit une population dynamique où *F* est un facteur de confusion.

Tableau 15- 10: Population

Strate 1 (F_1)

		E +	*E* —
M	+	400	500
	—	20 000	50 000

RR_1=2

Strate 2 (F_2)

		E +	*E* —
M	+	3200	1000
	—	80 000	50 000

RR_2=2

Global

		E +	*E* —
M	+	3600	1500
	—	100 000	100 000

RR_b = 2,4

De cette population, on tire un échantillon de 10 000 sujets exposés que l'on assortit pour le facteur F à 10 000 sujets non-exposés. Ainsi, pour les 2000 et les 8000 sujets exposés qui se trouvent dans les catégories ou strates F_1 et F_2, on choisit 2000 et 8000 sujets non-exposés dans les catégories correspondantes. Ces résultats figurent au tableau 15-11. L'assortiment a éliminé l'effet de confusion du facteur *F*. L'estimation RR_b n'est pas biaisée : $RR_b = 2$.

Tableau 15-11: Étude

Strate 1 *(F_1)*

		E +	*E* —
M	+	40	20
	—	2000	2000

$RR_1 = 2$

Strate 2 *(F_2)*

		E +	*E* —
M	+	320	160
	—	8000	8000

$RR_2 = 2$

Global

		E +	*E* —
M	+	360	180
	—	10 000	10 000

$RR_b = 2$

AU NIVEAU DE L'ANALYSE

Le contrôle de la confusion au niveau de l'analyse peut être fait par stratification ou par modélisation.

Par stratification
Les comparaisons entre groupes de sujets exposés et de sujets non-exposés sont faites séparément sur chacune des catégories ou strates de la variable *F* à contrôler. Ces comparaisons conduisent à des mesures spécifiques d'association. Si, par exemple, la variable *F* est l'âge, on calculera le risque relatif dans chaque groupe d'âge considéré. Les mesures spécifiques d'association peuvent être par la suite résumées en une seule mesure globale à partir d'une somme pondérée (comme on le verra au chapitre suivant). Notons que, lorsqu'il y a assortiment dans une étude de cohorte, on pratique généralement l'analyse stratifiée, non pour le contrôle de la confusion, mais plutôt pour des raisons de puissance statistique.

Par modélisation
Le contrôle de la confusion peut être fait à partir de modèles statistiques de régression. La présentation de ces techniques dépasse la portée de ce manuel.

Contrôle dans les études cas-témoins (mesure RC)

Dans les études cas-témoins, le contrôle de la confusion se pratique aussi au niveau de l'échantillonnage, mais par restriction seulement. Il peut être également exercé au niveau de l'analyse.

AU NIVEAU DE L'ÉCHANTILLONNAGE

Dans les études cas-témoins, à la différence des études de cohorte, l'assortiment n'assure pas le plein contrôle de la confusion, comme en témoigne l'exemple suivant.

Considérons la population totale décrite au tableau 15-12. Dans cette population, le facteur *F* est un facteur confondant pour l'association entre E et M. On tire un échantillon de 1200 cas. Les 300 cas de la catégorie F_1 et les 900 de la catégorie F_2 sont assortis à 300 et 900 témoins des catégories correspondantes de *F*. On obtient alors les résultats décrits au tableau 15-13. Le *RC,* en dépit de l'assortiment, demeure une estimation biaisée: $RC = 1{,}89$.

Au niveau de l'échantillonnage, pour les études cas-témoins, la restriction demeure la seule façon de contrôler l'effet de confusion. L'assortiment est néanmoins pratiqué dans les études cas-témoins pour des raisons de puissance statistique ou encore pour faciliter l'analyse stratifiée.

AU NIVEAU DE L'ANALYSE

Au niveau de l'analyse, le contrôle de la confusion peut être fait, comme pour les études de cohorte, soit par stratification, soit par modélisation.

Tableau 15-12

Strate 1 (F_1)

	E +	E —	
Cas	1000	2000	3000
Non-cas	12 000	48 000	60 000

$RC_1 = 2$

Strate 2 (F_2)

	E +	E —	
Cas	6000	3000	9000
Non-cas	14 000	14 000	28 000

$RC_2 = 2$

Global

	E +	E —	
Cas	7000	5000	12 000
Non-cas	26 000	62 000	88 000

$RC_b = 3{,}34$

Tableau 15-13

	E +	E —	
Cas	700	500	1200
Témoins	510	690	1200

$RC = 1{,}89$

RÉSUMÉ

Les biais qui affectent les mesures d'association *RR* et *RC* ont été classés suivant trois grandes catégories: les biais de sélection, les biais d'information et les biais ou effets de confusion. Le biais de sélection est une distorsion de l'estimation d'une mesure causée par les défauts de l'étude au niveau de la sélection des sujets. Il y a biais de sélection dans les études de cohorte si la probabilité de sélection des sujets exposés (des sujets non-exposés) est liée à la présence de la maladie; dans les études cas-témoins, on trouve ce biais si la probabilité de sélection des cas (des témoins) est liée à la présence de l'exposition au facteur étudié. Le biais d'information relève d'erreurs de classement des sujets. Ces erreurs peuvent affecter l'exposition, la maladie (ou d'autres variables). Le biais d'information pour les mesures d'association *RR* et *RC* est une distorsion de leur estimation causée par les erreurs de classement. Une erreur de classement est non-différentielle si le processus de classement a la même sensibilité et la même spécificité sur les deux groupes à comparer, autrement, elle est différentielle. Le biais de confusion dans l'estimation d'une association entre un facteur E et une maladie M peut être défini comme la distorsion attribuable à un tiers facteur F, appelé facteur confondant ou de confusion. La distinction est faite entre confusion et modification. La présence d'un facteur confondant entraîne un biais dans l'estimation d'une mesure globale brute d'association. La présence d'un facteur modifiant change les valeurs des mesures spécifiques d'association.

Symboles

Pr, $\hat{P}r$: prévalence relative vraie, estimation de la prévalence relative

Sn, Sp: sensibilité, spécificité

RR_b, RC_b: risque relatif brut, rapport des cotes brut

Formule

$$\hat{P}r = Pr \cdot Sn + (1 - Pr)(1 - Sp)$$

LECTURES SUGGÉRÉES

1. GOLDBERG, M. ***L'Épidémiologie sans peine,*** Paris, Éditions médicales Roland Bettex, 1986, pp.73-78, 135-141.
2. KLEINBAUM, D., KUPPER L.L. et MORGENSTERN, H. ***Epidemiologic Research,*** Lifetime Learning Publications, Belmont (USA), 1982, pp. 194-265.
3. PHILIPPE, P. ***Épidémiologie pratique,*** Montréal, Presses de l'Université de Montréal, 1985, pp. 91-108.
4. ROTHMAN, K.J. ***Modern Epidemiology,*** Boston, Little, Brown, 1986, pp. 77-113.
5. RUMEAU-ROUQUETTE, C., BREART G. et PADIEU, R. ***Méthodes en épidémiologie,*** Paris, Flammarion, 1985, pp.119-135.
6. ScHLESSELMAN, *J.J.* ***Case-Control Studies,*** Oxford University Press, Oxford, 1982, pp. 124-143.

ANNEXES DU CHAPITRE 15

A — Expressions formelles qui caractérisent le biais de sélection

B — Illustrations numériques des erreurs de classement non différentielle et différentielle

C — Exemples numériques pour distinguer les effets de confusion et de modification

D — Détection de la confusion par vérification formelle des associations

A — EXPRESSIONS FORMELLES QUI CARACTÉRISENT LE BIAIS DE SÉLECTION

Pour mieux concrétiser cette notion de biais de sélection, nous allons nous référer à des situations numériques pour le calcul du rapport des cotes *RC* dans les études cas-témoins.

Un investigateur veut conduire, dans une population, une étude cas-témoins sur l'association possible entre le facteur *E* et la maladie *M.* Nous connaissons, à l'insu de l'investigateur, les probabilités d'être sélectionné pour un cas exposé, un cas non-exposé, un témoin exposé et un témoin non-exposé. Désignons ces probabilités dans l'ordre par les symboles, α, β, -y, et δ.

α = Prob (être sélectionné pour un cas exposé)

β = Prob (être sélectionné pour un cas non-exposé)

y = Prob (être sélectionné pour un non-cas exposé)

δ = Prob (être sélectionné pour un non-cas non-exposé)

Ces quatre probabilités de sélection peuvent influencer l'estimation du *RC.*

Nous examinons maintenant trois études cas-témoins qui conduisent, une à une surestimation du *RC,* une autre à l'absence de biais et une dernière à une sous-estimation. Chacune des études cas-témoins réfère à des valeurs particulières des probabilités de sélection α, β, y et δ. Ces études sont toutes menées dans une population de 50 000 individus, où sont présents le facteur *E* et la maladie *M.* Nous supposons que, dans cette population, le facteur *E* n'est pas associé à *M,* la prévalence relative de la maladie *M* est de 1 % et celle du facteur *E* de 10 %. Cette population est décrite au tableau A15-1.

Tableau A15-1 : Population

		E +	E −	
M	+	50	450	500
	−	4950	44 550	49 500
		5000	45 000	50 000

$$RC = \frac{50 \times 44\,550}{450 \times 4950} = 1.$$

Surestimation du RC

Nous savons qu'un cas exposé a 90 % des chances d'être sélectionné (α = 0,90), alors que ces chances sont de 70 % pour un cas non-exposé (β = 0,70), ce qui veut dire qu'un cas exposé a 1,29 fois plus de chances d'être sélectionné pour l'étude qu'un cas nonexposé: α/β = 0,90/0,70 = 1,29. Nous savons aussi que les chances de sélection sont égales entre les témoins exposés (y) et les témoins non-exposés (δ): -y = δ. Nous résumons ces probabilités de sélection dans le tableau A15-2.

Tableau A15-2

	E +	*E* −
Cas	α = 0,90	β = 0,70
Témoins	γ	δ

Pour mener son étude cas-témoins, l'investigateur décide de retenir tous les cas qu'il peut trouver et de choisir un groupe témoin de taille équivalente à celle du groupe des cas. II a réussi à recruter 360 cas. De ceux-ci et à son insu, 45 sont des cas exposés (0,90 x 50) et

315 des cas non-exposés (0,70 x 450). L'investigateur choisit 360 témoins, ce qui correspond à une probabilité de $^{360}/_{49500}$, soit $^{2}/_{275}$, pour un non-malade d'être choisi comme témoin. Cette probabilité de sélection est la même, que le non-malade soit exposé ou non, ($y = \delta = {}^{2}/_{275}$). Ainsi, toujours à l'insu de l'investigateur, 36 témoins exposés ($^{2}/_{275}$ X 4950) et 324 témoins non-exposés ($^{2}/_{275}$ X 44 550) sont obtenus.

L'étude cas-témoins ainsi générée, décrite au tableau A15-3, fournit maintenant une estimation du *RC* de 1,29, ce qui ne représente sûrement pas la réalité. La sélection a conduit à un échantillon de cas où la proportion d'exposés est artificiellement élevée. Il y a ainsi surestimation du rapport des cotes *RC*.

Absence de biais du *RC*

Nous savons qu'un cas exposé a 90 % des chances d'être sélectionné ($\alpha = 0{,}90$), alors que ces chances sont de 70 % pour un cas non-exposé ($\beta = 0{,}70$). Contrairement à la situation précédente où y était égal à δ, cette fois un témoin exposé a aussi 1,29 fois plus de chance d'être sélectionné qu'un témoin non-exposé: $y/\delta = \alpha/\beta = 1{,}29$. Pour maintenir les groupes équilibrés (ce n'est pas obligatoire), même nombre de cas et de témoins, il faut prendre $y = {}^{9}/_{990}$ et $\delta = {}^{7}/_{990}$. En effet,

$$\frac{9}{990} \times 4950 + \frac{7}{990} \times 44\,550 = 360.$$

L'étude cas-témoins générée par l'application de ces probabilités de sélection est décrite au tableau A15-4. On obtient cette fois une estimation non-biaisée du *RC*.

Même sans la contrainte de définir des groupes équilibrés, si les rapports α/β et y/δ sont égaux, on trouvera encore une estimation non-biaisée du *RC*.

Sous-estimation du RC

Nous savons qu'un cas exposé a 90 % des chances d'être sélectionné ($\alpha = 0{,}90$), alors que ces chances sont de 70 % pour un cas non-exposé ($\beta = 0{,}70$), ce qui veut dire qu'un cas exposé à 1,29 fois plus de chances d'être sélectionné pour l'étude qu'un cas non-exposé: $\alpha/\beta = 0{,}90/0{,}70 = 1{,}29$. Nous savons de plus que le rapport des chances de sélection entre les témoins exposés (y) et non-exposés (δ) est ici égal à 11, ce qui veut dire qu'un témoin exposé a 11 fois plus de chances d'être sélectionné pour l'étude qu'un témoin

Tableau A15-3

	E +	E —	
Cas	45	315	360
Témoins	36	324	360

$$RC = \frac{45 \times 324}{36 \times 315} = 1{,}29.$$

Tableau A15-4

	E +	E —	
Cas	45	315	360
Témoins	45	315	360

$$RC = \frac{45 \times 315}{45 \times 315} = 1.$$

non-exposé. Pour maintenir les groupes équilibrés, même nombre de cas et de témoins, il faut prendre $y = {}^{1}/_{25}$ et $\delta = {}^{1}/_{275}$. L'étude cas-témoins générée par l'application de ces probabilités de sélection est décrite au tableau A15-5. On obtient cette fois une estimation du *RC* de 0,12. La sélection a conduit à une sous-estimation du *RC*.

Dans deux des trois situations décrites ci-dessus, les rapports α/β et y/δ sont différents. Pour celles-ci, on a observé un biais dans l'estimation du *RC*. Là où les deux rapports sont égaux, le *RC* a été estimé sans biais. De façon formelle, on peut démontrer aisément qu'il y a biais de sélection du *RC* dans les études cas-témoins si les rapports α/β et y/δ sont différents, c'est-à-dire si $\alpha\delta/\beta y$ 1. De plus, il est facile de démontrer qu'il y a surestimation si ce rapport est supérieur à 1 et sous-estimation s'il est inférieur à 1.

$$\begin{array}{rcl} & < & 1 \quad \text{sous-estimation} \\ \alpha\delta/\beta y = & 1 & \text{absence de biais} \\ & > & 1 \quad \text{surestimation} \end{array}$$

Notons que pour le *RC,* dans les études de cohorte, il y a biais de sélection si les rapports α/y et β/δ sont différents, c'est-à-dire encore ici si $\alpha\delta/\beta y \neq 1$.

Tableau A15-5

	E		
	+	−	
Cas	45	315	360
Témoins	198	162	360

$$RC = \frac{45 \times 162}{198 \times 315} = 0{,}12.$$

Pour l'estimation du risque relatif *RR* dans les études de cohorte, les conditions sont analogues à celles qui s'appliquent aux calculs du *RC,* si on remplace les probabilités de sélection y et δ par les probabilités marginales de sélection: v_1 et v_o. La probabilité v_1 représente la probabilité pour un sujet exposé d'être sélectionné, alors que v_o représente celle pour un sujet non-exposé. On trouve alors les conditions suivantes:

$$\begin{array}{rcl} & < & 1 \quad \text{sous-estimation} \\ \alpha v_o/\beta v_1 & = & 1 \quad \text{absence de biais} \\ & > & 1 \quad \text{surestimation} \end{array}$$

Notons que si le *RR* est le rapport de taux d'incidence mesurés dans une population dynamique, alors la valeur $\alpha v_o/\beta v_1$ est équivalente à celle de $\alpha\delta/\beta y$ pour le *RC* d'une étude cas-témoins conduite sur les cas incidents de cette même population. Si, par contre, le *RR* est le rapport d'incidences cumulatives, il n'existe pas de lien direct entre les valeurs $\alpha v_o/\beta v_1$ et $\alpha\delta/\beta y$. Cela s'explique par le fait que les valeurs v_1, et v_o sont les sommes pondérées respectivement de α et y, β et δ:

$$v_1 = \frac{A}{N_1}\alpha + \frac{C}{N_1}\gamma;$$

$$v_0 = \frac{B}{N_0}\beta + \frac{D}{N_0}\delta;$$

où A, B, C, D désignent, dans la population, respectivement et suivant la convention déjà établie, les malades exposés, les malades non-exposés, etc., et N_1 et N_o le total des sujets exposés et des sujets non-exposés.

B — ILLUSTRATIONS NUMÉRIQUES DES ERREURS DE CLASSEMENT NON-DIFFÉRENTIELLE ET DIFFÉRENTIELLE

Pour mieux comprendre le problème de ce biais d'information, nous allons décrire quelques exemples numériques. Dans un souci de simplification, ces exemples sont présentés à l'intérieur des limites suivantes:

— Le problème d'erreur de classement n'affecte que le critère de comparaison, la maladie, dans les études de cohorte, ou l'exposition, dans les études cas-témoins.

— Le critère de comparaison est dichotomique: malade, non-malade dans les études de cohorte; exposé, non-exposé dans les études cas-témoins.

— Le biais d'information est le seul qui puisse affecter les résultats.

Erreur de classement non-différentielle

ÉTUDE DE COHORTE

Considérons une étude de cohorte où un groupe de 1000 sujets exposés à E et un groupe de 1000 sujets non-exposés sont comparés pour la fréquence relative de la maladie M. Supposons que la fréquence de la maladie chez les sujets exposés est de 40 par 1000, alors qu'elle est de 20 par 1000 chez les sujets non-exposés. Alors, la situation réelle se présente comme celle décrite au tableau B15-1. La valeur réelle du risque relatif *RR* est de 2.

Supposons que le diagnostic de la maladie est fait à partir d'un test dont la sensibilité est de 0,90 et la spécificité est de 0,70. L'estimation de la fréquence relative de la maladie se trouve donc modifiée tant chez les sujets exposés que chez les sujets non-exposés. Les tableaux B15-2A et B décrivent alors l'erreur de classement induite. Celle-ci modifie les prévalences relatives chez les sujets exposés et les sujets non-exposés. Elles deviennent respectivement $^{324}/_{1000}$ et $^{312}/_{1000}$. À la place du tableau B15-1, on obtient le tableau B15-3. Dans ce cas, le *RR* est de 1,04. On remarque que la mesure *RR* s'est rapprochée de la valeur 1, qui correspond à la non-association.

Tableau B15-1

		E +	E −	
M	+	40	20	$RR = \frac{40/1000}{20/1000} = 2$
		1000	1000	

Tableau B15-2A

		Exposés M +	Exposés M −	
T	+	36	288	324
	−	4	672	676
		40	960	1000

Tableau B15-2B

		Non-exposés M +	Non-exposés M −	
T	+	18	294	312
	−	2	686	688
		20	980	1000

Tableau B15-3

		E +	E —
M	+	324	312
	—		
		1000	1000

$$RR = \frac{324/1000}{312/1000} = 1{,}04$$

ÉTUDE CAS-TÉMOINS

Considérons une étude cas-témoins où 200 cas sont comparés à 200 témoins pour le facteur *E*. Le tableau B15-4 décrit les résultats de l'étude faite sans biais d'information.

Maintenant, supposons que l'étude soit faite à partir d'un questionnaire Q qui permet de bien classer 80 % des sujets exposés et 90 % des sujets non-exposés. Cette situation, qui conduit à une modification de la fréquence relative des exposés tant chez les cas que chez les témoins, est décrite aux tableaux B15-5A et B. L'erreur de classement modifie les proportions d'exposés chez les cas et chez les témoins. Elles deviennent respectivement de $^{90}/_{200}$ et de $^{55}/_{200}$. A la place du tableau B15-4, on obtient le tableau B15-6. Le *RC* est alors égal à 2,16. Encore ici, la mesure *RC* s'est rapprochée de la valeur 1, qui correspond à la non-association.

Tableau B15-4

	M +	M —	
Cas	100	100	200
Témoins	50	150	200

$$RC = \frac{100 \times 150}{50 \times 100} = 3$$

Tableau B15-5A

		Cas E +	Cas E —	
Q	+	80	10	90
	—	20	90	110
		100	100	200

Tableau B15-5B

		Témoins E +	Témoins E –	
Q	+	40	15	55
	–	10	135	145
		50	150	200

Tableau B15-6

	M +	M —	
Cas	90	110	200
Témoins	55	145	200

$$RC = \frac{90 \times 145}{55 \times 110} = 2{,}16$$

Erreur de classement différentielle

Considérons l'exemple d'une étude cas-témoins où encore ici 200 cas sont comparés à 200 témoins pour le facteur *E*. Le tableau B15-7 décrit les résultats de l'étude conduite sans biais d'information.

Maintenant, supposons qu'il y ait erreur de classement différentielle, par exemple que le processus de classification Q est sensible à 0,90 et spécifique à 0,60 chez les cas, alors qu'il est sensible à 0,80 et spécifique à 0,90 chez les témoins. La situation pour chacun des deux groupes pourrait se présenter comme aux tableaux B15-8A et B. L'erreur de classement modifie le tableau B15-7 pour donner le tableau B15-9. Dans ce cas, le *RC* est égal à 4,90. Ici, l'erreur de classement différentielle entraîne une surestimation de la valeur du rapport des cotes *RC*.

Considérons une autre étude cas-témoins, du même type que la première, sauf que maintenant la condition maladie rend difficile l'observation de l'exposition, d'autant plus que cette exposition réfère à une habitude tabou (drogue, alcool...). Les sujets tendent à sous-déclarer leur consommation et, davantage les cas que les témoins. Cela signifie que la sensibilité est plus faible chez les cas que chez les témoins. Supposons, par ailleurs, une même spécificité pour les deux groupes. Pour concrétiser, disons une sensibilité de 0,70 et une spécificité de 0,90 chez les cas, une sensibilité de 0,80 et une spécificité de 0,90 chez les témoins. Alors, le tableau B15-10 (qui reprend le tableau B15-7) présente la situation initiale pour laquelle le *RC* est égal à 3. Les tableaux B15-11A et B décrivent les altérations induites par l'erreur de classement. Enfin, le tableau B15-12 reprend le tableau B15-10 modifié par l'erreur de classement.

Tableau B15-8A

		Cas E +	Cas E −	
Q	+	90	40	130
	−	10	60	70
		100	100	200

Tableau B15-8B

		Témoins E +	Témoins E —	
Q	+	40	15	55
	—	10	135	145
		50	150	200

Tableau B15-9

	M +	M —	
Cas	130	70	200
Témoins	55	145	200

$$RC = \frac{130 \times 145}{55 \times 70} = 4{,}90$$

Tableau B15-7

	M +	M —	
Cas	100	100	200
Témoins	50	150	200

$$RC = \frac{100 \times 150}{50 \times 100} = 3$$

Tableau B15-10

	M +	M —	
Cas	100	100	200
Témoins	50	150	200

$$RC = \frac{100 \times 150}{50 \times 100} = 3$$

Le *RC* calculé dans le tableau B15-12 est maintenant de 1,76. On observe que l'erreur de classement entraîne ici une sous-estimation du *RC*.

Tableau B15-11A

		Cas E +	Cas E —	
Q	+	70	10	80
	–	30	90	120
		100	100	200

Tableau B15-11B

		Témoins *E* +	Témoins *E* —	
Q	+	40	15	55
	—	10	135	145
		50	150	200

Tableau B15-12

	M +	M —	
Cas	*80*	120	200
Témoins	55	145	200

$$RC = \frac{80 \times 145}{55 \times 120} = 1{,}76$$

C — EXEMPLES NUMÉRIQUES POUR DISTINGUER LES EFFETS DE CONFUSION ET DE MODIFICATION

Nous simulons d'abord quatre situations numériques pour les études de cohorte, ensuite quatre autres pour les études cas-témoins.

Études de cohorte

Tableau C15-1: Ni confusion, ni modification

Strate 1 *(F1)*

		E +	*E* —	
M	+	20	20	
		500	1000	

$RR_1 = 2$

Strate 2 *(F_2)*

		E +	*E* —	
M	+	40	40	
		500	1000	

$RR_2 = 2$

Global

		E +	*E* —	
M	+	60	60	
		1000	2000	

$RR_b = 2$

Tableau C15-2 : Confusion mais pas de modification

Strate 1 (F_1)

M \ E	+	−
+	8	10
−	400	1000

$RR_1 = 2$

Strate 2 (F_2)

M \ E	+	−
+	64	20
−	1600	1000

$RR_2 = 2$

Global

M \ E	+	−
+	72	30
−	2000	2000

$RR_b = 2{,}4$

Tableau C15-3 : Confusion et modification

Strate 1 (F_1)

M \ E	+	−
+	20	20
−	500	1000

$RR_1 = 2$

Strate 2 (F_2)

M \ E	+	−
+	120	20
−	1000	500

$RR_2 = 3$

Global

M \ E	+	−
+	140	40
−	1500	1500

$RR_b = 3{,}5$

Tableau C15-4 : Modification mais confusion peu pertinente

Strate 1 (F_1)

		E +	E −
M	+	20	8
	−	500	400

$RR_1 = 2$

Strate 2 (F_2)

		E +	E −
M	+	60	24
	−	500	600

$RR_2 = 3$

Global

		E +	E −
M	+	80	32
	−	1000	1000

$RR_b = 2{,}5$

Études cas-témoins

Tableau C15-5 : Ni confusion ni modificatio

Strate 1 (F_1)

	E +	E −	
Cas	50	50	100
Témoins	50	100	150

$RC_1 = 2$

Strate 2 (F_2)

	E +	E −	
Cas	100	100	200
Témoins	50	100	150

$RC_2 = 2$

Global

	E +	E −	
Cas	150	150	300
Témoins	100	200	300

$RC_b = 2$

Tableau C15-6 : Confusion mais pas de modification

Strate 1 (F_1)

	E +	E −	
Cas	24	56	80
Témoins	36	168	204

$RC_1 = 2$

Strate 2 (F_2)

	E +	E −	
Cas	200	20	220
Témoins	80	16	96

$RC_2 = 2$

Global

	E +	E −	
Cas	224	76	300
Témoins	116	184	300

$RC_b = 4{,}68$

Tableau C15-7 : Confusion et modification

Strate 1 (F_1)

	E +	E −	
Cas	40	60	100
Témoins	50	150	200

$RC_1 = 2$

Strate 2 (F_2)

	E +	E −	
Cas	150	50	200
Témoins	50	50	100

$RC_2 = 3$

Global

	E +	E −	
Cas	190	110	300
Témoins	100	200	300

$RC_b = 3{,}45$

Tableau C15-8 : Modification mais confusion peu pertinente

Strate 1 (F_1)

	E +	E −	
Cas	50	50	100
Témoins	50	100	150

$RC_1 = 2$

Strate 2 (F_2)

	E +	E −	
Cas	150	50	200
Témoins	75	75	150

$RC_2 = 3$

Global

	E +	E −	
Cas	200	100	300
Témoins	125	175	300

$RC_b = 2{,}80$

D — DÉTECTION DE LA CONFUSION PAR VÉRIFICATION FORMELLE DES ASSOCIATIONS

Considérons séparément les études de cohorte et les études cas-témoins pour lesquelles nous allons préciser le terme association.

Études de cohorte

Dans les études de cohorte, l'association *(F→ E)* veut dire que la distribution de *F* chez les sujets exposés est différente de celle chez les sujets non-exposés. En termes numériques, on peut dire que le *RC* calculé entre ces deux variables, *F* et E, est différent de 1 : $RC_{FE} \neq 1$.

Pour les études d'incidence cumulative, l'association *(F→ M)* veut dire que, parmi les sujets non-exposés, la distribution de *F* chez les malades est différente de celle chez les non-malades.

Pour les études de taux d'incidence, l'association *(F→ M)* signifie que, parmi les sujets non-exposés, la distribution de *F* chez les malades est différente de celle chez les personnes-années à risque.

En termes numériques, on peut dire que le *RC* calculé entre *F* et *M* chez les sujets non-exposés est différent de 1 :

$$RC_{FM/E-} \neq 1.$$

Quand *F* est non modifiant, on peut tout aussi bien écrire pour les exposés :

$$RC_{FM/E+} \neq 1.$$

Études cas-témoins

Dans les études cas-témoins, l'association *(F→M)* veut dire que, parmi les sujets non-exposés, la distribution de *F* chez les cas est différente de celle chez les témoins:

$$RC_{FM/E} — \neq 1.$$

L'association *(F→E)* veut dire, que parmi les témoins, la distribution de *F* chez les sujets exposés est différente de celle chez les sujets non-exposés:

$$RC_{FM/E} — \neq\neq 1.$$

Quand *F* est non modifiant, on peut tout aussi bien écrire pour les cas:

$$RC_{FM/E} — \neq 1.$$

En résumé, dans les études de cohorte et les études cas-témoins, un facteur *F* non modifiant est confondant si on a les relations ou conditions suivantes:

Études de cohorte

- $RC_{FE} \neq 1$
- $RC_{FM/E} — \neq 1.$

Études cas-témoins

- $RC_{FE/M} — \neq 1.$
- $RC_{FM/E} — \neq 1.$

Considérons deux exemples, l'un d'une étude de cohorte, l'autre d'une étude cas-témoins.

EXEMPLE DANS UNE ÉTUDE DE COHORTE

L'étude de cohorte est celle décrite au tableau C15-2 et reproduite au tableau D15-1. On note que, suivant le critère de la comparaison de la mesure brute aux mesures spécifiques, le facteur F est confondant mais non modifiant. Nous allons vérifier que les deux conditions $RC_{FE} — \neq 1$ *et* $RC_{FM/E} — \neq 1$ sont satisfaites.

Tableau D15-1

Strate 1 (F_1)

		E +	E −
M	+	8	10
	−	400	1000

$RR_1 = 2$

Strate 2 (F_2)

		E +	E —
M	+	64	20
	−	1600	1000

$RR_2 = 2$

Global

		E +	E —
M	+	72	30
	−	2000	2000

$RR_b = 2{,}4$

Le tableau D15-2A décrit la distribution du facteur F chez les sujets exposés et les sujets non-exposés. La valeur 0,25 du RC_{FE} indique une association entre *F et E.* Le tableau D15-2B décrit, pour les sujets non-exposés, la distribu-

tion de *F* chez les malades et sur le total (personnes ou personnes-années). La valeur 0,50 du $RC_{FM/E-}$ indique une association entre *F* et *M* chez les sujets non-exposés.

EXEMPLE DANS UNE ÉTUDE CAS-TÉMOINS

L'étude cas-témoins est celle décrite au tableau C15-6 et reproduite au tableau D15-3. On note que, suivant le critère de la comparaison de la mesure brute aux mesures spécifiques, il y a confusion mais non modification. Nous allons vérifier que les deux conditions

$$RC_{FE/M-} \neq 1 \text{ et } RC_{FE/M-} \neq 1$$

sont satisfaites. Le tableau D15-4A décrit pour les témoins la distribution du facteur *F* chez les sujets exposés et les sujets non-exposés. La valeur 0,04 du $RC_{FE/M-}$ indique chez les témoins une association entre *F* et E. Le tableau D15-4B décrit pour les sujets non-exposés la distribution de *F* chez les cas et chez les témoins. La valeur 0,27 du $RC_{FM/E-}$ indique chez les sujets non-exposés une association entre *F* et *M*. Puisque

$$RC_{FE/M-} \neq 1 \text{ et } RC_{FM/E-} \neq 1,$$

le facteur *F* est confondant.

Tableau D15-3

Strate 1 (F_1)

	E +	E —	
Cas	24	56	80
Témoins	36	168	204
		$RC_1 = 2$	

Strate 2 (F_2)

	E +	E —	
Cas	200	20	220
Témoins	80	16	96
		$RC_2 = 2$	

Global

	E +	E —	
Cas	224	76	300
Témoins	116	184	300
		$RC_b = 4{,}68$	

Tableau D15-2A

		E +	E –
F	1	400	1000
	2	1600	1000
		$RC_{FE} = 0{,}25$	

Tableau D15-2B

		M+	Total
F	1	10	1000
	2	20	1000
		$RC_{FM/E-} = 0{,}50$	

Tableau D15-4A

		E +	E —
F	1	36	168
	2	80	16
		$RC_{FE/M—} = 0{,}04$	

Tableau D15-4B

		Cas	Témoins
F	1	56	168
	2	20	16
		$RC_{FM/E-} = 0{,}27$	

CHAPITRE 16

Ajustement des mesures

Ce chapitre porte sur la question de l'ajustement des mesures de fréquence, d'association et d'impact. On discute d'abord du problème de l'ajustement lorsqu'il s'agit de comparer deux mesures de fréquence. Plus spécifiquement, on aborde la question du choix d'une population-type. Ensuite, on décrit différents types d'ajustement des mesures d'association *RR* et *RC*. Ces ajustements sont regroupés suivant trois pratiques: poids définis à partir d'une distribution-type, poids définis à partir de la précision des estimations, poids de Mantel-Haenszel. Enfin, on présente l'ajustement des deux mesures d'impact: la fraction étiologique et la fraction prévenue. On insiste sur la notion de cohérence dans l'interprétation de mesures d'impact ajustées.

L'ajustement réfère à l'idée d'adapter un ensemble de mesures spécifiques à un système de poids en vue de résumer ces mesures spécifiques. S'il s'agit de mesures spécifiques à l'âge, on parle d'ajustement pour l'âge, s'il s'agit du sexe, d'ajustement pour le sexe, etc. Par exemple, si pour les trois catégories d'une variable on a les mesures spécifiques et le système de poids tels qu'ils sont décrits au tableau 16-1, alors l'expression

$$0{,}500 \times 0{,}005 + 0{,}300 \times 0{,}010 + 0{,}200 \times 0{,}015$$

décrit le procédé d'ajustement de ces trois mesures spécifiques sur le système de poids. De façon générale, si (m_i) est un ensemble de mesures spécifiques et (λi) un système de poids, $i = 1, ..., k$, alors $\Sigma\ \lambda_i m_i$ décrit l'ajustement de ces mesures sur le système de poids.

Tableau 16-1

	Catégorie		
	1	2	3
Mesures (m)	0,005	0,010	0,015
Poids (λ)	0,500	0,300	0,200

Conformément à cette définition de l'ajustement, la mesure qui résume ainsi les mesures spécifiques sera appelée *mesure ajustée* (sur le système de poids utilisé). Qu'il s'agisse de mesures de fréquence, de mesures d'association ou de mesures d'impact, le concept d'ajustement demeure le même: celui de choisir un système de poids et de l'utiliser pour résumer les mesures spécifiques en une mesure globale.

Au plan théorique, le choix d'un système de poids peut être arbitraire. En pratique, il est le plus souvent orienté par certaines règles d'usage et de bon sens. Un choix particulier de système de poids engendre une mesure ajustée (mesure-résumé) particulière.

Sous-jacent à la notion d'ajustement, le concept de *standardisation* traduit l'idée d'ajustement de deux ou plusieurs ensembles de mesures spécifiques (issus de deux ou plusieurs groupes différents) en vue de pratiquer des comparaisons. La standardisation définit une condition essentielle pour pouvoir comparer deux mesures ajustées. Deux mesures ajustées sur un même système de poids sont dites *standardisées.*

Dans ce qui suit, nous allons examiner certaines pratiques d'ajustement pour les mesures de fréquence, d'association et d'impact. Pour nos exemples, la variable d'ajustement utilisée sera généralement l'âge. Nous allons, pour certains types d'ajustement, poser le problème de la standardisation, à savoir si l'ajustement présenté conduit à des mesures standardisées.

AJUSTEMENT DES MESURES DE FRÉQUENCE

Le problème de l'ajustement des mesures de fréquence ne se pose vraiment que dans la perspective de la standardisation. On veut, par exemple, comparer la mortalité de deux populations au moyen de mesures globales. On ne peut pas le faire avec les taux bruts de mortalité si les deux populations ont des structures d'âge différentes. En effet, comme on l'a vu au chapitre 4 dans la section Somme pondérée des mesures, le taux brut de décès reflète non seulement la mortalité spécifique mais aussi la structure par âge

d'une population. Pour comparer globalement la mortalité de deux populations, il faudra convenir d'une mesure-résumé qui neutralise la différence entre les deux distributions d'âge.

Comparaison de deux mesures de fréquence

On veut comparer la mortalité des deux populations *A* et *B,* décrites au tableau 16-2. Dans ce tableau, pour la strate d'âge *i*, *ni* désigne l'effectif en personnes-années, *pi* l'effectif relatif, d_i le nombre de décès, t_i le taux de décès exprimé par 100 000 personnes-années.

La façon la plus élémentaire de comparer la mortalité est de le faire groupe d'âge par groupe d'âge, catégorie par catégorie, strate par strate. Ce sont des comparaisons spécifiques, faites par l'intermédiaire des mesures spécifiques, ici de taux spécifiques. Ainsi, remarquons que, pour chaque catégorie d'âge, le taux de la population *A* est supé- rieur à celui de la population *B :* 10 > 4, 30 > 20 et 60 > 50.

Le tableau 16-2 montre la difficulté qu'il y a à utiliser les taux bruts de décès pour comparer de façon valide la mortalité entre deux populations, deux groupes. En effet, dans ce tableau, à l'inverse des taux spécifiques, le taux brut de la population *A* (21) est plus faible que celui de la population *B* (28). Cette situation paradoxale s'explique par des différences dans la structure par âge. Le taux brut calculé sur *A* n'a pas le même système de poids que celui calculé sur *B*. On comprend qu'en utilisant les taux bruts à des fins de comparaison, on se trouve à comparer à la fois les taux spécifiques et les distributions d'âge.

Pour lever le paradoxe, c'est-à-dire rendre valide la comparaison, il est nécessaire de contrôler l'âge. Pour en neutraliser l'effet, nous allons recourir à la standardisation. En pratique, cela revient à choisir une *distribution-type (ou population-type ou* de référence) que l'on substitue aux distributions d'âge des populations à comparer. La distribution-type est plus ou moins arbitraire mais reprend la même échelle de classification que celle des populations à comparer. Prenons par exemple, pour les groupes d'âge 35-54 ans, 55-64 ans, 65-74 ans, la distribution-type décrite au tableau 16-3. Le système de poids correspondant est: 0,500; 0,300; 0,200.

Tableau 16-2

Âge (en années)	Population *A*				Population *B*			
	n_i	Pi	d_i	t_i	n_i	P_i	d_i	
35-54	600 000	0,60	60	10	50000	0,25	2	4
55-64	300 000	0,30	90	30	70000	0,35	14	20
65-74	100 000	0,10	60	60	80000	0,40	40	50
Total	**1 000 000**	**1,00**	**210**	21	**200000**	**1,00**	**56**	**28**

En ajustant les taux spécifiques respectivement des populations *A* et *B* sur ce nouveau système de poids (0,500; 0,300; 0,200), on trouve pour *A* et pour *B* des taux de 26 et 18 par 100 000 personnes-années. En effet,

$$0{,}5 \times 10 + 0{,}3 \times 30 + 0{,}2 \times 60 = 26$$

$$0{,}5 \times 4 + 0{,}3 \times 20 + 0{,}2 \times 50 = 18$$

Ces taux ajustés sur le même système de poids sont comparables, donc standardisés. La standardisation a conduit à une relation d'ordre entre les taux ajustés qui va dans le même sens que celle entre les taux spécifiques. Les taux spécifiques de *A* sont plus grands que ceux de *B,* le taux ajusté aussi. Bien entendu, la valeur numérique d'un taux ajusté dépend de la distribution-type choisie ou, si l'on veut, du système de poids utilisé. De ce point de vue, la comparaison des mesures ajustées est au fond de nature plutôt qualitative. Quelle que soit d'ailleurs la distribution-type retenue, la relation d'ordre entre les taux standardisés va toujours dans le même sens que celle des taux spécifiques, pourvu qu'il n'y ait pas d'enjambement entre les taux spécifiques. Cette dernière condition mérite d'être expliquée.

Dans notre exemple, tous les taux spécifiques de A sont supérieurs aux taux spécifiques correspondants de *B:* $10 > 4$, $30 > 20$, $60 > 50$. La relation d'ordre est la même pour toutes les catégories. Il n'y a donc pas d'enjambement entre les taux spécifiques. Si, par contre, la relation change avec les catégories, il y a enjambement. Le tableau 16-4 illustre cette situation. On a : $10 > 4$, $30 < 40$, $60 > 50$.

Dans ce cas, la standardisation ne veut plus rien dire puisque l'on peut démontrer ou vérifier facilement que la relation d'ordre entre les taux ajustés peut varier avec la distribution-type choisie. En effet, si la distribution-type conduit au système de poids (0,10; 0,80; 0,10), les taux ajustés pour les populations A et *B* du tableau 16-4 sont respectivement de 31 et 37,4 par 100 000 ($31 < 37{,}4$). Si, par ailleurs, le système de poids est cette fois de (0,10; 0,40; 0,50), les taux ajustés de A et de *B* sont respectivement 43 et 41,4 par 100 000 ($43 > 41{,}4$). L'ordre est inversé.

En cas d'enjambement, il est préférable de s'en tenir à des comparaisons spécifiques. Celles-ci ont d'ailleurs l'avantage d'apporter plus d'information. L'absence d'enjambement devient une condition à une bonne pratique de la standardisation si l'on décide de procéder à une comparaison globale.

Enfin, soulignons que la standardisation peut être pratiquée pour contrôler non seulement

Tableau 16-3: Population-type choisie

Âge (en années)	Personnes-années	p_i
35-54	500 000	0,500
55-64	300 000	0,300
65-74	200 000	0,200
Total	**1 000 000**	**1,000**

Tableau 16-4

Âge (en années)	Population A t_i	Population B t_i
35-54	10	4
55-64	30	40
65-74	60	50

l'effet de l'âge mais aussi de toute autre variable de confusion susceptible d'influencer la comparaison. Il peut s'agir, selon le problème étudié, du sexe, de l'occupation, de la région, des habitudes de vie, etc.

Choix de la distribution-type

Bien que l'on puisse théoriquement choisir n'importe quelle distribution-type, il est de coutume d'adopter comme population de référence ou bien la somme des populations à comparer, ou bien l'une d'entre elles ou encore une population extérieure de recensement... La pratique et le bon sens veulent que la distribution-type choisie corresponde le plus possible à la distribution d'âge (ou autre variable) qui caractérise la population dans laquelle est conduite l'étude. Les mesures ajustées ainsi obtenues reflètent davantage une situation concrète. Par exemple, pour des comparaisons régionales, on pourra utiliser la structure d'âge de la population totale (celle du pays).

AJUSTEMENT DES MESURES D'ASSOCIATION *RR* ET *RC*

Le problème de l'ajustement des mesures d'association, comme le risque relatif *RR ou* le rapport des cotes *RC,* se pose essentiellement dans le cadre du contrôle de la confusion. Comme on l'a vu au chapitre précédent, le contrôle d'un facteur F confondant peut s'exercer, entre autres, au niveau de l'analyse par stratification et ce, tant pour les études de cohorte que pour les études cas-témoins. Pour chacune des strates, on calcule une mesure d'association en comparant les sujets exposés aux sujets non-exposés, les cas aux témoins. La stratification révèle ainsi un ensemble de mesures d'association spécifiques. L'ajustement vise à résumer ces mesures spécifiques.

Dans ce qui suit, nous présentons certaines méthodes d'ajustement pour le risque relatif *RR* et le rapport des cotes *RC,* en nous limitant à celles qui sont les plus utilisées. La présentation débute par un bref retour sur les notions d'ajustement et de système de poids.

Mesure ajustée et système de poids

Supposons que l'analyse des données d'une étude de cohorte conduise à un risque relatif brut de 1,6 et que la stratification pour le facteur âge révèle un ensemble assez homogène de mesures spécifiques comme celui au tableau 16-5.

Il n'est pas déraisonnable d'expliquer la variation des mesures spécifiques par les seules fluctuations dues au hasard et, par ce fait, de penser qu'elles ne sont que des estimations différentes d'une même mesure d'association dont la valeur devrait se situer aux alentours de 1,9. (Il va sans dire qu'un tel jugement gagnerait en solidité s'il était appuyé par les résultats d'un test statistique.)

Tableau 16-5

Âge (en années)	RR_i
40-49	1,8
50-59	2,1
60-69	1,7
Total	1,6

L'homogénéité entre les mesures spécifiques favorise la présentation des résultats à partir d'une seule mesure globale d'association qui, au plan statistique, présente des avantages certains: simplification de la présentation des données et stabilité de la mesure globale plus grande que celle des mesures spécifiques. Pour des raisons de validité, la mesure globale brute 1,6 ne peut pas être utilisée pour résumer l'ensemble des trois mesures spécifiques. La valeur 1,6 se situe à l'extérieur de l'intervalle [1,7; 2,1]. La mesure brute est biaisée par la présence d'un effet de confusion dans les données. Il faudra convenir d'une mesure, ajustée pour l'âge, qui dans sa forme maintient le contrôle de la confusion exercé par la stratification.

Considérons une autre situation où la mesure brute est encore de 1,6 mais cette fois la stratification dévoile un ensemble hétérogène de mesures spécifiques tel qu'il est décrit au tableau 16-6.

La mesure brute est encore ici biaisée puisque sa valeur se situe à l'extérieur de l'intervalle [1,8; 5,8]. Par ailleurs, contrairement à la situation précédente, on ne peut plus expliquer la variation entre les mesures spécifiques par les seules fluctuations d'échantillonnage. (S'il nous était possible d'utiliser un test statistique pour comparer ces mesures, les résultats de ce test supporteraient probablement notre jugement.) L'explication la plus vraisemblable est celle de la présence d'une interaction entre l'âge et le facteur étudié. La stratification a révélé une situation de fait: que l'âge modifie la relation d'association qui existe entre le facteur E et la maladie M. En soi, cette observation peut être intéressante. Elle invite à une présentation détaillée des mesures spécifiques plutôt qu'une mesure globale qui voile complètement le phénomène de modification. En présence de modification, nous croyons que la présentation des mesures spécifiques est préférable. Cependant, si le chercheur décide tout de même de présenter ces résultats en utilisant une mesure globale (il peut avoir les meilleures raisons), il faudra encore ici convenir d'une mesure ajustée pour l'âge.

Tableau 16-6

Âge (en années)	RR_i
40-49	1,8
50-59	3,2
60-69	5,8
Total	1,6

Une fois prise la décision de présenter une mesure ajustée, il faut choisir un système de poids qui va permettre de remplacer les mesures spécifiques par une somme pondérée de celles-ci. Toute mesure ajustée RR_{ajus} (RC_{ajus}) est donc formellement décrite par la relation algébrique :

$$RR_{ajus} = \Sigma\lambda_i RR_i,$$

ou bien pour le RC_{ajus}

$$RC_{ajus} = \Sigma\lambda_i RC_i,$$

où (λ_i) décrit un système de poids.

Si le système de poids (λ_i) choisi était (0,3; 0,4; 0,3), la mesure ajustée pour les données du tableau 16-6 serait:

(0,3 x 1,8) + (0,4 x 3,2) + (0,3 x 5,8) = 3,56.

Bien qu'arbitraire, le choix du système de poids est en pratique guidé par certaines règles d'usage. On essaiera, par exemple, d'adapter le système de poids à la distribution du facteur (l'âge par exemple) telle qu'elle se présente dans l'un ou l'autre des groupes étudiés. Ce choix, rappelons-le, conduit à une mesure ajustée qui représente bien la situation dans la population où est menée l'étude. Par exemple, si la distribution du facteur d'un groupe étudié apparaît comme au tableau 16-7,

Tableau 16-7

Âge (en années)	Nombre
40-49	3 000
50-59	5 000
60-69	2 000
Total	10 000

on peut choisir comme système de poids

$$\frac{3000}{10\,000}; \frac{5000}{10\,000}; \frac{2000}{10\,000}$$

ou 0,3 ; 0,5 ; 0,2

L'ajustement des mesures spécifiques, décrites au tableau 16-6, sur ce système de poids apparaît comme:

(0,3 x 1,8) + (0,5 x 3,2) + (0,2 x 5,8) = 3,3

Si (w_i) représente une distribution quel- conque, on a alors le système de poids correspondant (λ_i) défini par:

$$\lambda_i = \frac{w_i}{\Sigma w_i}.$$

En d'autres circonstances, le choix peut être davantage guidé par des considérations statistiques. Ainsi, voudra-t-on utiliser un système de poids qui tienne compte de la précision de chacune des mesures spécifiques. Ici, la règle générale veut que le poids accordé à chacune des mesures spécifiques soit proportionnel à sa stabilité. En d'autres mots, dans la somme pondérée, on veut accorder un poids d'autant plus grand à une mesure spécifique qu'elle a une meilleure précision. On rappelle que la variance d'une mesure marque son instabilité, son manque de précision. Il est donc raisonnable de penser que l'inverse de la variance d'une mesure marque sa stabilité. Pour ce type d'ajustement, les poids λ_i seront donc définis de façon que chacun soit proportionnel à l'inverse de la variance $(1 / V_i)$ de la mesure spécifique correspondante. Alors les poids λ_i sont donnés par l'expression suivante:

$$\lambda_i = \frac{1/V_i}{\Sigma 1/V_i},$$

ou si l'on veut par

$$= \frac{w_i}{\Sigma w_i}, \text{ en posant } w_i = 1/V_i.$$

Aussi, peut-on regrouper les procédures d'ajustement suivant que les poids sont définis ou bien à partir d'une distribution-type ou bien à partir du critère de la précision. Nous présentons pour ces deux voies, les procédures les plus utilisées, d'abord pour le *RR,* ensuite pour le *RC.* Nous donnons ensuite la description d'un type particulier d'ajustement dit de Mantel-Haenszel. Nous terminons par l'ajustement dans les analyses appariées.

Ajustement par des poids définis à partir d'une distribution-type

AJUSTEMENT DU RISQUE RELATIF *RR*

Rappelons pour la strate i, la forme générale des tableaux dans les études de cohorte:

	Strate *i*		
	E		
	+	—	
Malades	a_i	b_i	M_{1i}
Personnes (— temps)	N_{1i}	N_{oi}	N_i

$$RR_i = \frac{a_i N_{0i}}{b_i N_{1i}}$$

Les expressions suivantes décrivent deux types d'ajustement du *RR* qui utilisent comme nous le verrons, une distribution-type:

- $$RR_s = \frac{\Sigma a_i N_{0i}/N_{1i}}{\Sigma b_i}$$

- $$RR_a = \frac{\Sigma a_i}{\Sigma b_i N_{1i}/N_{0i}}$$

Appliquées l'une et l'autre aux données du tableau 16-8 relatives à une étude de cohorte, on trouve :

$$RR_s = \frac{\frac{20 \times 1000}{500} + \frac{120 \times 500}{1000}}{20 + 20} = 2,5$$

$$RR_a = \frac{20 + 120}{\frac{20 \times 500}{1000} + \frac{20 \times 1000}{500}} = 2,8$$

On peut souligner que le RR_S est équivalent au rapport des deux mesures globales de fréquence R_1 chez les sujets exposés et R_o chez les sujets non-exposés, ajustées sur la distribution du facteur *F* chez les sujets non-exposés (ici la distribution-type). Considérons les données du tableau 16-8. Si on ajuste R_1 sur la distribution du facteur *F* (l'âge) des sujets non-exposés, à savoir (1000; 500), ou si l'on veut sur le système de

Tableau *16-8*

Strate 1 *(F_1)*

M	*E* +	*E* —
+	20	20
—	500	1000

$RR_1 = 2$

Strate 2 *(F_2)*

M	*E* +	*E* —
+	120	20
—	1000	500

$RR_2 = 3$

Global

M	*E* +	*E* —
+	140	40
—	1500	1500

$RR_b = 3,5$

poids ($^{1000}/_{1500}$; $^{500}/_{1500}$), le rapport de ces mesures donne le RR_S. Vérifions-le.

$$R_1 = \frac{1000}{1500} \times \frac{20}{500} + \frac{500}{1500} \times \frac{120}{1000} = \frac{20}{300}$$

$$R_0 = \frac{1000}{1500} \times \frac{20}{1000} + \frac{500}{1500} \times \frac{20}{500} = \frac{8}{300}$$

En faisant le rapport de R_1 à R0, on obtient la valeur de RR_S, à savoir 2,5. On peut facilement montrer que $RR_S = \Sigma\lambda_i RR_i$ où le poids λ_i est donné par bi/Σbi.

Le RR_S est une mesure standardisée dans le sens que des RR_S, calculés dans une même étude pour différents niveaux d'exposition au facteur E, peuvent validement être comparés et ainsi permettre, par exemple, l'identification d'une relation dose—effet. En effet, à des niveaux différents (par exemple, niveaux 1 et 2), les poids du RR_S du niveau 1 et du RR_S du niveau 2 ne sont pas différents. Ils sont tous les deux $b_i/\Sigma b_i$, ce qui rend les RR_S comparables.

On peut souligner que RR_a est aussi une mesure équivalente au rapport des deux mesures globales de fréquences R_1 chez les sujets exposés et R_0 chez les sujets non-exposés, mais cette fois ajustées sur la distribution du facteur (l'âge) chez les sujets exposés (ici la distribution-type). Pour être plus concret, si dans l'exemple du tableau 16-8, on ajuste R_0 sur la distribution d'âge des sujets exposés (c'est déjà fait pour R_1), à savoir (500; 1000), ou si l'on veut sur le système de poids ($^{500}/_{1500}$; $^{1000}/_{1500}$), le rapport de ces mesures donne le RR_a. Vérifions-le.

$$R_1 = \frac{500}{1500} \times \frac{20}{500} + \frac{1000}{1500} \times \frac{120}{1000} = \frac{28}{300}$$

$$R_0 = \frac{500}{1500} \times \frac{20}{1000} + \frac{1000}{1500} \times \frac{20}{500} = \frac{10}{300}$$

En faisant le rapport de R_1 à R_0, on obtient la valeur de RR_a, *à* savoir 2,8. Le groupe de comparaison est le groupe des sujets non-exposés, la distribution-type est celle du groupe des sujets exposés. On peut facilement montrer que $RR_a = \Sigma\lambda_i RR_i$

$$\text{si on pose} \quad \lambda_i = \frac{b_i N_{1i}}{N_{0i}} \Big/ \Sigma \frac{b_i N_{1i}}{N_{0i}},$$

Contrairement au RR_S, la mesure RR_a n'est pas standardisée. Si l'on calcule le RR_a dans une même étude pour deux niveaux différents d'exposition au facteur *E* (par exemple, niveau 1 et niveau 2), les poids sont:

$$\lambda_{1i} = \frac{b_i N_{1i}}{N_{0i}} \Big/ \Sigma \frac{b_i N_{1i}}{N_{0i}} \text{ pour le niveau 1}$$

$$\lambda_{2i} = \frac{b_i N_{2i}}{N_{0i}} \Big/ \Sigma \frac{b_i N_{2i}}{N_{0i}} \text{ pour le niveau 2.}$$

Il y a donc inégalité entre les systèmes de poids (λ_{1i}) et (λ_{2i}) si la distribution du facteur F à contrôler varie avec le niveau d'exposition. Dans ce cas, le RR_a du niveau 1 ne peut pas être validement comparé au RR_a du niveau 2.

CAS PARTICULIER DU SMR

Le SMR (sigle de l'expression anglaise» « standardized mortality ratio ») est une mesure fort utilisée en santé au travail. En raison de son usage répandu, cette mesure mérite qu'on la présente de façon particulière. D'abord nous la définirons, ensuite nous illustrerons le fait que, comme le RR_a, cette mesure ajustée n'est pas standardisée. Enfin, nous donnerons quelques indications relatives à son utilisation.

Pour un groupe de N_1 individus exposés à un facteur E (par exemple, un groupe de travailleurs d'une industrie), le SMR est défi-ni comme le rapport du nombre observé de décès *(a)* dans ce groupe au nombre attendu de décès *(A)*, ce dernier nombre étant calculé d'après l'hypothèse que le groupe des exposés soit affecté des mêmes taux spécifiques Rti que ceux de la population totale.

Remarquons d'abord que la définition du SMR repose sur la comparaison d'un groupe de sujets exposés à la population totale qui le contient. Sous cet aspect, le SMR se différencie du risque relatif qui, lui, est issu de la comparaison d'un groupe de sujets exposés à un groupe de sujets non-exposés, ces deux groupes étant mutuellement exclusifs. On rappelle qu'à la base, le *RR* se définit par le rapport du risque R_1 *(a/N₁)* chez les sujets exposés au risque R_0 *(b/N₀)* chez les sujets non-exposés :

$$RR = \frac{R_1}{R_0}.$$

Le SMR, lui, comme l'indique sa définition, est le rapport *a/A* (on écrit aussi *O/A):*

$$\text{SMR} = \frac{a \text{ (nombre observé)}}{A \text{ (nombre attendu)}}$$

Le nombre A est obtenu en appliquant le risque R_t de la population totale au groupe des N_1 sujets exposés:

$$A = R_t \cdot N_1$$

Ainsi,

$$\text{SMR} = \frac{a}{A} = \frac{a}{R_t \cdot N_1} = \frac{R_1}{R_t}$$

Si on observe dans un groupe de 3000 travailleurs 16 décès par une maladie spécifiée et si le risque R_t dans la population est de 3 décès pour 1000 personnes-années (R_t = 0,003), alors le nombre observé de décès est de 16 et le nombre attendu de 9 décès (0,003 x 3000). Le SMR pour cette maladie est donc donné par $\frac{16}{9}$, c'est-à-dire 1,8. (Souvent dans la littérature le SMR est exprimé en pourcentage. Suivant cette convention, on pourrait écrire ici SMR = 180).

Le SMR se définit donc comme le rapport du risque R_1 chez les sujets exposés au risque R_t de la population totale. Si on représente par p1 la taille relative du groupe des sujets exposés à la population totale (par exemple d'une industrie de 500 travailleurs dans une population de 50 000, p_1 = 0,01), le risque R_t peut s'écrire :

$$R_t = p_1 R_1 + (1 - p_1) R_0.$$

On a donc:

$$\text{SMR} = \frac{R_1}{R_t} = \frac{R_1}{p_1 R_1 + (1 - p_1) R_0}$$

$$= \frac{RR}{p_1 (RR - 1) + 1}$$

où $RR = \frac{R_1}{R_0}$

Cette dernière expression permet d'établir le lien formel qui existe entre le SMR et le *RR*. Ainsi découvre-t-on que si $RR > 1$, le SMR est toujours formellement une sous-estimation du *RR*. Il est une bonne estimation du *RR* si le produit p_1 *(RR — 1)* est faible, c'est-à-dire si *RR* n'est pas trop grand et/ou si p_1 est faible. Cette dernière con-

dition, généralement satisfaite en santé au travail, est en pratique suffisante.

Tel qu'il est présenté dans la littérature, le SMR est une mesure ajustée, au même titre que le RR_a décrit précédemment. Considérons une étude où un groupe de sujets exposés est comparé à la population totale. Supposons que le contrôle d'un facteur confondant requiert la stratification des données pour ce facteur. Chaque strate i engendre une mesure spécifique SMR_i comme :

$$SMR_i = \frac{a_i}{A_i} = \frac{a_i}{R_{ti} N_{1i}}$$

Pour l'ensemble des strates, on a:

$$SMR = \frac{\Sigma a_i}{\Sigma A_i} = \frac{\Sigma a_i}{\Sigma R_{ti} \cdot N_{1i}}.$$

Cette dernière expression est celle du RR_a dans laquelle on a remplacé le risque b_i/N_{0i} des sujets non-exposés par le risque R_{ti} de la population totale.

Comme le RR_a, le SMR n'est pas malgré son nom, une mesure standardisée (il faudrait l'appeler AMR pour « adjusted mortality ratio »). Si on calcule le SMR dans une même étude pour deux niveaux différents d'exposition au facteur E, ou encore pour deux groupes de travailleurs (comme souvent on le fait en santé au travail), les SMR résultants ne sont pas comparables au plan de la validité. On comprend que le SMR est fonction de la distribution (N_{2i}) ou (N_{1i}) du facteur F à contrôler, suivant qu'il s'agisse du groupe 2 ou du groupe 1. Illustrons ce fait par un exemple numérique. Considérons deux groupes de travailleurs, G_1 et G_2 que l'on décide de comparer à la population P. Le tableau 16-9 décrit le nombre de décès, l'effectif et le taux de décès par année pour les deux groupes d'âge, jeune et âgé, pour chacun des groupes, G_1 et G_2 et pour la population P.

Alors les SMR_1 et SMR_2 pour les groupes G_1 et G_2 de travailleurs sont donnés respectivement par:

$$SMR_1 = \frac{3 + 16}{0{,}005 \times 200 + 0{,}010 \times 800}$$

$$= \frac{19}{9}$$

$$= 2{,}1$$

$$SMR_2 = \frac{12 + 4}{0{,}005 \times 800 + 0{,}010 \times 200}$$

$$= \frac{16}{6}$$

$$= 2{,}7$$

Bien que les taux spécifiques des deux groupes G_1 et G_2 soient identiques, les SMR_1 et SMR_2 sont manifestement différents. Ce fait est directement imputable à la différence entre les distributions d'âge des deux groupes

Tableau 16-9

Groupe		G_1	G_2	P
Jeune	décès	3	12	1000
	effectif	200	800	200 000
	taux	0,015	0,015	0,005
Âgé	décès	16	4	1000
	effectif	800	200	100 000
	taux	0,020	0,020	0,010

G_1 et G_2: (200; 800) pour G_1 contre (800; 200) pour G_2.

L'utilisation du SMR trouve certains avantages tant au plan pratique qu'au plan statistique. S'il est difficile d'identifier un groupe de sujets non-exposés adéquat comme groupe de comparaison, la population générale peut être alors un substitut intéressant, d'autant plus que généralement ses taux spécifiques sont directement accessibles par les statistiques vitales. Il faut ajouter que ces mêmes taux spécifiques jouissent au plan statistique d'une bonne stabilité parce qu'ils sont généralement calculés à partir de grands effectifs. Ces avantages ne doivent pas faire oublier deux limites importantes de cette mesure : le fait qu'elle ne soit pas standardisée, c'est-à-dire qu'elle ne se prête pas à des comparaisons, et sa grande sensibilité au biais de sélection décrit comme « l'effet de bonne santé » (« healthy worker effect »).

AJUSTEMENT DU RAPPORT DES COTES RC

Rappelons pour le strate i la forme générale des tableaux dans les études cas-témoins.

Strate i

		E +	E —	
M	+	a_i	b_i	
	—	c_i	d_i	M_{0i}
		N_{ii}	N_{0i}	N_i

$$RC_i = \frac{a_i d_i}{b_i c_i}$$

Les expressions suivantes décrivent deux types d'ajustement du *RC,* équivalents aux ajustements décrits pour le *RR.*

- $$RC_s = \frac{\Sigma \frac{a_i d_i}{c_i}}{\Sigma b_i}$$

- $$RC_a = \frac{\Sigma a_i}{\Sigma \frac{b_i c_i}{d_i}}$$

Appliquées l'une et l'autre aux données du tableau 16-10, relatives à une étude cas-témoins, on trouve :

$$RC_s = \frac{\frac{40 \times 150}{50} + \frac{150 \times 50}{50}}{60 + 50} = 2{,}45$$

et

$$RC_a = \frac{40 + 150}{\frac{60 \times 50}{150} + \frac{50 \times 50}{50}} = 2{,}71$$

Tableau 16-10

Strate 1 *(F_1)*

	E +	E —	
Cas	40	60	100
Témoins	50	150	200
		$RC_1 = 2$	

Strate 2 *(F_2)*

	E +	E —	
Cas	150	50	200
Témoins	50	50	100
		$RC_2 = 3$	

Global

	E +	E —	
Cas	190	110	300
Témoins	100	200	300
		$RC_b = 3{,}45$	

Le RC_S, comme le RR_S, est une mesure ajustée et standardisée aux fins de comparaisons internes entre plusieurs niveaux d'exposition. Les poids λ_i, donnés par $b_i/\Sigma b_i$, ne varient pas avec les niveaux d'exposition au sein d'une même étude. Le RC_a, comme le RR_a, est une mesure ajustée mais non standardisée. Elle ne se prête pas à des comparaisons valides. Les poids λ_i donnés par $\frac{b_i c_i}{d_i} \Big/ \Sigma \frac{b_i c_i}{d_i}$ comprennent un élément (c_i) qui dépend des niveaux d'exposition.

Ajustement par des poids définis à partir du critère de la précision des estimations

Rappelons que, suivant cette méthode, le système de poids est défini à partir de l'inverse de la variance. Pour des raisons statistiques (expliquées au chapitre suivant), il est préférable d'aborder le problème avec la transformation logarithmique du *RR (ou RC),* c'est-à-dire avec le ln *RR (ou* ln *RC).* La notation ln désigne le logarithme naturel ou népérien, c'est-à-dire de base e.

L'ajustement est alors pratiqué sur le logarithme de *RR (ou RC). A* la fin de l'opération, on applique au ln *RR* (ou ln *RC)* ajusté la transformation exponentielle (inverse du logarithme) pour finalement obtenir l'ajustement du *RR (ou RC)* désiré. Suivant cette méthode, on calcule d'abord:

$$\ln RR_v = \Sigma \frac{w_i}{\Sigma w_i} \ln RR_i = \frac{\Sigma w_i \ln RR_i}{\Sigma w_i}$$

où $w_i = 1/\text{Variance } (\ln RR_i)$

et ensuite, $RR_v = e^{\left(\frac{\Sigma w_i \ln RR_i}{\Sigma w_i}\right)}$

où l'indice v réfère à l'ajustement par l'inverse de la variance. (Dans ces formules, on peut remplacer RR_v et RR_i par RC_v et RC_i).

Même s'il n'est pas apparent sous cette forme, le RR_v est une mesure-résumé des mesures spécifiques RR_i. En utilisant les propriétés du logarithme, on peut transformer les expressions ci-dessus pour obtenir l'expression équivalente :

$$RR_v = (RR_1^{w_1} \times RR_2^{w_2} \times \ldots \times RR_k^{w_k})^{1/\Sigma w_i}.$$

Cette fois le *RR,* apparaît clairement comme une moyenne géométrique des RR_i, pondérée en fonction des w_i. Le RR_v est une mesure qui résume les RR_i au moyen d'une relation géométrique.

Nous présentons maintenant, suivant le critère de la précision, à l'aide d'exemples, les mesures ajustées pour le rapport des taux d'incidence et le rapport des incidences cumulatives dans les études de cohorte et pour le *RC* dans les études cas-témoins ou les études transversales.

AJUSTEMENT DU RISQUE RELATIF *RR* COMME RAPPORT DES TAUX D'INCIDENCE

Si le *RR* (RR_i) est un rapport de taux d'incidence, on peut démontrer que dans l'hypothèse d'une non-association ($RR = 1$):

$$\text{Variance } (\ln RR_i) = N^2_1/N_{1i}N_{0i}M_{1i}.$$

Il s'ensuit que dans la formule générale

$$RR_v = e^{\left(\frac{\Sigma w_i \ln RR_i}{\Sigma w_i}\right)}$$

les w_i sont donnés par:

$$w_i = \frac{N_{1i}N_{0i}M_{1i}}{N_i^2}.$$

Considérons à nouveau les données du tableau 16-8. Nous supposons que les dernières lignes réfèrent à des personnes-années, de sorte que les rapports $^{20}/_{500}$, $^{20}/_{1000}$... correspondent à des taux d'incidence.
Pour la première et la deuxième strate, on a respectivement:

$$w_1 = \frac{500 \times 1000 \times 40}{1500^2} = 8{,}889$$
$$w_2 = \frac{1000 \times 500 \times 140}{1500^2} = 31{,}111$$

Par conséquent,

$$\ln RR_v = \frac{\Sigma w_i \ln RR_i}{\Sigma w_i}$$

$$= \frac{8{,}889\ (\ln\ 2) + 31{,}111\ (\ln\ 3)}{8{,}889 + 31{,}111}$$

$$= 1{,}0085$$

Ainsi, $RR_v = e^{1{,}0085} = 2{,}74.$

Il existe une autre formule pour la variance du ln RR_i lorsque le RR_i est un rapport de taux d'incidence. Elle est valable sans qu'il soit nécessaire de faire l'hypothèse de non-association. Cette formule est:

$$\text{Variance } (\ln RR_i) = \frac{1}{a_i} + \frac{1}{b_i}$$

Il s'ensuit que les w_i sont donnés par:

$$w_i = \frac{1}{\frac{1}{a_i} + \frac{1}{b_i}} = \frac{a_i b_i}{a_i + b_i}$$

Pour la première et la deuxième strate du même tableau 16-8, on a respectivement:

$$w_1 = \frac{20 \times 20}{20 + 20} = 10$$

$$w_2 = \frac{120 \times 20}{120 + 20} = 17{,}143$$

Par conséquent,

$$\ln RR_v = \frac{\Sigma w_i \ln RR_i}{\Sigma w_i}$$

$$= \frac{10\ (\ln\ 2) + 17{,}143\ (\ln\ 3)}{10 + 17{,}143}$$

$$= 0{,}9492$$

Ainsi, $RR_v = e^{0{,}9492} = 2{,}58$

AJUSTEMENT DU RISQUE RELATIF *(RR)* COMME RAPPORT DES INCIDENCES CUMULATIVES

Si le RR (RR_i) est un rapport d'incidences cumulatives, on peut démontrer que, dans l'hypothèse d'une non-association:

Variance (ln RR_i) = $M0_iN_i/N_{ii}N0_iM_{ii}$,

où $M_{0i}=N_i-M_{1i}$.

Il s'ensuit que dans la formule générale:

$$RR_v = e^{\left(\frac{\Sigma w_i \ln RR_i}{\Sigma w_i}\right)}$$

les w_i sont donnés cette fois par:

$$w_i = \frac{N_{1i}N_{0i}M_{1i}}{M_{0i}N_i}, \quad \text{où } M_{0i} = N_i - M_{1i}.$$

Considérons à nouveau les données du tableau 16-8. Nous supposons que les dernières lignes réfèrent cette fois à des personnes de sorte que les rapports $^{20}/_{500}$, $^{20}/_{1000}$... correspondent à des incidences cumulatives. Pour la première et la deuxième strate, on a respectivement:

$$w_1 = \frac{500 \times 1000 \times 40}{1460 \times 1500} = 9{,}132$$

$$w_2 = \frac{1000 \times 500 \times 140}{1360 \times 1500} = 34{,}314$$

Par conséquent,

$$\ln RR_v = \frac{\Sigma w_i \ln RR_i}{\Sigma w_i}$$

$$= \frac{9{,}132\ (\ln\ 2) + 34{,}314\ (\ln\ 3)}{9{,}132 + 34{,}314}$$

$$= 1{,}0134$$

Ainsi, $RR_v = e^{1{,}0134} = 2{,}75$

II existe une autre formule pour la variance du ln RR_i lorsque le RR_i est un rapport d'incidences cumulatives. Elle n'exige pas que soit faite l'hypothèse de non-association. Cette formule est:

$$\text{Variance } (\ln RR_i) = \frac{c_i}{a_i N_{1i}} + \frac{d_i}{b_i N_{0i}}$$

Il s'ensuit que les w_i sont donnés par:

$$w_i = \frac{1}{\frac{c_i}{a_i N_{1i}} + \frac{d_i}{b_i N_{0i}}} = \frac{a_i b_i N_{1i} N_{0i}}{a_i d_i N_{1i} + b_i c_i N_{0i}}$$

Pour la première et la deuxième strate, toujours du tableau 16-8, on a respectivement:

$$w_1 = \frac{20 \times 20 \times 500 \times 1000}{20 \times 980 \times 500 + 20 \times 480 \times 1000}$$

$$= 10{,}309$$

$$w_2 = \frac{120 \times 20 \times 1000 \times 500}{120 \times 480 \times 1000 + 20 \times 880 \times 500}$$

$$= 18{,}072$$

Par conséquent,

$$\ln RR_v = \frac{\Sigma w_i \ln RR_i}{\Sigma w_i}$$

$$= \frac{10{,}309\ (\ln\ 2) + 18{,}072\ (\ln\ 3)}{10{,}309 + 18{,}072}$$

$$= 0{,}9513$$

Ainsi, $RR_v = e^{0{,}9513} = 2{,}59$

AJUSTEMENT DU RAPPORT DES COTES RC DANS LES ÉTUDES CAS-TÉMOINS OU LES ÉTUDES TRANSVERSALES

On peut démontrer que la variance de ln RC_i est donnée par:

$$\text{Variance } (\ln RC_i) = \frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}$$

Il s'ensuit que dans la formule générale:

$$RC_v = e^{\left(\frac{\Sigma w_i \ln RC_i}{\Sigma w_i}\right)}$$

les w_i sont donnés par:

$$w_i = \frac{1}{\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}}.$$

Appliquée aux données du tableau 16-10, la mesure conduit au résultat suivant:

$$w_1 = \frac{1}{\frac{1}{40} + \frac{1}{60} + \frac{1}{50} + \frac{1}{150}}$$

$$= 14{,}634$$

$$w_2 = \frac{1}{\frac{1}{150} + \frac{1}{50} + \frac{1}{50} + \frac{1}{50}}$$

$$= 15$$

On a par conséquent,

$$\ln RC_v = \frac{\Sigma w_i \ln RC_i}{\Sigma w_i}$$

$$= \frac{14{,}634\ (\ln\ 2) + 15\ (\ln\ 3)}{14{,}634 + 15}$$

$$= 0{,}8984$$

Ainsi, $RC_v = e^{0{,}8984} = 2{,}46$

On peut noter en terminant que le RR_v *(ou* RC_v*)* n'est pas standardisé puisque les poids qui le définissent font intervenir des éléments (a_i, N_{1i}, M_{1i}) qui pourraient varier avec les niveaux d'exposition.

Ajustement de Mantel-Haenszel

Ce type d'ajustement a d'abord été défini pour le *RC* dans les études cas-témoins. Comme l'indiquent les auteurs, ce procédé utilise des poids qui tiennent compte à la fois de la précision des estimations et de leur importance. La mesure globale résultante RC_{MH} a certaines propriétés intéressantes, notamment sa flexibilité de calcul dans le cas où les tableaux présentent des cellules nulles.

La mesure RC_{MH} a la forme:

$$RC_{MH} = \frac{\Sigma a_i d_i / N_i}{\Sigma b_i c_i / N_i}.$$

Appliquée aux données du tableau 16-10, la mesure RC_{MH} conduit au résultat suivant:

$$RC_{MH} = \frac{\frac{40 \times 150}{300} + \frac{150 \times 50}{300}}{\frac{60 \times 50}{300} + \frac{50 \times 50}{300}}$$

$$= 2{,}45.$$

Il est facile de vérifier que RC_{MH} est bien une mesure ajustée car elle peut s'exprimer dans la forme $\Sigma \lambda_i RC_i$ si on pose:

$$\lambda_i = \frac{b_i c_i}{N_i} \Big/ \Sigma \frac{b_i c_i}{N_i}.$$

En effet,

$$RC_{MH} = \frac{\Sigma a_i d_i / N_i}{\Sigma b_i c_i / N_i}$$

$$= \frac{\Sigma \frac{b_i c_i}{N_i} \times \frac{a_i d_i}{b_i c_i}}{\Sigma \frac{b_i c_i}{N_i}}$$

$$= \frac{\Sigma \frac{b_i c_i}{N_i} RC_i}{\Sigma \frac{b_i c_i}{N_i}}$$

L'ajustement de Mantel-Haenszel adapté au *RR* pour les études de cohorte conduit à

une mesure ajustée RR_{MH} qui a la forme suivante:

$$RR_{MH} = \frac{\Sigma a_i N_{0i}/N_i}{\Sigma b_i N_{1i}/N_i}.$$

Appliquée aux données du tableau 16-8, la mesure RR_{MH} conduit au résultat suivant:

$$RR_{MH} = \frac{\frac{20 \times 1000}{1500} + \frac{120 \times 500}{1500}}{\frac{20 \times 500}{1500} + \frac{20 \times 1000}{1500}}$$

$$= 2{,}67.$$

Il est facile de vérifier que RR_{MH} est une mesure ajustée qui résume les risques relatifs spécifiques RR_i si l'on utilise les poids $\frac{b_i N_{1i}/N_i}{\Sigma b_i N_{1i}/N_i}$ et si l'on se rappelle que $RR_i = \frac{a_i N_{0i}}{b_i N_{1i}}$.

On peut noter en terminant que le RC_{MH} et le RR_{MH} ne sont pas standardisés puisque les poids qui les définissent font intervenir des éléments (c_i, N_{1i}) qui pourraient varier avec les niveaux d'exposition.

Ajustement dans les analyses appariées

Quand l'appariement a été pratiqué au niveau de l'échantillonnage, l'investigateur peut, au niveau de l'analyse, décider de pratiquer une analyse par paire (ou analyse appariée). Nous ne considérons ici que les analyses appariées 1:1, c'est-à-dire celles où les paires sont (1 cas, 1 témoin) ou (1 sujet exposé, 1 sujet non-exposé).

AJUSTEMENT DU RISQUE RELATIF *(RR)*

Considérons une étude de cohorte où chaque sujet exposé est apparié à un sujet non-exposé bien spécifique. Ainsi, chacun des 100 sujets exposés est assorti à un sujet non-exposé. L'observation est faite pour déterminer si chacun des 200 sujets à l'étude est affecté ou non par la maladie *M.* Supposons que l'on observe parmi les 100 paires ainsi constituées :

10 paires où le sujet exposé et le sujet non-exposé sont affectés par *M;*

50 paires où ni le sujet exposé ni le sujet non-exposé n'est affecté par *M;*

30 paires où seul le sujet exposé est affecté par *M;*

10 paires où seul le sujet non-exposé est affecté par *M.*

Ces données peuvent être placées dans un tableau 2 x 2 comme le tableau 16-11:

Tableau 16-11

		Exposés M		
		+	—	Total
Non-exposés M	+	10	10	20
	—	30	50	80
		Total 40	60	100

Les nombres qui apparaissent dans ce tableau décrivent les paires (sujet exposé, sujet non-exposé) et non les individus. La mesure du *RR* dans un tel tableau est obtenue en faisant le rapport du risque chez les sujets exposés ($^{40}/_{100}$) au risque chez les sujets non-exposés ($^{20}/_{100}$), soit $^{40}/_{20}$ (= 2).

Pour la forme générale, si on désigne par:

r, le nombre de paires où les deux sujets sont affectés par *M;*

s, le nombre de paires où seul le sujet non-exposé est affecté par *M;*

t, le nombre de paires où seul le sujet exposé est affecté par *M;*

u, le nombre de paires où ni le sujet exposé ni le sujet non-exposé n'est affecté par *M.*

On aura le tableau 16-12:

Tableau 16-12

		Exposés *M* +	Exposés *M* —	Total
Non-exposés *M*	+	*r*	*s*	*r*+*s*
	—	*t*	*u*	*t*+*u*
Total				*r*+*t* *s*+*u*r+*s*+*t*+*u*

La mesure *RR* est alors donnée par:

$$RR = \frac{r + t}{r + s}.$$

On peut montrer que le *RR* ainsi calculé est équivalent à un RR_{MH}.

D'abord, il est naturel de concevoir une paire (sujet exposé, sujet non-exposé) comme une strate où se retrouvent uniquement deux sujets: 1 sujet exposé et 1 sujet non-exposé. S'il y a *n* paires, il y a *n* strates. Pour la paire *i*, la strate correspondante peut se décrire à l'aide du tableau 2 x 2 déjà connu:

Strate *i*

		E +	*E* —	
M	+	a_i	b_i	
	—	c_i	d_i	
		1	1	2

OÙ $a_i = 0$ *ou* 1,

$b_i = 0$ ou 1

$c_i = 1$ ou 0

$d_i = 1$ou 0

$N_{1i} = N_{oi} = 1$

$N_i = 2$

Il y a *r* strates où $a = 1$, $b = 1$, $c = 0$, $d = 0$. Le tableau 16-13 donne le tableau 2 x 2 correspondant.

Il y a *s* strates où $a = 0$, $b = 1$, $c = 1$, $d = 0$. Le tableau 16-14 donne le tableau 2 x 2 correspondant.

Il y a *t* strates où $a = 1$, $b = 0$, $c = 0$, $d = 1$. Le tableau 16-15 donne le tableau 2 x 2 correspondant.

Il y a u strates où $a = 0$, $b = 0$, $c = 1$, $d = 1$. Le tableau 16-16 donne le tableau 2 x 2 correspondant.

En appliquant l'ajustement de Mantel-Haenszel sur les n strates, on a:

$$RR_{MH} = \frac{\Sigma a_i N_{0i}/N_i}{\Sigma b_i N_{1i}/N_i}$$

$$= \frac{r \cdot \frac{1 \times 1}{2} + s \cdot \frac{0 \times 1}{2} + t \cdot \frac{1 \times 1}{2} + u \cdot \frac{0 \times 1}{2}}{r \cdot \frac{1 \times 1}{2} + s \cdot \frac{1 \times 1}{2} + t \cdot \frac{0 \times 1}{2} + u \cdot \frac{0 \times 1}{2}}.$$

$$= \frac{r \times ½ + t \times ½}{r \times ½ + s \times ½}$$

$$= \frac{r + t}{r + s}$$

Le rapport $\frac{r + t}{r + s}$ est un RR ajusté selon la méthode de Mantel-Haenszel adaptée aux études de cohorte.

Tableau 16-13

r strates

		E +	E −	
M	+	1	1	
	−	0	0	
		1	1	2

Tableau 16-14

s strates

		E +	E −	
M	+	0	1	
	−	1	0	
		1	1	2

Tableau 16-15

t strates

		E +	E −	
M	+	1	0	
	−	0	1	
		1	1	2

Tableau 16-16

u strates

		E +	E −	
M	+	0	0	
	−	1	1	
		1	1	2

AJUSTEMENT DU RAPPORT DES COTES (RC)

Considérons une étude cas-témoins où chacun des 100 cas est apparié à un témoin bien spécifique. L'étude est ainsi constituée de 100 paires (cas, témoin).

L'observation, faite pour déterminer si chacun des 200 sujets à l'étude a été exposé ou non au facteur *E,* conduit aux résultats suivants :

— 10 paires où le cas et le témoin sont exposés à *E;*

— 50 paires où ni le cas ni le témoin n'est exposé à *E;*

— 30 paires où seul le cas est exposé à *E;*

— 10 paires où seul le témoin est exposé à *E.*

Ces données peuvent être placées dans un tableau 2 x 2 comme le tableau 16-17:

Tableau 16-17

			Cas E +	Cas E −	Total
Témoins	E	+	10	10	20
		−	30	50	80
		Total	40	60	100

Les nombres qui apparaissent dans ce tableau décrivent les paires (cas, témoin) et non les individus. La mesure du *RC* dans un tel tableau est obtenue en faisant le rapport du nombre de paires (30) où seul le cas est exposé au nombre de paires (10) où seul le témoin est exposé.

$$RC = \frac{30}{10}$$

Pour la forme générale, si on désigne par:

r, le nombre de paires où les deux sujets sont exposés à *E;*

s, le nombre de paires où seul le témoin est exposé à *E;*

t, le nombre de paires où seul le cas est exposé à *E;*

u, le nombre de paires où ni le cas ni le témoin n'est exposé à *E.*

On aura le tableau 16-18:

Tableau 16-18

		Cas *E* +	Cas *E* —	Total
Témoins *E*	+	*r*	*s*	*r+s*
	—	*t*	*u*	*t+u*
	Total	*r+t*	*s+u*	*r+s+t+u*

La mesure *RC* est alors donnée par:

$$RC = \frac{t}{s}.$$

On peut montrer que le *RC* ainsi calculé est équivalent à un RC_{MH}.

D'abord, il est naturel de concevoir une paire (cas, témoin) comme une strate sur laquelle on trouve uniquement deux sujets:

1 cas et 1 témoin. S'il y a *n* paires, il y a *n* strates. Pour la paire i, la strate correspondante peut se décrire à l'aide du tableau 2 x 2 suivant:

		Strate *i* *E* +	Strate *i* *E* —	
M	+	a_i	b_i	1
	—	c_i	d_i	1
				2

OÙ $a_i = 0$ ou 1,

$b_i = 1$ou 0

$c_i = 0$ou1

$d_i = 1$ou 0

$M_{1i} = Mo_i = 1$

$N_i = 2.$

Il y a r strates où $a = 1$, $b = 0$, $c = 1$, $d = 0$. Le tableau 16-19 donne le tableau 2 x 2 correspondant.

Tableau 16-19

		r strates *E* +	*r* strates *E* —	
M	+	1	0	1
	—	1	0	1
				2

Tableau 16-20

		s strates *E* +	*s* strates *E* —	
M	+	0	1	1
	—	1	0	1
				2

Tableau 16-21

		t strates *E* +	*t* strates *E* —	
M	+	1	0	1
	—	0	1	1
				2

Tableau 16-22

		u strates *E* +	*u* strates *E* —	
M	+	0	1	1
	—	0	1	1
				2

Il y a s strates où $a = 0$, $b = 1$, $c = 1$, $d = 0$. Le tableau 16-20 donne le tableau 2 *x* 2 correspondant.

Il y a t strates où $a = 1$, $b = 0$, $c = 0$, $d = 1$. Le tableau 16-21 donne le tableau 2 x 2 correspondant.

Il y a u strates où $a = 0$, $b = 1$, $c = 0$, $d = 1$. Le tableau 16-22 donne le tableau 2 x 2 correspondant.

En appliquant l'ajustement de Mantel-Haenszel sur les *n* strates, on a:

$$RC_{MH} = \frac{\Sigma a_i d_i / N_i}{\Sigma b_i c_i / N_i}$$

$$= \frac{r \cdot \frac{1 \times 0}{2} + s \cdot \frac{0 \times 0}{2} + t \cdot \frac{1 \times 1}{2} + u \cdot \frac{0 \times 1}{2}}{r \cdot \frac{0 \times 1}{2} + s \cdot \frac{1 \times 1}{2} + t \cdot \frac{0 \times 0}{2} + u \cdot \frac{1 \times 0}{2}}.$$

$$= \frac{t \times \frac{1}{2}}{s \times \frac{1}{2}}$$

$$= \frac{t}{s}$$

Le rapport $\frac{t}{s}$ est un *RC* ajusté selon la méthode de Mantel-Haenszel.

AJUSTEMENT DES MESURES D'IMPACT

Nous allons aborder successivement dans cette section l'ajustement des fractions étiologiques et des fractions prévenues. Le type d'ajustement approprié à ces mesures est celui qui rend *cohérente* l'interprétation de la mesure globale avec celle des mesures spécifiques. Le type d'ajustement défini pour les fractions étiologiques est heureusement transposable, ou presque, aux fractions prévenues.

Ajustement des fractions étiologiques

Pour discuter l'ajustement des fractions étiologiques tant chez les sujets exposés que dans la population totale, nous allons nous référer constamment aux données du tableau 16-8, que nous reproduisons ici au tableau 16-23.

Tableau 16-23

Strate 1 (F_1)

		E +	E −
M	+	20	20
	−	500	1000

$RR_1 = 2$

Strate 2 (F_2)

		E +	E −
M	+	120	20
	−	1000	500

$RR_2 = 3$

Global

		E +	E −
M	+	140	40
	−	1500	1500

$RR_b = 3{,}5$

AJUSTEMENT DE LA FRACTION ÉTIOLOGIQUE CHEZ LES SUJETS EXPOSÉS : FE_I

Pour chacune des deux strates et globalement, les fractions étiologiques chez les sujets exposés sont:

$$FE_{11} = FE_{1\ (\text{strate } 1)} = \frac{2-1}{2} = 0{,}50 \text{ (ou } \tfrac{1}{2})$$

$$FE_{12} = FE_{1\ (\text{strate } 2)} = \frac{3-1}{3} = 0{,}67 \text{ (ou } \tfrac{2}{3})$$

$$FE_1 = FE_{1\ (\text{globale})} = \frac{3{,}5-1}{3{,}5} = 0{,}71 \text{ (ou } \tfrac{5}{7})$$

où FE_1, en l'absence d'un deuxième indice, désigne la fraction globale, c'est-à-dire celle où les strates 1 et 2 sont réunies. Si l'on interprète ces fractions étiologiques, on peut dire qu'il y a, parmi les cas exposés:

10 (= 0,50 *x* 20) cas attribuables au facteur *E* dans la première strate;

80 (= 0,67 x 120) cas attribuables au facteur *E* dans la deuxième strate;

100 (= 0,71 x 140) cas attribuables au facteur *E* dans la totalité des cas exposés.

Cette interprétation souffre d'une incohérence. Au total, parmi les 140 cas, on devrait dénombrer, non pas 100 cas attribuables au facteur *E,* mais 90 (10 + 80). Cette incohérence est provoquée en partie par le fait que la fraction étiologique globale FE_1 est calculée à partir d'une mesure *RR* biaisée (le RR_b). Toutefois, même si on décide de calculer FE_1 à partir d'un *RR* ajusté, il n'est pas sûr que, de ce fait, la cohérence soit atteinte. C'est le cas notamment si l'on utilise la mesure ajustée standardisée RR_S.

Étant donné qu'ici

$$RR_s = \frac{\Sigma a_i N_{0i}/N_{1i}}{\Sigma b_i}$$

$$= \frac{\frac{20 \times 1000}{500} + \frac{120 \times 500}{1000}}{20 + 20} = 2{,}5$$

la fraction étiologique chez la totalité des cas exposés devient $\frac{2{,}5-1}{2{,}5} = 0{,}60$. Cette estimation n'est pas cohérente avec les mesures spécifiques puisqu'elle donne un nombre de 84 cas attribuables (0,60 × 140) plutôt que 90.

Le type d'ajustement cohérent est celui dont le système de poids est basé sur la distribution du facteur confondant *F* parmi les cas exposés. Dans notre exemple (tableau 16-23) où cette distribution est (20; 120), le système de poids correspondant est: ($^{20}/_{140}$; $^{120}/_{140}$). On obtient comme ajustement:

$$FE_1 = (^{20}/_{140})\ FE_{11} + (^{120}/_{140})\ FE_{12}$$
$$= (^{20}/_{140})\ \text{X}\ 0{,}50 + (^{120}/_{140})\ \text{x}\ 0{,}67$$
$$= 0{,}64\ (\text{ou } ^{9}/_{14}).$$

Cette dernière fraction étiologique chez la totalité des cas exposés est, dans son interprétation, cohérente avec les fractions étiologiques spécifiques. En effet, elle conduit à une estimation de 90 cas attribuables ($^{9}/_{14}$ X 140). De façon générale, une fraction étiologique cohérente chez la totalité des cas exposés est donnée par la formule :

$$FE_1 =$$

où (a_i) représente, la distribution du facteur confondant chez les cas exposés.

Il est intéressant de remarquer que la FE_1 (cohérente) est ajustée au RR_a (ou SMR), c'est-à-dire que:

$$FE_1 = \Sigma \frac{a_i}{\Sigma a_i} FE_{1i} = \frac{RR_a - 1}{RR_a}.$$

En voici la démonstration.

$$\Sigma \frac{a_i}{\Sigma a_i} FE_{1i} = \Sigma \left(\frac{a_i}{\Sigma a_i} \cdot \frac{RR_i - 1}{RR_i} \right)$$

$$= \frac{\Sigma a_i (1 - 1/RR_i)}{\Sigma a_i}$$

$$= \frac{\Sigma a_i \left(1 - \frac{b_i N_{1i}}{a_i N_{0i}}\right)}{\Sigma a_i}$$

$$= \frac{\Sigma a_i - \Sigma \frac{b_i N_{1i}}{N_{0i}}}{\Sigma a_i}$$

$$= \frac{\left(\Sigma a_i \Big/ \Sigma \frac{b_i N_{1i}}{N_{0i}}\right) - 1}{\Sigma a_i \Big/ \Sigma \frac{b_i N_{1i}}{N_{0i}}}$$

$$= \frac{RR_a - 1}{RR_a}$$

En définitive, la cohérence est atteinte si la fraction étiologique chez les sujets exposés FE_1 est exprimée par $\frac{RR_a - 1}{RR_a}$

Dans notre exemple (toujours le tableau 16-23), on a :

$$RR_a = \frac{\Sigma a_i}{\Sigma \frac{b_i N_{1i}}{N_{0i}}}$$

$$= \frac{20 + 120}{\frac{20 \times 500}{1000} + \frac{20 \times 1000}{500}} = 2{,}8$$

et

$$FE_1 \text{ (cohérente)} = \frac{2{,}8 - 1}{2{,}8} = 9/14.$$

On observe bien la cohérence avec les fractions étiologiques spécifiques puisque ($^9/_{14}$) x 140 = 90 (= 10 + 80).

La démonstration de la cohérence de l'ajustement RC_a avec la fraction étiologique FE_1 dans les études cas-témoins est analogue. On n'a qu'à substituer dans la démonstration précédente ci à N_{1i} et d_i à N_{0i}.

AJUSTEMENT DE LA FRACTION ÉTIOLOGIQUE TOTALE : FE_t

Utilisons encore ici les données du tableau 16-23. Supposons, aux fins de l'exemple, que ces données reflètent, en toutes bonnes probabilités, la situation réelle de la population dans laquelle est conduite l'étude, ce qui serait le cas si l'étude était conduite sur la population entière ou sur un échantillon représentatif de celle-ci. Dans ce contexte, la proportion p_{ci} de sujets exposés chez les cas (première strate, deuxième strate et globalement ou en totalité) estimée par l'étude ne résulte pas d'un artifice, mais représente la réalité. De même, la proportion p_1 de sujets exposés dans l'échantillon serait une estimation de la proportion réelle de sujets exposés dans la population.

Pour chacune des deux strates et globalement, nous avons les fractions étiologiques totales suivantes:

$$FE_{t1} = \frac{20}{40} \times \frac{2-1}{2} = 0{,}25 \text{ (ou } ¼)$$

$$FE_{t2} = \frac{120}{140} \times \frac{3-1}{3} = 0{,}57 \text{ (ou } ⁴⁄₇)$$

$$FE_t = \frac{140}{180} \times \frac{3{,}5-1}{3{,}5} = 0{,}56 \text{ (ou } ⁵⁄₉).$$

Si l'on interprète ces fractions étiologiques, on peut dire qu'il y a, parmi les cas,

10 (= $^1/_4$ x 40) cas attribuables au facteur *E* dans la première strate;

80 (= $^4/_7$x 140) cas attribuables au facteur *E* dans la deuxième strate;

100 (= $^5/_9$ x 180) cas attribuables au facteur *E* dans la totalité des cas.

Cette interprétation présente aussi une incohérence. Au total, parmi les 180 cas, on devrait dénombrer, non pas 100 cas attribuables au facteur *E,* mais 90 (10 + 80). Cette incohérence est induite en partie par le fait que la fraction étiologique globale FE_t est calculée à partir d'une mesure *RR* biaisée. Même calculée à partir d'un *RR* ajusté, il n'est pas sûr que la cohérence soit atteinte. Si on utilise, par exemple, la mesure *RR, (= 2,5),* on a,

$$FE_t = \frac{140}{180} \times \frac{2{,}5-1}{2{,}5} = 0{,}47 \text{ (ou } ⁷⁄₁₅).$$

Cette estimation n'est pas non plus cohérente avec les mesures spécifiques puisqu'elle conduit à un nombre de 84 cas attribuables plutôt que 90.

Le type d'ajustement cohérent est celui basé sur la distribution du facteur confondant parmi les cas. En effet, dans l'exemple au tableau 16-23, cette distribution est (40; 140); le système de poids correspondant est ($^{40}/_{180}$; $^{140}/_{180}$). On obtient comme ajustement,

$$FE_t = (^{40}⁄_{180})\, FE_{t1} + (^{140}⁄_{180})\, FE_{t2}$$

$$= (^{40}⁄_{180}) \times 0{,}25 + (^{140}⁄_{180}) \times 0{,}57$$

$$= 0{,}50 \text{ (ou } ½).$$

Cette dernière fraction étiologique chez la totalité des cas est, dans son interprétation, cohérente avec les fractions étiologiques spécifiques. En effet, elle conduit à une estimation de 90 cas attribuables (0,50 x 180). De façon générale, ici aussi, une fraction étiologique cohérente chez la totalité des cas est donnée par la formule :

$$FE_t \text{ (cohérente)} = \Sigma \frac{M_{1i}}{\Sigma M_{1i}} FE_{ti}$$

où (M_{1i}) représente la distribution du facteur confondant chez les cas.

Il est intéressant de remarquer qu'encore ici la FE_t (cohérente) est ajustée au RR_a (ou SMR), c'est-à-dire que :

$$FE_t \text{ (cohérente)} = \Sigma \frac{M_{1i}}{\Sigma M_{1i}} FE_{ti} = p_{c1} \frac{RR_a - 1}{RR_a}$$

où $p_{c1} = \dfrac{\Sigma a_i}{\Sigma M_{1i}}$.

En voici la démonstration.

$$\Sigma \frac{M_{1i}}{\Sigma M_{1i}} FE_{ti} = \Sigma \left(\frac{M_{1i}}{\Sigma M_{1i}} \cdot p_{C1i} \frac{RR_i - 1}{RR_i} \right),$$

où $p_{C1i} = \frac{a_i}{M_{1i}}$, c'est-à-dire la proportion de sujets exposés parmi les cas dans la strate i:

$$= \Sigma \left(\frac{M_{1i}}{\Sigma M_{1i}} \cdot \frac{a_i}{M_{1i}} \cdot \frac{RR_i - 1}{RR_i} \right)$$

$$= \frac{\Sigma\, a_i (1 - 1/RR_i)}{\Sigma M_{1i}}$$

$$= \frac{\Sigma\, a_i \left(1 - \frac{b_i N_{1i}}{a_i N_{0i}} \right)}{\Sigma M_{1i}}$$

$$= \frac{\Sigma a_i - \Sigma \frac{b_i N_{1i}}{N_{0i}}}{\Sigma M_{1i}}$$

$$= \frac{\left(\Sigma a_i \Big/ \Sigma \frac{b_i N_{1i}}{N_{0i}} \right) - 1}{\Sigma M_{1i} \Big/ \Sigma \frac{b_i N_{1i}}{N_{0i}}}$$

$$= \frac{RR_a - 1}{\frac{\Sigma M_{1i}}{\Sigma a_i} RR_a}$$

$$= \frac{RR_a - 1}{(1/p_{C1})\, RR_a}$$

$$= p_{C1} \left(\frac{RR_a - 1}{RR_a} \right)$$

Dans l'exemple au tableau 16-23 où l'on a $RR_a = 2{,}8$ et $p_{cI} = {}^{140}/_{180}$, on obtient:

$$FE_t \text{ (cohérente)} = ({}^{140}/_{180}) \times \left(\frac{2{,}8 - 1}{2{,}8} \right) = {}^{1}/_{2}.$$

On observe bien la cohérence avec les fractions étiologiques spécifiques puisque $^1/_2$ x 180 = 90 (=10+80).

On doit remarquer que la cohérence est atteinte si la fraction étiologique totale FE_t est exprimée sous la forme

$$p_{C1} \left(\frac{RR_a - 1}{RR_a} \right),$$

mais qu'elle ne l'est plus si elle est exprimée sous la forme

$$\frac{p_1 (RR_a - 1)}{p_1 (RR_a - 1) + 1}.$$

On peut facilement le vérifier à l'aide de notre exemple. En effet,

$$\frac{p_1 (RR_a - 1)}{p_1 (RR_a - 1) + 1} = \frac{{}^{1}/_{2} (2{,}8 - 1)}{{}^{1}/_{2} (2{,}8 - 1) + 1}$$

$$= 0{,}47 \text{ (ou } {}^{9}/_{19}).$$

Cette valeur 0,47 n'est pas cohérente avec les valeurs des FE_{ti} puisque 0,47 x 180 = 85,26 ($\neq$ 10 + 80).

Ajustement des fractions prévenues

Nous allons maintenant aborder la question de l'ajustement des fractions prévenues tant chez les sujets exposés que chez la population totale, en considérant le tableau 16-24 qui décrit les résultats d'une étude sur l'efficacité d'un vaccin *V* contre la maladie *M*. Ces résultats sont stratifiés suivant un certain facteur *F* (disons l'âge).

Tableau 16-24

Strate 1 (F_1)

M \ V	+	—
+	100	100
—	1000	500

$Ef_1 = 2$

Strate 2 (F_2)

M \ V	+	—
+	100	600
—	500	1000

$Ef_2 = 3$

Global

M \ V	+	—
+	200	700
—	1500	1500

$Ef_b = 3{,}5$

AJUSTEMENT DE LA FRACTION PRÉVENUE CHEZ LES SUJETS EXPOSÉS: FP_1

Au tableau 16-24, pour chacune des deux strates et globalement, les fractions prévenues chez les

$$FP_{11} = \frac{2-1}{2} = 0{,}50 \text{ (ou } ½)$$

$$FP_{12} = \frac{3-1}{3} = 0{,}67 \text{ (ou } ⅔)$$

$$FP_1 = \frac{3{,}5-1}{3{,}5} = 0{,}71 \text{ (ou } ⅝)$$

Pour l'interprétation de ces fractions prévenues, établissons que:

— le nombre de cas observés ou apparus parmi les vaccinés est donné par a (par exemple, dans la première strate, $a = 100$);

— le nombre de cas potentiels est le nombre de cas qui seraient apparus chez les vaccinés s'il n'y avait pas eu vaccination. Avec a cas observés parmi les vaccinés et un vaccin d'une efficacité *Ef,* le nombre de cas potentiels est bien sûr donné par $a \cdot Ef$

ou de façon équivalente par $\frac{a}{1 - FP_1}$

— le nombre de cas prévenus ou évités chez les vaccinés parmi les cas potentiels est évidemment donné par la différence entre le nombre de cas potentiels et le nombre de cas observés, soit $a \cdot Ef - a$ c'est-à-dire $a(Ef - 1)$ ou de façon équivalente par $\frac{a}{1 - FP_1} \cdot FP_1$

Remarquons que le nombre de cas potentiels est égal au nombre de cas observés plus le nombre de cas évités.

Dans la première strate, parmi les 200 $(= \frac{100}{1 - ½})$ cas potentiels, il y en a 100 qui sont apparus et 100 $(= \frac{100}{1 - ½} \times ½)$ qui ont pu être prévenus par l'action du vaccin.

Dans la deuxième strate, parmi les 300 ($= \frac{100}{1 - 2/3}$) cas potentiels, il y en a 100 qui sont apparus et 200 ($= \frac{100}{1 - 2/3} \times 2/3$) qui ont pu être prévenus par l'action du vaccin.

Globalement, parmi les 700 ($= \frac{200}{1 - 5/7}$) cas potentiels, il y en a 200 qui sont apparus et 500 ($= \frac{100}{1 - 5/7} \times 5/7$) qui ont pu être prévenus par l'action du vaccin.

L'interprétation de FP_1 souffre d'une incohérence. A partir de ce qui est observé sur les strates, le nombre total de cas potentiels chez les vaccinés devrait être de 500 (= 200 + 300) et non pas de 700. Le nombre de cas évités par l'action du vaccin devrait être de 300 (= 100 + 200) et non de 500. Ces incohérences sont induites en partie par le fait que la fraction prévenue globale FP_1 est calculée à partir d'une mesure Ef biaisée (le $Ef_b = 3{,}5$).

Le type d'ajustement cohérent est celui dont le système de poids est basé sur la distribution du facteur confondant parmi les cas potentiels vaccinés. Dans l'exemple au tableau 16-24 où cette distribution est (200; 300), le système de poids correspondant est de ($^{200}/_{500}$; $^{300}/_{500}$). On obtient comme ajustement:

$$FP_1 = (^2/_5)\, FP_{11} + (^3/_5)\, FP_{12}$$
$$= {}^2/_5 \times 0{,}50 + {}^3/_5 \times 0{,}67 = 0{,}60 \text{ (ou } {}^3/_5).$$

Cette dernière fraction prévenue chez la totalité des cas exposés est, dans son interprétation, cohérente avec les fractions prévenues spécifiques. En effet, elle conduit à une estimation de 300 cas prévenus ($3/5 \times 500$). De façon générale, une fraction prévenue chez les exposés cohérente est donnée par la formule:

$$FP_1 \text{ (cohérente)} = \Sigma \frac{a_i Ef_i}{\Sigma a_i Ef_i} FP_{1i}$$

où $a_i Ef_i$ représente bien chez les exposés le nombre de cas potentiels pour chacune des strates.

Il est intéressant de remarquer que lorsqu'on utilise la mesure ajustée Ef_a, c'est-à-dire $\frac{1}{RR_a}$, la fraction prévenue FP_1 qui en résulte est cohérente. Vérifions-le pour notre exemple.

$$Ef_a = \frac{1}{RR_a}$$

$$= \frac{\Sigma \frac{b_i N_{1i}}{N_{0i}}}{\Sigma a_i}$$

$$= \frac{\frac{100 \times 1000}{500} + \frac{600 \times 500}{1000}}{100 + 100}$$

$$= 2{,}5.$$

Cet ajustement engendre la fraction prévenue $\frac{2{,}5 - 1}{2{,}5}$ (= 0,60 ou 3/5). Comme on vient de le voir, cette fraction prévenue FP_1 est cohérente avec les fractions prévenues spécifiques FP_{1i}.

La fraction cohérente FP_1 satisfait la relation:

$$FP_1 = \frac{Ef_a - 1}{Ef_a}.$$

Cette relation montre que l'ajustement de Ef_a conduit à une fraction prévenue FP_1 cohérente.

On peut facilement démontrer la relation en notant d'abord que:

$$\frac{FP_1}{1 - FP_1} = Ef - 1$$

$$\text{et,} \quad \frac{1}{1 - FP_1} = Ef$$

En voici la démonstration.

$$FP_1 = \Sigma \left(\frac{a_i Ef_i}{\Sigma a_i Ef_i} \right) FP_{1i}$$

$$= \Sigma \frac{a_i Ef_i}{\Sigma a_i Ef_i} \cdot \frac{(Ef_i - 1)}{Ef_i}$$

$$= \frac{\Sigma a_i Ef_i - \Sigma a_i}{\Sigma a_i Ef_i}$$

$$= \frac{\Sigma \frac{b_i N_{1i}}{N_{0i}} - \Sigma a_i}{\Sigma \frac{b_i N_{1i}}{N_{0i}}}$$

$$= \frac{\frac{\Sigma b_i N_{1i} / N_{0i}}{\Sigma a_i} - 1}{\frac{\Sigma b_i N_{1i} / N_{0i}}{\Sigma a_i}}$$

$$= \frac{Ef_a - 1}{Ef_a}.$$

La démonstration de la cohérence de l'ajustement de Ef_a, c'est-à-dire $\frac{1}{RC_a}$, avec la fraction prévenue FP_1 dans les études cas-témoins est analogue. On n'a qu'à substituer dans la démonstration c_i à N_{1i} et d_i à N_{0i}.

AJUSTEMENT DE LA FRACTION PRÉVENUE TOTALE : FP_t

Utilisons encore ici les données du tableau 16-24. Pour des raisons qui vont se clarifier par la suite, nous allons utiliser uniquement l'expression :

$$\frac{p_{c1}(Ef - 1)}{p_{c1}(Ef - 1) + 1}$$

pour exprimer la fraction prévenue totale FP_t.

Pour chacune des deux strates et globalement, les fractions prévenues totales sont:

$$FP_{t1} = \frac{\frac{1}{2}(2 - 1)}{\frac{1}{2}(2 - 1) + 1} = 0{,}33 \ (\tfrac{1}{3})$$

$$FP_{t2} = \frac{\frac{1}{7}(3 - 1)}{\frac{1}{7}(3 - 1) + 1} = 0{,}22 \ (\tfrac{2}{9})$$

$$FP_t = \frac{\frac{2}{9}(3{,}5 - 1)}{\frac{2}{9}(3{,}5 - 1) + 1} = 0{,}36 \ (\tfrac{5}{14})$$

Pour l'interprétation de ces fractions prévenues, établissons que

— le nombre de cas observés est donné par M_1 (par exemple, dans la première strate, $M_1 = 200$)

— le nombre de cas potentiels est estimé par $\frac{M_1}{1 - FP_t}$ ou de façon équivalente par $a(Ef - 1) + M_1$;

— le nombre de cas prévenus ou évités est estimé par la différence entre le nombre de cas potentiels et le nombre de cas observés, soit $\frac{M_1}{1 - FP_t} - M_1$, c'est-à-dire $[a(Ef - 1) + M_1] FP_t$ ou de façon équivalente par $a(Ef - 1)$.

Remarquons qu'encore ici le nombre de cas potentiels est égal à la somme des cas observés et des cas évités.

Suivant ces relations, on interprète, pour chacune des strates et globalement, les fractions prévenues FP_t de la façon suivante:

Dans la première strate, parmi les 300 $\left(= \frac{200}{1 - 1/3}\right)$ cas potentiels, il y en a 200 qui sont apparus et 100 $\left(= \frac{100}{1 - 1/3} \times 1/3\right)$ qui ont pu être prévenus par l'action du vaccin.

Dans la deuxième strate, parmi les 900 $\left(= \frac{700}{1 - 2/9}\right)$ cas potentiels, il y en a 700 qui sont apparus et 200 $\left(= \frac{700}{1 - 2/9} \times 2/9\right)$ qui ont pu être prévenus par l'action du vaccin.

Globalement, il y a 1400 cas $\left(= \frac{900}{1 - 5/14}\right)$ potentiels; 900 sont apparus et 500 $\left(= \frac{100}{1 - 5/14} \times 5/14\right)$ ont pu être prévenus ou évités par l'action du vaccin.

L'interprétation de FP_t souffre d'une incohérence. A partir de ce qui est observé sur les strates, le nombre total de cas potentiels devrait être de 1200 (= 300 + 900) au lieu de 1400. Le nombre de cas prévenus par l'action du vaccin devrait être de 300 (= 100 + 200) et non de 500. Ces incohérences résultent en partie du fait que la fraction prévenue globale FP_t est calculée à partir d'une mesure Ef biaisée.

Le type d'ajustement qui conduit à la cohérence est celui dont le système de poids est basé sur la distribution du facteur confondant parmi les cas potentiels. En effet, dans l'exemple, cette distribution est (300; 900); le système de poids correspondant est ($^{300}/_{1200}$; $^{900}/_{1200}$). On a comme ajustement,

$$FP_t = (^{300}/_{1200})\, FP_{t1} + (^{900}/_{1200})\, FP_{t2}$$

$$= 1/4 \times 0{,}33 + 3/4 \times 0{,}22 = 0{,}25 \text{ (ou } 1/4).$$

Cette dernière fraction prévenue chez la totalité des cas est cohérente, dans son interprétation, avec les fractions prévenues spécifiques. En effet, elle donne une estimation de 300 cas évités sur 1200 cas potentiels ($^1/_4$ x 1200).

De façon générale, une fraction prévenue totale cohérente est donnée par la formule :

$$FP_t \text{ (cohérente)} = \frac{\Sigma\,[a_i\,(Ef_i - 1) + M_{1i}]\,FP_{ti}}{\Sigma\,[a_i\,(Ef_i - 1) + M_{1i}]}$$

où $[a_i\,(Ef_i - 1) + M_{1i}]$ représente le nombre de cas potentiels sur chacune des strates.

Il est intéressant de remarquer que lorsque l'on utilise la mesure ajustée Ef_a, la fraction prévenue FP_t qui en résulte est cohérente. Vérifions-le pour notre exemple.

Nous rappelons que Ef_a = 2,5. Cet ajustement engendre la fraction prévenue

$$\frac{2/9\,(2{,}5 - 1)}{2/9\,(2{,}5 - 1) + 1} (= 0{,}25 \text{ ou } 1/4).$$

Comme on vient de le voir, cette fraction prévenue FP_t est cohérente avec les fractions prévenues spécifiques FP_{ti}.

Cette relation qui existe entre l'ajustement du Ef_a et la cohérence de la fraction étiologique FP_t peut facilement être démontrée.

Nous savons que

$$FP_t \text{ (cohérente)} = \frac{\Sigma\,[a_i(Ef_i - 1) + M_{1i}]\,FP_{ti}}{\Sigma\,[a_i(Ef_i - 1) + M_{1i}]}.$$

et que

$$[a_i(Ef_i - 1) + M_{1i}]\,FP_{ti} = a_i(Ef_i - 1).$$

Ainsi,

$$FP_t = \frac{\Sigma\, a_i(Ef_i - 1)}{\Sigma\,[a_i(Ef_i - 1) + M_{1i}]}$$

$$= \frac{\Sigma\, a_i Ef_i - \Sigma a_i}{\Sigma\, a_i Ef_i - \Sigma a_i + M_1}$$

$$= \frac{\Sigma\, b_i N_{1i}/N_{0i} - \Sigma a_i}{\Sigma\, b_i N_{1i}/N_{0i} - \Sigma a_i + M_1}$$

$$= \frac{\dfrac{\Sigma b_i N_{1i}/N_{0i}}{\Sigma a_i} - \dfrac{\Sigma a_i}{\Sigma a_i}}{\dfrac{\Sigma b_i N_{1i}/N_{0i}}{\Sigma a_i} - \dfrac{\Sigma a_i}{\Sigma a_i} + \dfrac{M_1}{\Sigma a_i}}$$

$$= \frac{Ef_a - 1}{Ef_a - 1 + \dfrac{M_1}{\Sigma a_i}}$$

$$= \frac{p_{C1}(Ef_a - 1)}{p_{C1}(Ef_a - 1) + 1}, \text{ puisque } \frac{M_1}{\Sigma a_i} = \frac{1}{p_{C1}}$$

La démonstration de la cohérence de l'ajustement Ef_a avec la fraction prévenue FP_f dans les études cas-témoins est analogue. On n'a qu'à substituer dans la démonstration c_i à N_{1i} et di à N_{oi}.

On doit remarquer que la cohérence est atteinte si la fraction prévenue est exprimée à l'aide de Ef_a sous la forme

$$\frac{p_{C1}(Ef_a - 1)}{p_{C1}(Ef_a - 1) + 1}$$

mais ne l'est plus si elle est présentée sous la forme

$$\frac{Ef_a - 1}{Ef_a} \cdot p_1$$

On peut facilement le vérifier à l'aide de notre exemple. Rappelons que $Ef_a = 2{,}5$ *et* $p_1 = {}^1/_2$. Ainsi,

$$FP_t = \frac{(Ef_a - 1)}{Ef_a} \cdot p_1$$

$$= \frac{(2{,}5 - 1)}{2{,}5} \times \tfrac{1}{2} = 0{,}30\ (= {}^3/_{10}).$$

Cette valeur 0,30 n'est pas cohérente avec les valeurs des FP_{ti} puisqu'elle donne 360(= 0,30 x 1200) cas évités plutôt que 300.

RÉSUMÉ

Une mesure ajustée est une somme pondérée de mesures spécifiques. L'ajustement est la construction d'une mesure ajustée. Pour les mesures de fréquence, l'ajustement est généralement pratiqué en vue d'établir des comparaisons. Dans ce cas, les mesures à comparer sont ajustées sur un même système de poids; elles sont alors dites standardisées. Pour les mesures de fréquence, le système de poids est généralement défini à partir d'une population-type (distribution-type). L'ajustement des mesures *RR* et *RC* permet de résumer en une mesure d'association globale les mesures spécifiques d'association. Les procédures d'ajustement se regroupent suivant que les poids sont définis à partir d'une distribution-type ou bien à partir du critère de la précision. Dans la première catégorie, on trouve pour le *RR* les ajustements du type RR_a (ou SMR) ou RR_s

et pour le *RC* les ajustements du type RC_a et RC_S. La seconde catégorie regroupe les ajustements basés sur le critère de la précision des mesures spécifiques. Les poids sont alors définis en fonction de l'inverse de la variance. L'ajustement de Mantel-Haenszel constitue une méthode qui relève à la fois des deux catégories. À l'exception du RR_S (ou RC_s), les ajustements décrits ne conduisent pas à des mesures standardisées. Ces mesures ajustées ne se prêtent pas à des comparaisons internes. Le type d'ajustement approprié aux mesures d'impact est celui qui rend cohérente l'interprétation de la mesure globale avec celle des mesures spécifiques. Pour la fraction étiologique ou la fraction prévenue chez les sujets exposés, le type d'ajustement cohérent est celui dont le système de poids est basé sur la distribution du facteur confondant chez les cas exposés. Pour la fraction étiologique ou la fraction prévenue totale, le type d'ajustement cohérent est celui basé sur la distribution du facteur confondant chez tous les cas.

Symboles

(λ_i) : système de poids

(w_i) : distribution quelconque

V_i : variance sur la strate *i*

R_1, *Ro*, R_t*:* risque chez les sujets exposés, risque chez les sujets non-exposés, risque dans la population totale

p_1 : proportion de sujets exposés dans la population totale

p_{cI}*:* proportion de sujets exposés parmi les cas.

RR_i, RC_i : risque relatif spécifique à la strate *i*, rapport des cotes spécifique à la strate *i*

RR_{ajus}, RC_{ajus}: risque relatif ajusté, rapport des cotes ajusté

RR_a, RC_a*:* risque relatif ajusté, rapport des cotes ajusté, chacun par la méthode de la distribution-type

RR_s, *RCs :* risque relatif standardisé, rapport des cotes standardisé, chacun par la méthode de la distribution-type

SMR: « standardized mortality ratio »

RR_v, RC_v*:* risque relatif ajusté, rapport des cotes ajusté, chacun par un critère de précision (variance)

RR_{MH}, RC_{MH}*:* risque relatif ajusté de Mantel-Haenszel, rapport des cotes ajusté de Mantel-Haenszel

FE_I (cohérente), FE_{Ii}*:* fraction étiologique (cohérente) chez les sujets exposés, chez les sujets exposés spécifique à la strate *i*

FE_t (cohérente), FE_{ti}*:* fraction étiologique (cohérente) totale, spécifique à la strate *i*

Ef : mesure d'efficacité d'un facteur préventif

Ef_a*:* mesure d'efficacité ajustée suivant une méthode analogue à celle du RR_a

FP_1 (cohérente), FP_{Ii}*:* fraction prévenue (cohérente) chez les sujets exposés, chez les sujets exposés spécifique à la strate *i*

FP_t (cohérente), FP_{ti}*:* fraction prévenue (cohérente) totale, spécifique à la strate *i*.

Formules

$$\lambda_i = \frac{w_i}{\Sigma w_i}, \qquad \lambda_i = \frac{1/V_i}{\Sigma 1/V_i}$$

$$RR_{\text{ajus}} = \Sigma \lambda_i RR_i$$

$$RC_{\text{ajus}} = \Sigma \lambda_i RC_i$$

$$RR_a = \frac{\Sigma a_i}{\Sigma b_i N_{1i}/N_{0i}}$$

$$RC_a = \frac{\Sigma a_i}{\Sigma \dfrac{b_i c_i}{d_i}}$$

$$RR_s = \frac{\Sigma a_i N_{0i}/N_{1i}}{\Sigma b_i}$$

$$RC_s = \frac{\Sigma \dfrac{a_i d_i}{c_i}}{\Sigma b_i}$$

$$\text{SMR} = \frac{a \text{ (nombre observé)}}{A \text{ (nombre attendu)}} = \frac{R_1}{R_t}$$

où $R_t = p_1 R_1 + (1 - p_1) R_0$

$$\text{SMR} = \frac{RR}{p_1 (RR - 1) + 1}$$

$$RR_v = e^{\left(\frac{\Sigma w_i \ln RR_i}{\Sigma w_i}\right)}$$

où, si RR est un rapport de taux d'incidence,

$$w_i = \frac{N_{1i} N_{0i} M_{1i}}{N_i^2} \text{ ou } = \frac{a_i b_i}{a_i + b_i}$$

où, si RR est un rapport de proportions, d'incidences cumulatives,

$$w_i = \frac{N_{1i} N_{0i} M_{1i}}{M_{0i} N_i} \text{ ou } = \frac{a_i b_i N_{1i} N_{0i}}{a_i d_i N_{1i} + b_i c_i N_{0i}}$$

$$RC_v = e^{\left(\frac{\Sigma w_i \ln RC_i}{\Sigma w_i}\right)}$$

où $w_i = \dfrac{1}{\dfrac{1}{a_i} + \dfrac{1}{b_i} + \dfrac{1}{c_i} + \dfrac{1}{d_i}}$.

$$RR_{\text{MH}} = \frac{\Sigma a_i N_{0i}/N_i}{\Sigma b_i N_{1i}/N_i}.$$

$$RC_{\text{MH}} = \frac{\Sigma a_i d_i/N_i}{\Sigma b_i c_i/N_i}.$$

$RR_{\text{MH}} = \dfrac{r + t}{r + s}$, en analyse appariée

$RC_{\text{MH}} = \dfrac{t}{s}$, en analyse appariée

$$FE_1 \text{ (cohérente)} = \Sigma \frac{a_i}{\Sigma a_i} FE_{1i} = \frac{RR_a - 1}{RR_a}.$$

$$FE_t \text{ (cohérente)} = \Sigma \frac{M_{1i}}{\Sigma M_{1i}} FE_{1i} = p_{c1} \frac{RR_a - 1}{RR_a}$$

$$FP_1 \text{ (cohérente)} = \Sigma \frac{a_i Ef_i}{\Sigma a_i Ef_i} FP_{1i} = \frac{Ef_a - 1}{Ef_a}.$$

$$FP_t \text{ (cohérente)} = \Sigma \frac{[a_i (Ef_i - 1) + M_{1i}]}{\Sigma [a_i (Ef_i - 1) + M_{1i}]} FP_{ti}$$

$$= \frac{p_{c1} (Ef_a - 1)}{p_{c1} (Ef_a - 1) + 1}$$

LECTURE SUGGÉRÉE

KLEINBAUM, D., KUPPER L.L. et MORGENSTERN. H. ***Epidemiologic Research,*** Belmont (USA), Lifetime Learning Publications, 1982, chapitre 17, pp.340-351.

ANNEXE DU CHAPITRE 16

Résumé des principaux types d'ajustement pour les mesures d'association *RR* et *RC*

Nous regroupons ici les différents types d'ajustement pour le *RR* et le *RC* suivant les trois grandes voies d'ajustement. Nous rappelons que celles-ci se caractérisent par la façon de définir les poids: à partir d'une distribution-type, à partir de la précision des estimations, la méthode de Mantel-Haenszel. Pour chacune des méthodes, nous indiquons la mesure, la formule d'ajustement, le fait que la mesure soit standardisée ou non, et un bref commentaire sur l'ajustement.

Poids définis à partir d'une distribution-type

Mesure	*Formule*	*Standardisée*	*Commentaires*
RR_s	$\frac{\Sigma \frac{a_i N_{0i}}{N_{1i}}}{\Sigma b_i}$	Oui	Se prête aux comparaisons internes d'une étude (par exemple, entre plusieurs niveaux d'exposition).
RR_a	$\frac{\Sigma a_i}{\Sigma \frac{b_i N_{1i}}{N_{0i}}}$	Non	Mesure désignée par SMR lorsque le groupe de comparaison est la population entière. Mesure fort utilisée en santé au travail. Elle a l'avantage d'une bonne stabilité statistique.
SMR	$\frac{a}{A} = \frac{\Sigma a_i}{\Sigma A_i}$	Non	Mesure qui correspond directement à la mesure RR_S des études de cohorte.
RC_s	$\frac{\Sigma \frac{a_i d_i}{c_i}}{\Sigma b_i}$	Oui	Mesure qui correspond directement à la mesure du RR_a des études de cohorte. Peu utilisée.
RC_a	$\frac{\Sigma a_i}{\Sigma \frac{b_i c_i}{d_i}}$	Non	

Poids définis à partir du critère de la précision

Mesure	*Formule*	*Standardisée*	*Commentaires*
RR_V	$e^{\left(\frac{\Sigma w_i \ln RR_i}{\Sigma w_i}\right)}$.	Non	Ajustement valable pour le *RR* où w_i représente l'inverse de la variance du $\ln RR_i$:
RC_V	$e^{\left(\frac{\Sigma w_i \ln RC_i}{\Sigma w_i}\right)}$	Non	Ajustement valable pour le *RC* où w_i représente l'inverse de la variance du $\ln RC_i$:

Ajustement de Mantel-Haenszel

Mesure	*Formule*	*Standardisée*	*Commentaires*
RR_{MH}	$\dfrac{\Sigma \frac{a_i N_{0i}}{N_i}}{\Sigma \frac{b_i N_{1i}}{N_i}}$	Non	Peu utilisée dans les études de cohorte.
RC_{MH}	$\dfrac{\Sigma \frac{a_i d_i}{N_i}}{\Sigma \frac{b_i c_i}{N_i}}$	Non	Très utilisée dans les études cas-témoins. Utilisable même si, dans la stratification, certains tableaux présentent des cellules nulles.

Ajustement dans l'analyse appariée

Mesure	*Formule*	*Standardisée*	*Commentaires*
RR_{MH}	$\dfrac{r+t}{r+s}$	Non	Correspond à l'ajustement par la méthode de Mantel-Haenszel adaptée à la mesure du *RR*.
RC_{MH}	$\dfrac{t}{s}$	Non	Correspond à l'ajustement par la méthode de Mantel-Haenszel pour le *RC*.

CHAPITRE 17

Intervalle de confiance

Le présent chapitre insiste d'abord sur le fait que la méthode de l'intervalle de confiance, ou estimation par intervalle, prolonge celle de l'estimation ponctuelle en tentant de cerner la valeur inconnue d'un paramètre. La moyenne est le paramètre retenu pour introduire l'idée d'intervalle de confiance. Le chapitre donne en approximation normale les intervalles de confiance de plusieurs paramètres: moyenne, médiane, proportion, taux, risque relatif, SMR, rapport des cotes, coefficient de corrélation linéaire, coefficient de corrélation intra-classe, mesure d'accord kappa et probabilité cumulative de survie.

Quand une étude, pour des raisons pratiques, est menée auprès d'un échantillon plutôt que sur la population entière, les résultats conduisent plutôt à une estimation qu'à la vraie valeur du paramètre. Que le paramètre étudié soit un risque relatif, une moyenne, une mesure d'impact ou autre, son estimation est généralement effectuée sur une échelle quantitative continue et compte ainsi une infinité de valeurs possibles. La valeur estimée ou observée d'un paramètre apparaît comme un point sur une droite numérique qui rassemble toutes les valeurs possibles d'estimation, y compris la valeur vraie du paramètre.

La valeur estimée est en quelque sorte une *estimation ponctuelle* qui s'avère à elle seule une donnée insuffisante ou incomplète pour décrire correctement le paramètre que l'on cherche à cerner. L'estimation ponctuelle permet de déterminer, en termes plus ou moins vagues, l'ordre de grandeur du paramètre. En l'absence de tout biais, quelle confiance peut-on accorder à une telle estimation? Pour répondre à cette question de fiabilité, on propose de construire, autour de la valeur estimée, un intervalle ayant une probabilité prédéterminée de recouvrir la valeur paramétrique. C'est l'intervalle *de confiance;* il permet de cadrer, avec une certaine confiance, un ensemble de valeurs susceptibles de contenir la vraie valeur (inconnue) du paramètre. La largeur de l'intervalle reflète la précision avec laquelle a été estimé le paramètre.

En pratique dans l'*estimation par intervalle,* l'investigateur fixe la probabilité qui l'intéresse. La connaissance de celle-ci précède le calcul de l'intervalle. La probabilité la plus souvent retenue est 0,95 mais rien n'empêche de choisir d'autres valeurs, comme 0,99 ou même 0,90.

Quelle que soit la probabilité adoptée, elle détermine le *niveau de confiance* que l'on accorde à un intervalle sur son aptitude à contenir la valeur vraie d'un paramètre. Plus spécifiquement, si l'on a convenu d'une probabilité égale à 0,95 (ou 95 %), on parle d'intervalle de confiance à 0,95 ou au niveau 0,95 (95 %).

La recherche d'un intervalle ayant de bonnes chances (90 %, 95 %, 99 %) et non toutes les chances (100 %) de recouvrir la vraie valeur d'un paramètre peut surprendre. L'intervalle qui recouvre à coup sûr (niveau 100 %) la tension artérielle moyenne vraie s'étend théoriquement de 0 à l'infini. C'est un intervalle trop peu spécifique de la tension artérielle pour présenter un quelconque intérêt puisqu'il s'applique à n'importe quel paramètre à valeurs positives.

Dans ce qui suit, nous présentons des formules d'intervalles de confiance pour un certain nombre de paramètres (mesures) définis dans les chapitres précédents. Nous ne donnons que des intervalles approximatifs. Ils suffisent à faire comprendre le concept d'intervalle de confiance. De plus, leurs calculs sont relativement simples contrairement en général à ceux des intervalles exacts. La qualité de l'approximation diminue toutefois avec le nombre de sujets, c'est-à-dire la taille de l'échantillon. L'approximation appliquée ici repose sur l'utilisation du modèle de la distribution normale. Pour qui sent le besoin de connaître cette distribution, elle est présentée dans l'annexe de ce chapitre.

Nous commençons par la présentation de l'intervalle de confiance d'une moyenne arithmétique. Nous utilisons ce paramètre pour expliquer, avec un certain souci du détail, les idées qui fondent la construction de tout intervalle

de confiance en approximation normale. Ces idées qui s'appliquent à d'autres paramètres sont essentiellement qu'une estimation ponctuelle d'un paramètre est une valeur d'une variable dont la distribution est proche du modèle normal.

ESTIMATION PAR INTERVALLE D'UNE MOYENNE (ARITHMÉTIQUE)

Il est d'abord important de souligner le fait que la moyenne m ou ma mesurée sur un échantillon (on dit aussi moyenne d'échantillon ou échantillonnale) change d'un échantillon à un autre. Au chapitre 3, on a obtenu pour la tension artérielle une moyenne m égale à 150 mmHg avec un échantillon de 121 patients. Un autre échantillon de 121 patients aurait pu donner une autre moyenne *m,* par exemple 154 mmHg. Il *y* a variation des moyennes échantillonnales *m*. La moyenne échantillonnale est une variable. Ce fait est propre aux estimations ponctuelles de n'importe quel paramètre. On peut alors se poser trois questions sur les estimations m du paramètre.

— Autour de quelle valeur fluctuent les moyennes *m*?

— Quelle est la dispersion des moyennes *m*?

— Quelle est la distribution des moyennes *m*?

Les réponses à ces questions sont essentielles au calcul d'un intervalle de confiance pour la moyenne; elles le sont aussi pour tout autre paramètre. Nous allons successivement répondre à ces trois questions pour la moyenne.

Moyenne des moyennes échantillonnales

La moyenne M_m des moyennes m mesurées sur tous les échantillons possibles de taille *n* tirés d'une population, est égale à la moyenne μ de la population, c'est-à-dire à la moyenne vraie. En des termes moins déroutants, les moyennes échantillonnales m fluctuent autour de la moyenne *p*. de la population. Pour s'en convaincre, réfléchissons à partir d'un exemple simple. Considérons une population de quatre personnes ($N = 4$) : *A, B, C, D,* dont le poids respectif est: 70, 71, 73, 78 kg. Le poids moyen μ de la population est de 73 kg. Énumérons tous les échantillons possibles de taille 2 (2 personnes distinctes). Leur composition figure au tableau 17-1 avec leur moyenne respective *m*.

Tableau 17-1

Échantillon	Données	Moyenne *m*
A,B	70; 71	70,5
A,C	70;73	71,5
A,D	70;78	74
B,C	71; 73	72
B,D	71;78	74,5
C,D	73;78	75,5

La moyenne M_m de toutes les moyennes *m* est égale à :

$$M_m = \frac{70{,}5 + 71{,}5 + \ldots + 75{,}5}{6}$$

$$= \frac{438}{6}$$

$$= 73 \text{ kg}.$$

Ainsi $M_m = \mu$. Ce résultat, toujours vrai quelles que soient les tailles de la population et de

l'échantillon, montre bien que les moyennes m fluctuent autour de μ.

Erreur-type de la moyenne échantillonnale

S'il y a variation des moyennes échantillonnales m, il y a lieu de s'interroger sur leur dispersion. Comme à l'accoutumé, nous mesurons par un écart-type la dispersion ou variation des *m*. L'usage veut qu'on appelle cet écart-type l'erreur-type de la moyenne, dénotée σ*m*. L'erreur-type nous renseigne sur la précision de la mesure. Pour les données du tableau 17-1, on trouve :

$$\sigma_m = \sqrt{\frac{\Sigma (m - \mu)^2}{6}}$$

$$= 1{,}7795$$

Ce calcul n'est pas très commode puisqu'en pratique on ne disposera que d'un échantillon, en conséquence d'une seule valeur m. Heureusement, la théorie statistique vient à notre rescousse pour nous proposer la relation suivante qui est valide dans tous les cas:

$$\sigma_m = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N - n}{N - 1}} \qquad [1]$$

où σ est l'écart-type de la variable étudiée (ici le poids),

n est la taille de l'échantillon,

N est la taille de la population.

Dans notre exemple sur le poids des quatre personnes, où

$$\sigma = \sqrt{\frac{\Sigma (x - \mu)^2}{N}} = 3{,}0822$$

$$n = 2$$

$$N = 4$$

On vérifie bien la relation [1]:

$$\sigma_m = \frac{3{,}0822}{\sqrt{2}} \sqrt{\frac{4 - 2}{4 - 1}}$$

$$= 1{,}7795$$

La relation [1] peut être simplifiée lorsque la population est infinie ($N = \infty$). Dans ce cas, le facteur $\sqrt{\frac{N - n}{N - 1}}$ est toujours égal à 1. En effet:

$$\frac{N - n}{N - 1} = \frac{1 - n/N}{1 - 1/N} = \frac{1 - 0}{1 - 0} = 1.$$

La formule pour l'erreur-type σ_m se réduit alors à:

$$\sigma_m = \frac{\sigma}{\sqrt{n}} \qquad [2]$$

Lorsque la taille de l'échantillon est relativement petite par rapport à celle de la population, la relation [2] remplace aussi, mais en approximation, la relation [1]. Par exemple, si $n = 3000$ et $N = 1000\,000$, on a :

$$\sqrt{\frac{N - n}{N - 1}} = \sqrt{\frac{1000\,000 - 3000}{1000\,000 - 1}} = 0{,}9985.$$

Dans ce cas,

$$\sigma_m \simeq \frac{\sigma}{\sqrt{n}}$$

Les relations [1] ou [2] entre la taille de l'échantillon et l'erreur-type sont très intéressantes. On observe que σ_m diminue lorsque n augmente. Ce que l'on savait par intuition se confirme. N'est-ce pas que les moyennes échantillonnales varient moins d'un échantillon à l'autre si les échantillons comprennent plus de sujets? La moyenne m s'approche vraisemblablement de lorsque l'erreur-type de la moyenne m diminue. A la limite, si l'on prélevait tous les sujets qui composent la population, on aurait $n = N$ et, conséquemment $\sigma_m = 0$. L'erreur d'échantillonnage s'éclipse totalement puisqu'en prenant toute la population, il n'y a plus d'échantillonnage.

Distribution de la moyenne échantillonnale

La moyenne m mesurée sur un échantillon est une variable. Sa valeur varie d'un échantillon à l'autre. On peut donc parler de la distribution (d'échantillonnage) des moyennes m. Un théorème, nommé *théorème de la limite centrale* permet d'affirmer que si la taille de l'échantillon est suffisamment grande, les moyennes échantillonnales m suivent une distribution assez voisine de la distribution normale. Plus n est grand, plus la distribution d'échantillonnage des m s'approche du modèle normal. Si la variable étudiée x est elle-même de distribution normale, la distribution d'échantillonnage des m est normale peu importe la taille d'échantillon n. Énoncé sous forme symbolique, ce théorème dit que :

$$m \sim \mathcal{N}(\mu, \sigma_m^2)$$

c'est-à-dire que m suit approximativement (~) une distribution normale (N) de moyenne p. et de variance (σ^2_m). Lorsque la variable x est elle-même normale, la variable m l'est aussi, sans qu'il soit nécessaire de penser en termes d'approximation. La courbe en cloche (figure 17-1) caractérise donc la distribution d'échantillonnage de m.

En situation d'approximation, la distribution d'échantillonnage de la moyenne m obéit raisonnablement au modèle normal pourvu que la taille n de l'échantillon ne soit pas trop petite. En pratique, la conformité au modèle normal est satisfaisante dès lors que n dépasse 30.

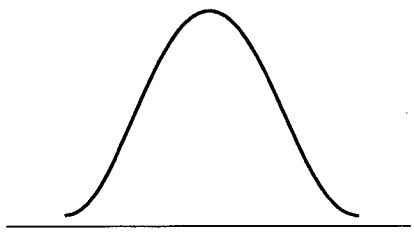
Intervalle de confiance pour une moyenne

Nous venons d'apprendre que m suit (approximativement) une distribution normale de moyenne p. et d'erreur-type σ_m. Si on soustrait p. de la variable m et que l'on divise ensuite la différence par σ_m, on obtient la nouvelle variable :

$$Z = \frac{m - \mu}{\sigma_m}$$

Cette variable Z suit (approximativement) une distribution normale elle aussi, mais cette fois de moyenne 0 et de variance 1.

Figure 17-1



C'est la *distribution normale centrée réduite* (expliquée en annexe à ce chapitre). On peut écrire :

$$Z = \frac{m - \mu}{\sigma_m} \sim \mathcal{N}(0,1),$$

Le tableau A17-1 qui figure en annexe nous permet alors d'écrire qu'il y a une probabilité de 95,4 % (100 − 4,6 %) que $\frac{m - \mu}{\sigma_m}$ tombe entre −2 et 2, c'est-à-dire :

$$\text{Prob}\left(-2 < \frac{m - \mu}{\sigma_m} < 2\right) = 95{,}4\ \%.$$

On peut aussi dire qu'il y a une probabilité de 95 % que :

$$-1{,}96 < \frac{m - \mu}{\sigma_m} < 1{,}96$$

ce qu'on peut écrire :

$$\text{Prob}\left(-1{,}96 < \frac{m - \mu}{\sigma_m} < 1{,}96\right) = 0{,}95.$$

De façon équivalente, on a :

$$\text{Prob}\,(m - 1{,}96\,\sigma_m < \mu < m + 1{,}96\,\sigma_m) = 0{,}95.$$

Il y a une probabilité de 95 % que l'intervalle qui s'étend de m — 1,96 σ_m à m+ 1,96 σ_m recouvre la valeur inconnue C'est, en approximation normale, l'intervalle de confiance à 95 % pour la moyenne de population μ. Si l'on souhaite un niveau de confiance de 99 %, il suffit de remplacer le coefficient 1,96 par 2,58 (plus exactement 2,576). Pour un niveau de confiance quelconque 1 — α, la formule générale de l'intervalle de confiance (IC) d'une moyenne est :

$$IC\colon\ (m - z_{\alpha/2}\sigma_m\,;\, m + z_{\alpha/2}\sigma_m)$$

où m — zα/2σ_m et m + zα/2σ_m sont respectivement ce qu'on appelle les limites de confiance inférieure et supérieure. On peut écrire :

$$\underline{m} = m - z_{\alpha/2}\sigma_m$$

$$\overline{m} = m + z_{\alpha/2}\sigma_m$$

L'indice α/2 rattaché à z se comprend à partir de la figure 17-2.

En pratique, σ est souvent inconnu, donc aussi σ_m. On remplace alors l'écart-type de population σ par son estimation ponctuelle, soit l'écart-type échantillonnai. Rappelons que :

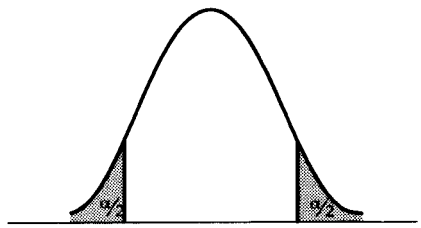
$$s^2 = \frac{\Sigma\,(x_i - m)^2}{n - 1}.$$

Dans l'intervalle de confiance, σ_m sera remplacé à son tour par s_m qui satisfait l'une ou l'autre des deux relations suivantes :

$$s_m = \frac{s}{\sqrt{n}}\sqrt{\frac{N - n}{N - 1}} \qquad \text{(population finie)}$$

$$\text{ou,} \qquad s_m = \frac{s}{\sqrt{n}} \qquad \text{(population infinie ou grande population)}$$

Figure 17-2



Lorsque n dépasse 30, les perturbations causées par la substitution de σ_m par s_m ne sont pas trop fortes et la conformité au modèle de la distribution normale est passablement conservée pour l'expression $\frac{m - \mu}{s_m}$. Ainsi, l'intervalle

$$IC: (m - z_{\alpha/2}s_m;\ m + z_{\alpha/2}s_m)$$

demeure en ce cas un intervalle de confiance au niveau $1 - \alpha$, tout-à-fait acceptable.

Considérons l'exemple de la tension artérielle des 121 patients pour laquelle on a trouvé $m = 150$ mmHg et s = 10,9 mmHg. Si l'échantillon des 121 provient d'une grande population, on a :

$$s_m = \frac{s}{\sqrt{121}} = {}^{10,9}\!/_{11}.$$

L'intervalle de confiance à 95 % pour la tension artérielle moyenne p. est donnée par:

$$IC: (150 - 1{,}96 \times {}^{10,9}\!/_{11};\ 150 + 1{,}96 \times {}^{10,9}\!/_{11})$$

$$: (150 - 1{,}9;\ 150 + 1{,}9)$$

$$: (148{,}1;\ 151{,}9).$$

En l'absence de biais, avec un échantillon représentatif, on peut dire qu'il y a 95 chances sur 100 que l'intervalle qui va de 148,1 à 151,9 mmHg contienne la tension moyenne vraie μ.

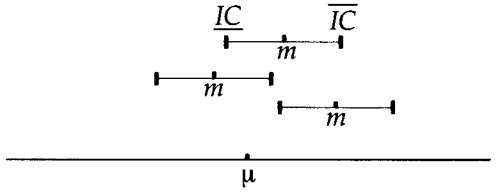
Nous allons énoncer maintenant quelques remarques qui s'appliquent bien entendu à l'intervalle de confiance d'une moyenne, mais aussi à tous les intervalles de confiance pour quelque paramètre que ce soit.

- Référons-nous à l'intervalle de confiance à 95 % que l'on vient de calculer pour la tension artérielle moyenne, soit:

IC: (148,1;151,9).

C'est une erreur de dire qu'il y a une probabilité de 95 % que la tension artérielle moyenne μ de la population se trouve ou tombe dans l'intervalle (148,1; 151,9). Pour une population donnée, la moyenne μ est une constante dans l'esprit de la statistique classique (fréquentiste). La moyenne μ est ou n'est pas dans l'intervalle (148,1; 151,9). Elle ne peut pas s'y trouver avec telle ou telle probabilité. L'interprétation erronée que nous venons de mentionner est très répandue. Elle s'explique principalement par la propension à voir la moyenne μ comme une variable. Ce sont plutôt les limites inférieure et supérieure $\underline{IC}$ et $\overline{IC}$ de l'intervalle de confiance qui varient, comme on tente de l'exprimer à la figure 17-3.

Figure 17-3



La moyenne μ en tant que constante est un point fixe situé quelque part sur la droite numérique alors que les intervalles de confiance sont susceptibles d'occuper des positions différentes suivant l'échantillon prélevé, donc suivant la valeur de m que l'on en tire. Avec un autre échantillon de 121 patients, les limites de confiance $\underline{IC}$ et $\overline{IC}$ seraient probablement différentes de 148,1 et 151,9. L'énoncé probabiliste (ici 0,95) renvoit

plutôt aux chances qu'a l'intervalle défini par les limites de confiance de contenir la moyenne μ. En définitive, quand on dit que (148,1; 151,9) est un intervalle de confiance à 95 %, cela peut-être interprété au sens fréquentiste. Sur 100 échantillons aléatoires de même taille prélevés sur une même population, environ 95 d'entre eux recouvriront la moyenne μ de cette population.

- L'intervalle de confiance d'une moyenne, par son étendue même, définit pour le paramètre un espace de valeurs compatibles avec les données observées. Si, sur la base de données se rapportant à la tension artérielle systolique, on a trouvé (148,1; 151,9) comme intervalle de confiance à 95 % pour la tension moyenne, on peut dire que toute valeur de cet intervalle est une estimation compatible avec les données observées. Si l'intervalle de confiance à 95 % d'un risque relatif est (2,3; 8,1), une valeur 3 pour le *RR* est compatible avec les données; par contre la valeur 1 ne l'est pas. Les données observées sont, pour ainsi dire, cohérentes avec l'idée d'un ris-que relatif supérieur à 1, c'est-à-dire d'une association entre le facteur et la maladie.
- La formule d'intervalle de confiance pour une moyenne:

$$(m - z_{\alpha/2}\sigma_m;\ m + z_{\alpha/2}\sigma_m)$$

qui peut s'écrire de façon plus détaillée:

$$\left(m - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}};\ m + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right)$$

ou,

$$\left(m - z_{\alpha/2}\frac{\sigma}{\sqrt{n}};\ m + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

montre que l'élément $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$ (ou $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$) détermine la largeur de l'intervalle pour une moyenne. On voit que la largeur augmente avec $z_{\alpha/2}$ (donc avec le niveau de confiance 1 — α) et avec l'écart- type σ. Par contre, elle diminue lorsque la taille *n* de l'échantillon augmente. C'est un résultat général bien sûr et qui va selon le bon sens : un intervalle de confiance, pour tout paramètre, devient plus large quand on exige un niveau de confiance plus élevé et plus étroit quand on dispose d'un plus grand nombre de sujets. En l'absence de biais, la largeur d'un intervalle de confiance (l'estimation par intervalle) nous renseigne sur la précision des mesures.

Avant de passer à d'autres paramètres, notons que la forme, décrite pour l'intervalle de confiance de niveau 1 — α d'une moyen- ne, s'applique chaque fois que les estimations (estim.) d'un paramètre se distribuent exactement ou approximativement comme dans le modèle normal. Cette forme est:

$$(\text{estim.} - z_{\alpha/2}\sigma_{\text{estimation}};\ \text{estim.} + z_{\alpha/2}\sigma_{\text{estimation}}).$$

Cette formule est importante puisqu'elle va orienter, en approximation normale, les estimations par intervalle qui suivront.

INTERVALLE DE CONFIANCE POUR UNE MÉDIANE

La distribution d'échantillonnage d'une médiane *(mé)* est proche d'une distribution normale si la variable étudiée x est distribuée normalement et si la taille de l'échantillon est suffisamment grande ($n > 60$). L'intervalle de confiance pour la médiane au niveau $1 - \alpha$ est alors :

$$IC : (mé - z_{\alpha/2}\sigma_{mé};\ mé + z_{\alpha/2}\sigma_{mé})$$

où $\sigma_{mé} \simeq 1{,}253\,\sigma_m$.

On a,

$$\underline{mé} = mé - z_{\alpha/2}\sigma_{mé}$$

$$\overline{mé} = mé + z_{\alpha/2}\sigma_{mé}$$

C'est un intervalle peu intéressant et on lui préférera l'intervalle de confiance de la moyenne. En effet, l'intervalle de confiance de la médiane proposé ici ne s'applique qu'à la médiane d'une variable x distribuée normalement. Or, dans ce cas, la médiane coïncide théoriquement avec la moyenne. Et comme $\sigma_m < \sigma_{mé}$, l'intervalle de confiance pour la moyenne a l'avantage d'être moins large pour un même niveau de confiance.

INTERVALLE DE CONFIANCE POUR UNE PROPORTION

Le symbole *t* est utilisé pour représenter la proportion (ou le pourcentage) d'individus qui, dans une population, présentent une certaine caractéristique : être fumeur par exemple. Le symbole *p* désigne la proportion (ou le pourcentage) observé sur un échantillon.

Calculée sur un échantillon aléatoire, la proportion *p* est une estimation ponctuelle de *n*. Nous pourrions vérifier, comme on l'a fait pour la moyenne observée m, que la moyenne de toutes les estimations *p* est égale à л. L'erreur-type σ_p de la proportion échantillonnale *p* est donnée par les formules suivantes:

$$\sigma_p = \sqrt{\frac{\pi\,(1 - \pi)}{n}}\ \sqrt{\frac{N - n}{N - 1}}\ \text{(population finie)}$$

$$\text{ou, } \sigma_p = \sqrt{\frac{\pi\,(1 - \pi)}{n}}\ \text{(population infinie ou grande population)}$$

La distribution d'échantillonnage d'une proportion *p* n'obéit pas au modèle de la distribution normale. Toutefois, si la taille *n* de l'échantillon est suffisamment grande et si p n'est ni trop près de 0 ni de 1 (plus spécifiquement selon la pratique si $np > 5$ *et* $nq > 5$, où $q = 1 - p$), la distribution de *p* obéit approximativement au modèle normal. Dans ces conditions, on peut écrire:

et l'intervalle de confiance au niveau $1 - \alpha$ pour л est donné par:

$$IC : (p - z_{\alpha/2}\sigma_p;\ p + z_{\alpha/2}\sigma_p)$$

On a,

$$\underline{p} = p - z_{\alpha/2}\sigma_p$$

$$\overline{p} = p + z_{\alpha/2}\sigma_p$$

Il faut toujours en pratique estimer σ_p puisque l'élément л qui entre dans son calcul est toujours inconnu. L'estimation ponctuelle s_p de σ_p est donnée par l'une ou l'autre des deux formules suivantes:

$$s_p = \sqrt{\frac{pq}{n-1}}\ \sqrt{\frac{N-n}{N-1}} \quad \text{(population finie)}$$

$$\text{ou,} \quad s_p = \sqrt{\frac{pq}{n-1}} \quad \text{(population infinie ou grande population)}$$

En pratique donc, si les conditions de l'approximation normale sont vérifiées, un intervalle de confiance acceptable au niveau 1 — α pour la proportion л est donné par l'expression:

$$IC: (p - z_{\alpha/2}s_p;\ p + z_{\alpha/2}s_p).$$

Au cours d'une enquête portant sur 50 282 naissances, on a diagnostiqué 404 cas de malformation cardiaque. La prévalence relative à la naissance de malformation cardiaque est alors estimée à $^{404}/_{50\,282}$ ou 0,008. La prévalence relative *Pr* est une proportion *p.* On a :

$$p = Pr = \frac{404}{50282}$$

et

$$s_p = \sqrt{\frac{pq}{n-1}}$$

$$= \sqrt{\frac{\frac{404}{50282} \times \frac{49878}{50282}}{50281}}$$

$$= 0{,}000398$$

Puisque $np > 5$ *et* $nq > 5$, l'intervalle de confiance à 95 % pour la prévalence relative est donné par:

IC:(0,008 — 1,96 x 0,000398;
0,008 + 1,96 x 0,000398)

:(0,0072; 0,0088)

: (0,007 ; 0,009)

En l'absence de tout biais, il y a environ 95 chances sur 100 que l'intervalle qui s'étend de 0,007 à 0,009 contienne la vraie valeur de la prévalence relative de malformation cardiaque à la naissance.

INTERVALLE DE CONFIANCE POUR UN TAUX

Si $t = a/N$ symbolise un taux observé, donc soumis aux fluctuations d'échantillonnage, un intervalle de confiance approximatif au niveau 1 — α est donné par la formule suivante:

$$IC: \left(t - z_{\alpha/2}\frac{t}{\sqrt{a}};\ t + z_{\alpha/2}\frac{t}{\sqrt{a}}\right)$$

On a,

$$\underline{t} = t - z_{\alpha/2}\frac{t}{\sqrt{a}}$$

$$\bar{t} = t + z_{\alpha/2}\frac{t}{\sqrt{a}}$$

II y a eu au cours d'une année 150 décès dans un groupe de 20 000 personnes. Le taux de décès observé est alors de 7,5 décès par 1000 personnes-années :

$$t = \frac{a}{N} = \frac{150 \text{ décès}}{20\,000 \text{ personnes-années}}\,1000$$

= 7,5 décès par 1000 personnes-années

L'intervalle de confiance approximatif à 95 % pour le taux de décès exprimé par 1000 personnes-années est donné par:

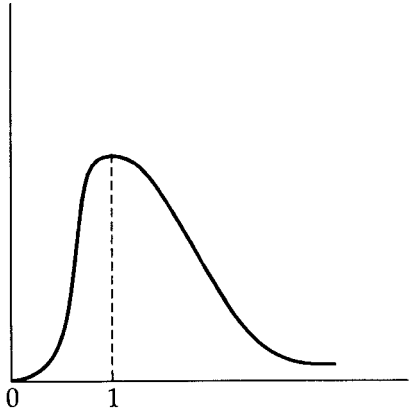
$$IC: \left(7{,}5 - 1{,}96 \times \frac{7{,}5}{\sqrt{150}}; 7{,}5 + 1{,}96 \times \frac{7{,}5}{\sqrt{150}}\right)$$

$$: (6{,}3;\ 8{,}7)$$

INTERVALLE DE CONFIANCE POUR UN RISQUE RELATIF

La distribution d'échantillonnage du risque relatif observé *RR* n'obéit pas au modèle de la distribution normale. Il est assez facile de concevoir qu'une distribution d'échantillonnage du *RR* n'est pas symétrique, mais plutôt asymétrique à droite. Les valeurs observées du *RR* ne peuvent pas être inférieures à 0 mais peuvent s'étendre pratiquement sans limite à droite. La courbe de distribution du risque relatif *RR* peut ressembler alors à celle à la figure 17-4.

Figure 17-4



En présence de distributions asymétriques à droite, les statisticiens utilisent couramment la transformation logarithmique (dénotée ln pour logarithme naturel ou népérien) en vue d'obtenir une courbe, sinon parfaitement, du moins assez symétrique. Le tableau 17-2A montre clairement que la branche droite de la courbe devient moins prononcée en prenant le logarithme.

Tableau 17-2

A		**B**	
RR	ln *RR*	*RR*	ln *RR*
1	0	1	0
5	1,609	$^1/_5$	-1,609
9	2,197	$^1/_9$	-2,197
13	2,565	$^1/_{13}$	-2,565
17	2,833	$^1/_{17}$	-2,833

D'une ligne à l'autre au tableau 17-2A, le *RR* augmente plus vite que le ln *RR.* En utilisant le logarithme, les valeurs correspondant à des risques supérieurs à 1 se resserrent, la branche droite de la distribution asymétrique se rétrécit. Par contre, en utilisant le logarithme, la branche gauche de la figure 17-4 s'allonge comme en témoignent les valeurs au tableau 17-2B mais jamais autant que l'était la branche droite. Alors que le *RR* de valeur 17, par exemple, est ramené d'assez loin vers 2,833 par la transformation ln, le risque relatif *RR* de valeur $^1/_{17}$ (= 0,0588) n'est pas projeté sur une aussi longue distance par la même transformation, puisque ln $^1/_{17}$ = -2,833. Dans le premier cas, la distance est de 14,167 (17 – 2,833); dans le deuxième cas, elle est de 2,892 (0,059 + 2,833).

La distribution d'échantillonnage du ln *RR* observé sera plus symétrique que celle de *RR,* virtuellement plus proche du modèle normal. Ainsi, le calcul en approximation normale de l'intervalle de confiance pour un risque relatif

pourra mieux se faire par l'intermédiaire d'une transformation logarithmique.

La formule :

$$(\ln RR - z_{\alpha/2}\, s_{\ln RR};\ \ln RR + z_{\alpha/2}\, s_{\ln RR})$$

donne un intervalle de confiance approximatif au niveau 1 – α pour le logarithme du risque relatif. Pour retrouver le risque relatif, il suffit d'appliquer la transformation inverse du logarithme, c'est-à-dire une transformation exponentielle. Ce faisant, on obtient:

$$(e^{\ln RR - z_{\alpha/2}\, s_{\ln RR}};\ e^{\ln RR + z_{\alpha/2}\, s_{\ln RR}})$$

ou $(RR\, e^{-z_{\alpha/2}\, s_{\ln RR}};\ RR\, e^{+z_{\alpha/2}\, s_{\ln RR}})$

Les expressions qui permettent de calculer l'erreur-type du ln *RR,* soit $s_{\ln RR}$, sont différentes selon que le *RR* est un rapport de taux ou de proportions. Nous allons distinguer les deux cas.

Si *RR* est un rapport de taux : $RR = t_1/t_o$ où $t_1 = a/N_1$ et $t_o = b/N_o$, on peut prendre:

$$s_{\ln RR} = \sqrt{\frac{1}{a} + \frac{1}{b}}$$

L'intervalle de confiance approximatif au niveau 1 — α est alors:

$$IC:\ (RR\, e^{-z_{\alpha/2}\sqrt{1/a + 1/b}};\ RR\, e^{+z_{\alpha/2}\sqrt{1/a + 1/b}})$$

On a,

$$\underline{RR} = RR\, e^{-z_{\alpha/2}\sqrt{1/a + 1/b}}$$

$$\overline{RR} = RR\, e^{+z_{\alpha/2}\sqrt{1/a + 1/b}}$$

Des 150 décès observés au cours d'une année dans le groupe de 20 000 personnes, 80 sont survenus chez les hommes qui formaient un sous-groupe de 9000 personnes. Les taux de décès observés chez les hommes (h) et les femmes (f) sont respectivement:

$$t_h = t_1 = \frac{a_1}{N_1} = \frac{80}{9000} = 0{,}00889$$

et

$$t_f = t_0 = \frac{a_0}{N_0} = \frac{70}{11000} = 0{,}00636$$

Le rapport des taux t_h et t_f est un risque relatif. On a:

$$RR = \frac{0{,}00889}{0{,}00636} = 1{,}40$$

La force de mortalité est approximativement 1,4 fois plus grande chez les hommes. L'intervalle de confiance à 95 % pour le risque relatif est donné par:

$$IC:\ (1{,}40\, e^{-1{,}96\sqrt{1/80 + 1/70}};\ 1{,}40\, e^{+1{,}96\sqrt{1/80 + 1/70}})$$

$$: (1{,}02;\ 1{,}93)$$

- *Si RR* est un rapport de proportions (par exemple d'incidences cumulatives ou de prévalences relatives) : $RR = p_l/p_o$ où $p_l = a/N_1$ et $p_o = b/N_o$, on peut prendre:

$$s_{\ln RR} = \sqrt{\frac{c}{aN_1} + \frac{d}{bN_0}}$$

où $c = N_1 - a$ et $d = N_0 - b$

L'intervalle de confiance approximatif au niveau 1 — α est alors :

$$IC: \left(RR\, e^{-z_{\alpha/2}\sqrt{\frac{c}{aN_1}+\frac{d}{bN_0}}} ; \right.$$

$$\left. RR\, e^{+z_{\alpha/2}\sqrt{\frac{c}{aN_1}+\frac{d}{bN_0}}} \right)$$

On a,

$$\underline{RR} = RR\, e^{-z_{\alpha/2}\sqrt{\frac{c}{aN_1}+\frac{d}{bN_0}}}$$

$$\overline{RR} = RR\, e^{+z_{\alpha/2}\sqrt{\frac{c}{aN_1}+\frac{d}{bN_0}}}$$

Dans une étude transversale qui a porté sur 20 000 personnes (9000 hommes et 11 000 femmes), on a observé 500 cas d'une certaine maladie dont 355 étaient des femmes. On estime donc les deux proportions, ici les deux prévalences relatives, à:

$$p_f = p_1 = \frac{355}{11\,000} = 0{,}0323$$

$$p_h = p_0 = \frac{145}{9000} = 0{,}0161$$

Le rapport des deux prévalences relatives est un risque relatif. On a :

$$RR = \frac{p_f}{p_h} = \frac{0{,}0323}{0{,}0161} = 2$$

Au moment de l'enquête, il y avait deux fois plus de femmes atteintes que d'hommes, toute proportion gardée. L'intervalle de confiance approximatif à 95 % est donné par:

$$IC: \left(2e^{-1{,}96\sqrt{\frac{10645}{355\times11000}+\frac{8855}{145\times9000}}} ; \right.$$

$$\left. 2e^{+1{,}96\sqrt{\frac{10645}{355\times11000}+\frac{8855}{145\times9000}}} \right)$$

: (1,65; 2,42)

En l'absence de tout biais, il y a 95 chances sur 100 que l'intervalle qui s'étend de 1,65 à 2,42 contienne la vraie valeur du risque relatif. Cet intervalle est compatible au niveau de confiance 0,95 avec l'idée d'un risque relatif supérieur à 1.

INTERVALLE DE CONFIANCE POUR LE SMR

Le SMR a été défini comme le rapport du nombre de décès observé *(a)* chez les individus exposés à un facteur au nombre de décès attendu *(A)*. Le SMR, soit *a/A,* est bien entendu lui aussi soumis aux fluctuations d'échantillonnage. Dans ce qui suit, nous supposons que le nombre attendu *(A)*, le dénominateur du SMR, est connu sans erreur d'échantillonnage. En d'autres termes, nous supposons que la variance de *A* est nulle. Cette supposition est assez habituelle. Elle repose sur le fait que les taux spécifiques qui interviennent dans le calcul de *A* proviennent d'une population standard et peuvent être alors considérés comme des valeurs paramétriques. On peut aussi justifier cette supposition par la grande stabilité dont jouissent les taux calculés dans une population à large effectif.

Cette supposition étant faite, il s'ensuit que les fluctuations d'échantillonnage du SMR sont réduites à celles de son numérateur *(a)*. La distribution d'échantillonnage de

a n'obéit pas au modèle de la distribution normale. Toutefois, si l'on admet que l'apparition d'un décès est un événement relativement rare (c'est notre deuxième supposition), il est possible de démontrer que la racine carrée de a suit approximativement une distribution normale avec une erreur-type constante égale à 0,50. Un intervalle de confiance approximatif pour la racine carrée du numérateur du SMR est alors donné par l'expression :

$$IC: \quad (\sqrt{a} - z_{\alpha/2} \times 0{,}50; \sqrt{a} + z_{\alpha/2} \times 0{,}50)$$

On a,

$$\underline{IC} = \sqrt{a} - z_{\alpha/2} \times 0{,}50$$

$$\overline{IC} = \sqrt{a} + z_{\alpha/2} \times 0{,}50$$

Les limites de confiance inférieure et supérieure du numérateur du SMR sont obtenues en élevant au carré respectivement IC et IC. Celles du SMR sont obtenues en divisant par A les résultats de ces élévations au carré. On obtient en définitive :

$$\underline{\text{SMR}} = \frac{(\sqrt{a} - z_{\alpha/2} \times 0{,}50)^2}{A}$$

$$\overline{\text{SMR}} = \frac{(\sqrt{a} + z_{\alpha/2} \times 0{,}50)^2}{A}$$

avec comme intervalle de confiance du SMR:

$$IC: \left(\frac{(\sqrt{a} - z_{\alpha/2} \times 0{,}50)^2}{A}; \frac{(\sqrt{a} + z_{\alpha/2} \times 0{,}50)^2}{A} \right)$$

Si le nombre de décès observé (a) est égal à 16 et le nombre attendu (A) à 9, le SMR est égal à 1,8. Pour un niveau de confiance de 95 %.

$$\underline{\text{SMR}} = \frac{(\sqrt{16} - 1{,}96 \times 0{,}50)^2}{9} = 1{,}01$$

$$\overline{\text{SMR}} = \frac{(\sqrt{16} + 1{,}96 \times 0{,}50)^2}{9} = 2{,}76$$

Un intervalle de confiance à 95 % approximatif est donné par:

(1,01; 2,76)

C'est l'intervalle de confiance basé sur la transformation « racine carrée ».

INTERVALLE DE CONFIANCE POUR UN RAPPORT DES COTES

Comme pour le risque relatif, la distribution d'échantillonnage du rapport des cotes observé *RC* n'obéit pas, non plus, au modèle normal. La transformation logarithmique permet de s'y rapprocher. Ce qui a été dit pour le risque relatif peut être repris pour le rapport des cotes en utilisant la formule appropriée de l'erreur-type du ln *RC*. On rappelle que le rapport des cotes a la forme:

$$RC = \frac{ad}{bc}.$$

L'erreur-type du ln RC est approximée par :

$$s_{\ln RC} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

Un intervalle de confiance approximatif au niveau 1 – α pour le rapport des cotes est alors donné par la formule :

$$IC: (RC\, e^{-z_{\alpha/2}\sqrt{1/a+1/b+1/c+1/d}};$$

$$RC\, e^{+z_{\alpha/2}\sqrt{1/a+1/b+1/c+1/d}})$$

On a,

$$\underline{RC} = RC\, e^{-z_{\alpha/2}\sqrt{1/a+1/b+1/c+1/d}}$$

$$\overline{RC} = RC\, e^{+z_{\alpha/2}\sqrt{1/a+1/b+1/c+1/d}}$$

Les résultats d'une étude cas-témoins sur l'association entre un facteur E et une certaine maladie sont décrits au tableau 17-3.

Tableau 17-3

	M		
	+	—	
Cas	40	160	200
Témoins	20	180	200

$RC = 2{,}25$

L'intervalle de confiance approximatif à 95 % pour le rapport des cotes est donné par :

$$(2{,}25\, e^{-1{,}96\sqrt{1/40+1/160+1/20+1/180}};\ 2{,}25\, e^{+1{,}96\sqrt{1/40+1/160+1/20+1/180}})$$

$: (1{,}26; 4{,}01)$

En l'absence de biais, cet intervalle de confiance est compatible au niveau de confiance 0,95 avec l'idée d'une association réelle entre le facteur E et la maladie en question.

INTERVALLE DE CONFIANCE POUR UN COEFFICIENT DE CORRÉLATION LINÉAIRE

La distribution d'échantillonnage du coefficient de corrélation linéaire observé r obéit plutôt mal au modèle de la distribution normale, spécialement quand le paramètre ϱ c'est-à-dire le coefficient de corrélation linéaire vraie, est différent de 0. Toutefois, il existe une transformation de r qui donne naissance à une nouvelle variable invariablement symbolisée par z_r, qui elle se plie mieux aux exigences du modèle normal. Cette transformation dite de Fisher est décrite par l'équation suivante:

$$z_r = \frac{1}{2}\ln\frac{1+r}{1-r}.$$

Il a été démontré que z_r suit d'assez près une distribution normale, même pour un petit nombre n de sujets. La variance de z_r est donnée par $\frac{1}{n-3}$. En théorie, la variance est un peu différente, mais, en pratique, $\frac{1}{n-3}$ convient tout à fait, à moins d'exiger une très haute précision. On en déduit en bonne approximation la formule suivante de l'intervalle de confiance au niveau $1-\alpha$ pour le paramètre transformé z_ϱ (le paramètre dont z_r est une estimation ponctuelle).

$$IC: (z_r - z_{\alpha/2}\sqrt{\frac{1}{n-3}};\ z_r + z_{\alpha/2}\sqrt{\frac{1}{n-3}})$$

On a,

$$\underline{z_r} = z_r - z_{\alpha/2}\sqrt{\frac{1}{n-3}}$$

$$\overline{z_r} = z_r + z_{\alpha/2}\sqrt{\frac{1}{n-3}}$$

où il faut noter que le z dans z_r ne doit pas être confondu avec celui de $z\alpha/2$.

Cet intervalle de confiance n'est pas celui du coefficient de corrélation linéaire ϱ, mais bien du coefficient transformé z_ϱ. Pour obtenir l'intervalle de confiance du coefficient de corrélation linéaire, il faut appliquer la transformation inverse qui nous permet de passer cette fois-ci de z_r à r. Nous présentons en résumé les trois étapes qui conduisent au calcul en approximation normale de l'intervalle de confiance du coefficient de corrélation linéaire:

— calcul du z_r à l'aide de la transformation de Fisher;

— calcul de l'intervalle de confiance: $(\underline{z_r}, \overline{z_r})$;

— calcul de l'intervalle de confiance: $(\underline{r}, \overline{r})$

où, $\underline{r} = \dfrac{e^{2\underline{z}r} - 1}{e^{2\underline{z}r} + 1}$ et $\overline{r} = \dfrac{e^{2\overline{z}r} - 1}{e^{2\overline{z}r} + 1}$.

Ces dernières relations s'obtiennent facilement.

Avec $$z_r = \frac{1}{2} \ln \frac{1 + r}{1 - r},$$

on a, $$\frac{1 + r}{1 - r} = e^{2z_r},$$

d'où, $$r = \frac{e^{2z_r} - 1}{e^{2z_r} + 1}.$$

On trouve dans les livres de base en statistique des tables toutes faites qui permettent de passer de *r à* z_r *et,* à l'inverse, de z_r *- à r.*

À partir d'un échantillon de 25 sujets, on a estimé à 0,87 le coefficient de corrélation linéaire r entre l'âge et la tension artérielle systolique. Il s'ensuit que:

$$z_r = \frac{1}{2} \ln \frac{1 + 0{,}87}{1 - 0{,}87} = 1{,}333$$

L'intervalle de confiance à 95 % de z_ϱ est donné par:

$$IC: \left(1{,}333 - 1{,}96 \sqrt{\frac{1}{25 - 3}} \; ; \; 1{,}333 + 1{,}96 \sqrt{\frac{1}{25 - 3}}\right)$$

: (0,915; 1,751)

Ainsi, $\underline{z_r} = 0{,}915$ et $\overline{z_r} = 1{,}751$

On déduit que:

$$\underline{r} = \frac{e^{2 \times 0{,}915} - 1}{e^{2 \times 0{,}915} + 1} = 0{,}72$$

et $$\overline{r} = \frac{e^{2 \times 1{,}751} - 1}{e^{2 \times 1{,}751} + 1} = 0{,}94$$

L'intervalle de confiance approximatif à 95 % pour le coefficient de corrélation linéaire est donné par :

IC: (0,72; 0,94).

En l'absence de biais, il y a environ 95 chances sur 100 que l'intervalle qui va de 0,72 à 0,94 contienne ϱ.

INTERVALLE DE CONFIANCE POUR UN COEFFICIENT DE CORRÉLATION INTRA-CLASSE

Le coefficient de corrélation intra-classe r_I est une sorte de coefficient de corrélation. Il s'ensuit que le calcul de l'intervalle de confiance du coefficient intra-classe peut

passer, lui aussi, par la transformation de Fisher. On démontre que :

$$z_{r_I} = \frac{1}{2} \ln \frac{1 + r_I}{1 - r_I},$$

suit approximativement une distribution normale avec une variance égale à $\frac{1}{n - 3/2}$. L'intervalle de confiance prend la forme:

$$IC: \left(z_{r_I} - z_{\alpha/2} \frac{1}{\sqrt{n - 3/2}} \; ; z_{r_I} + z_{\alpha/2} \frac{1}{\sqrt{n - 3/2}}\right)$$

Avec $n = 5$ (5 sujets) et $r_I = 0{,}294$, on trouve en approximation pour le coefficient de corrélation intra-classe un intervalle de confiance à 95 % de:

IC: (-0,63; 0,87)

Compte tenu du petit nombre de sujets, les limites de cet intervalle sont assez approximatives. Il n'empêche que cet intervalle, qui comprend des valeurs négatives et positives, ne permet pas d'avancer facilement l'idée d'un coefficient de corrélation intra-classe vrai réellement supérieur à 0, donc de conclure à un accord réel même médiocre.

Les calculs ont été faits comme suit, dans l'ordre :

— $z_{r_I} = \frac{1}{2} \ln \frac{1 + 0{,}294}{1 - 0{,}294}$

$= 0{,}30294$

— $(0{,}30294 - 1{,}96 \times \frac{1}{\sqrt{5 - 3/2}}; 0{,}30294 + 1{,}96 \times \frac{1}{\sqrt{5 - 3/2}})$

$(-0{,}74472;\ 1{,}35060)$

— $\underline{r} = \frac{e^{2(-0{,}74472)} - 1}{e^{2(-0{,}74472)} + 1} = -0{,}632$

$\bar{r} = \frac{e^{2(1{,}35060)} - 1}{e^{2(1{,}35060)} + 1} = 0{,}874$

INTERVALLE DE CONFIANCE POUR UNE MESURE D'ACCORD KAPPA

Comme toute mesure calculée sur les données observées, l'estimation de x est soumise aux fluctuations d'échantillonnage. Son erreur-type est connue. Pour x quelconque, elle est donnée par la formule suivante due à Fleiss, Cohen et Everitt:

$$s_\varkappa = \frac{\sqrt{A + B - C}}{(1 - p_C)\sqrt{n}}$$

où $A = \sum_i p_{ii} [1 - (p_{i\cdot} + p_{\cdot i})(1 - \varkappa)]^2$

$B = (1 - \varkappa)^2 \sum_i \sum_j p_{ij} (p_{\cdot i} + p_{j\cdot})^2, \quad i \neq j$

et $C = [\varkappa - p_C (1 - \varkappa)]^2$

Les symboles p_{ii}, p_{ij}, $p_{i\cdot}$, $p_{\cdot j}$ etc. trouvent leur signification aux tableaux 9-4 et 9-5. Un intervalle de confiance approximatif, au niveau $1 - \alpha$, de la mesure d'accord kappa est donné par:

$$IC: (\varkappa - z_{\alpha/2} s_\varkappa;\ \varkappa + z_{\alpha/2} s_\varkappa)$$

On a, $\underline{\varkappa} = \varkappa - z_{\alpha/2} s_\varkappa$

$\bar{\varkappa} = \varkappa + z_{\alpha/2} s_\varkappa$

Considérons l'exemple aux tableaux 17-4 et 17-5 suivants.

Tableau 17-4

		Observateur O_1		
		C_1	C_2	Total
Observateur O_2	C_1	9	1	10
	C_2	0	10	10
	Total	9	11	20

Tableau 17-5

		Observateur O_1		
		C_1	C_2	Total
Observateur O_2	C_1	0,45	0,05	0,50
	C_2	0	0,50	0,50
	Total	0,45	0,55	1,00

On trouve:

$p_O = 0,95$

$pc = 0,50$

x = 0,90

$A = 0,45[1 - (0,50 + 0,45)(1 - 0,90)]^2$
$+ 0,50[1 - (0,50 + 0,55)(1 - 0,90)]^2$
$= 0,76907375$

$B = (1 - 0,90)^2 [0,05 (0,45 + 0,50)^2$
$+ 0 (0,55 + 0,50)^2]$
$= 0,00045125$

$C = [0,90 - 0,50 (1 - 0,90)]^2$
$= 0,7225$

Enfin,

$$s_\kappa = \frac{\sqrt{0,76907375 + 0,00045125 - 0,7225}}{(1 - 0,50)\sqrt{20}}$$

$$= 0,0970$$

L'intervalle de confiance à 95 % est approximativement donné par:

IC: (0,90 -1,96 x 0,0970; 0,90+ 1,96 x 0,0970)

: (0,71; 1,09)

ce qui peut être réduit à :

(0,71; 1)

puisque le kappa ne dépasse pas la valeur 1.

INTERVALLE DE CONFIANCE POUR UNE PROBABILITÉ (CUMULATIVE) DE SURVIE

Les probabilités cumulatives de survie $S(t_i)$, calculées à partir d'une table de survie, sont soumises aux habituelles fluctuations d'échantillonnage. La formule classique:

$$IC: (S(t_i) - z_{\alpha/2}\, s_{S(t_i)};\ S(t_i) + z_{\alpha/2}\, s_{S(t_i)})$$

où $^s s(t_i)$ désigne l'erreur-type de $S(t_i)$, conduit à un calcul approximatif d'intervalle de confiance pour une probabilité cumulative de survie. Le calcul de $^s s(t_i)$ utilise habituellement la formule de Greenwood, que les probabilités cumulatives de survie $S(t_i)$ aient été estimées par la méthode actuarielle ou celle de Kaplan-Meier. Cette formule approximative de l'erreur-type s'écrit:

$$s_{S(t_i)} = S(t_i)\left[\frac{D_0}{O_0(O_0 - D_0)} + \ldots + \frac{D_{i-1}}{O_{i-1}(O_{i-1} - D_{i-1})}\right]^{1/2}$$

$$\text{ou,} \quad s_{S(t_i)} = S(t_i)\left[\sum_{j=0}^{i-1} \frac{D_j}{O_j(O_j - D_j)}\right]^{1/2}.$$

Si l'on se réfère au tableau 11-4 sur la mesure de la probabilité de survie, nous avons :

$$S(240) = 0{,}6309 \text{ et,}$$

$$s_{S(240)} = 0{,}6309\left[\frac{1}{12(12-1)} + \frac{0}{10{,}5(10{,}5-0)} + \frac{1}{10(10-1)} + \frac{2}{8{,}5(8{,}5-2)}\right]^{1/2}$$

$$= 0{,}1478$$

Pour la probabilité cumulative de survie à 240 jours, on obtient l'intervalle de confiance à 95 % approximatif suivant:

IC: (0,6309 - 1,96 x 0,1478 ; 0,6309 + 1,96 x 0,1478)

: (0,34; 0,92)

La nature approximative de cet intervalle est spécialement apparente par le fait que les limites de confiance, dans notre exemple 0,34 et 0,92, sont symétriques par rapport à l'estimation ponctuelle 0,63 de la probabilité cumulative de survie à 240 jours. Or, nous savons de la théorie que la distribution d'échantillonnage d'une probabilité de survie, comme d'une proportion, n'est pas symétrique de façon générale.

RÉSUMÉ

L'estimation ponctuelle d'un paramètre varie d'un échantillon à l'autre. L'estimation de la moyenne a été utilisée pour illustrer cette variabilité et faire apparaître l'idée d'intervalle de confiance. L'estimation par intervalle d'un paramètre (moyenne, risque relatif...) consiste à trouver un intervalle susceptible de contenir, avec une certaine probabilité, la vraie valeur (inconnue) du paramètre. La probabilité est fixée par l'investigateur, le plus souvent à 0,95 ou 0,99. Elle définit le niveau de confiance de l'intervalle. En approximation normale, un intervalle de confiance prend la forme suivante:

(estim. — $Z\alpha/_2\sigma$estim.; estim. + $Z\alpha/_2\sigma$estim.)

Pour un niveau de confiance $1 - \alpha$ égal à 0,95 ou 0,99, les valeurs de $z\ \alpha/_2$ sont respectivement égales à 1,96 et 2,576. On note que sa construction utilise toujours, évidemment, l'estimation ponctuelle du paramètre, c'est-à-dire sa valeur observée. Pour un niveau de confiance et un paramètre donnés, la largeur d'un intervalle de confiance diminue lorsque la taille de l'échantillon augmente. La précision est alors meilleure. Les intervalles de confiance en approximation normale ont été donnés pour une moyenne arithmétique, médiane, proportion, taux, risque relatif, SMR, rapport des cotes, coefficient de corrélation linéaire, coefficient intra-classe, mesure d'accord kappa, probabilité cumulative de survie.

Symboles

n, N: taille de l'échantillon, de la population

μ, m: moyenne vraie, estimation ponctuelle de la moyenne

mé: médiane estimée

π, p: proportion vraie, estimation de la proportion

σ, s: écart-type vrai, estimation de l'écart-type

σ_m, s_m: erreur-type vraie de la moyenne, estimation de l'erreur-type de la moyenne

σ_p, s_p: erreur-type vraie de la proportion, estimation ponctuelle de l'erreur-type de la proportion

$t = a/N$: taux estimé

RR, *RC*, SMR: risque relatif estimé, rapport des cotes estimé, SMR estimé

$s_{\ln RR}$, $s_{\ln RC}$: estimation de l'erreur-type du $\ln RR$, estimation de l'erreur-type du $\ln RC$

ϱ, r: coefficient de corrélation linéaire vraie, estimation du coefficient

r_I: coefficient de corrélation intra-classe estimé

$\varkappa$, s_k: accord kappa estimé, estimation de l'erreur-type de kappa

$S(t_i)$, $s_{S(t_i)}$: probabilité cumulative de survie au temps t_i, estimation de son erreur-type

$\mathcal{N}(\mu,\sigma^2)$, $\mathcal{N}(0,1)$: distribution normale de moyenne μ et de variance σ^2, distribution normale centrée réduite

IC: intervalle de confiance

$\underline{IC}$, $\overline{IC}$: limite inférieure, supérieure de l'intervalle de confiance

$1 - \alpha$: niveau de confiance

$z_{\alpha/2}$: valeur lue dans la table normale centrée réduite qui correspond à un niveau de confiance $(1 - \alpha)$

Formules

1. Formules des erreurs-types

$$\sigma_m = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} \quad \text{ou} \quad \frac{\sigma}{\sqrt{n}}$$

$$s_m = \frac{s}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} \quad \text{ou} \quad \frac{s}{\sqrt{n}}$$

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} \cdot \sqrt{\frac{N-n}{N-1}} \quad \text{ou} \quad \sqrt{\frac{\pi(1-\pi)}{n}}$$

$$s_p = \sqrt{\frac{pq}{n-1}} \cdot \sqrt{\frac{N-n}{N-1}} \quad \text{ou} \quad \sqrt{\frac{pq}{n-1}}$$

$$s_{\ln RR} = \sqrt{\frac{1}{a} + \frac{1}{b}} \quad \text{(pour le rapport de taux)}$$

$$s_{\ln RR} = \sqrt{\frac{c}{aN_1} + \frac{d}{bN_0}} \quad \text{(pour le rapport de proportions)}$$

$$s_{\ln RC} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

$$z_r = \frac{1}{2} \ln \frac{1 + r}{1 - r}, \quad r = \frac{e^{2z_r} - 1}{e^{2z_r} + 1}.$$

$$z_{r_I} = \frac{1}{2} \ln \frac{1 + r_I}{1 - r_I},$$

$$s_\varkappa = \frac{\sqrt{A + B - C}}{(1 - p_C)\sqrt{n}}$$

où $A = \sum_i p_{ii} [1 - (p_{i\cdot} + p_{\cdot i})(1 - \varkappa)]^2$

$B = (1 - \varkappa)^2 \sum_i \sum_j p_{ij} (p_{\cdot i} + p_{j\cdot})^2, \quad i \neq j$

et $C = [\varkappa - p_C (1 - \varkappa)]^2$

$$s_{S(t_i)} = S(t_i) \left[\frac{D_0}{O_0 (O_0 - D_0)} + \ldots + \frac{D_{i-1}}{O_{i-1} (O_{i-1} - D_{i-1})} \right]^{1/2}$$

2. Formules des intervalles de confiance en approximation normale

moyenne :

$IC : (m - z_{\alpha/2}\sigma_m ;\ m + z_{\alpha/2}\sigma_m)$

$: (m - z_{\alpha/2} s_m ;\ m + z_{\alpha/2} s_m)$

médiane :

$IC : (mé - z_{\alpha/2}\sigma_{mé} ;\ mé + z_{\alpha/2}\sigma_{mé})$

$: (mé - z_{\alpha/2} s_{mé} ;\ mé + z_{\alpha/2} s_{mé})$

où $s_{mé} \approx 1{,}253\ s_m$

proportion :

$IC : (p - z_{\alpha/2}\sigma_p ;\ p + z_{\alpha/2}\sigma_p)$

$: (p - z_{\alpha/2} s_p ;\ p + z_{\alpha/2} s_p)$

taux :

$$IC : \left(t - z_{\alpha/2} \frac{t}{\sqrt{a}} ;\ t + z_{\alpha/2} \frac{t}{\sqrt{a}}\right)$$

risque relatif :

$$IC : (RR\, e^{-z_{\alpha/2} s_{\ln RR}} ;\ RR\, e^{+z_{\alpha/2} s_{\ln RR}})$$

rapport des cotes :

$$IC : (RC\, e^{-z_{\alpha/2} s_{\ln RC}} ; RC\, e^{+z_{\alpha/2} s_{\ln RC}})$$

SMR :

$$IC : \left(\frac{(\sqrt{a} - z_{\alpha/2} \times 0{,}50)^2}{A} ; \frac{(\sqrt{a} + z_{\alpha/2} \times 0{,}50)^2}{A}\right)$$

coefficient de corrélation linéaire :

$$IC : \left(\underline{r} = \frac{e^{2\underline{z}r} - 1}{e^{2\underline{z}r} + 1} ;\ \bar{r} = \frac{e^{2\bar{z}r} - 1}{e^{2\bar{z}r} + 1} \right)$$

où $\underline{z}_r = z_r - z_{\alpha/2} \sqrt{\frac{1}{n - 3}}$

et $\bar{z}_r = z_r + z_{\alpha/2}\sqrt{\frac{1}{n-3}}$

coefficient de corrélation intra-classe:

$$IC: \left(\underline{r}_l = \frac{e^{2\underline{z}} r_l - 1}{e^{2\underline{z}} r_l + 1};\ \bar{r}_l = \frac{e^{2\bar{z}} r_l - 1}{e^{2\bar{z}} r_l + 1} \right)$$

où $\underline{z}_{r_l} = z_{r_l} - z_{\alpha/2}\sqrt{\frac{1}{n-3/2}}$

et $\bar{z}_{r_l} = z_{r_l} + z_{\alpha/2}\sqrt{\frac{1}{n-3/2}}$

kappa:

$IC: (\kappa - z_{\alpha/2} s_\kappa;\ \kappa + z_{\alpha/2} s_\kappa)$

probabilité cumulative de survie:

$IC: (S(t_i) - z_{\alpha/2} s_{S(t_i)};\ S(t_i) + z_{\alpha/2}\, s_{S(t_i)})$

LECTURES SUGGÉRÉES

1. FLEISS, J.L. ***Statistical Methods for Rates and Proportions,*** New York, John Wiley & Sons, 1981, textes dispersés.
2. KLEINBAUM, D., KUPPER L.L. et MORGENSTERN H. ***Epidemiologic Research,*** Belmont (USA), Lifetime Learning Publications, 1982, chapitre 15, pp. 296-306.
3. ROTHMAN, *K.J.* ***Modern Epidemiology,*** Boston, Little, Brown, 1986, chapitres 11 et 12, pp. 208-220.
4. SCHERRER, B. ***Biostatistique,*** Chicoutimi, Gaétan Morin Éditeur, 1984, chapitres 8 et 10, pp.261-290, 317-366.

ANNEXE DU CHAPITRE 17

Distribution normale ou campaniforme

La distribution normale occupe une place centraie tant dans la théorie que dans la pratique de la statistique.

Modèle de distribution

Nous savons que la distribution des valeurs observées d'une variable quantitative continue peut être représentée graphiquement par un histogramme ou un polygone de fréquences. Si idéalement on pouvait augmenter indéfiniment le nombre de ces valeurs, tout en réduisant l'intervalle de classes, le polygone de fréquences deviendrait à la limite parfaitement lisse. Il prendrait la forme d'une courbe.

De manière à concrétiser un peu cette tendance vers la forme courbe, considérons l'exemple donné au chapitre 3 sur la distribution des tensions artérielles systoliques des 121 patients de 50 ans ou plus. Le polygone de fréquences de cette distribution est repris à la figure A17-1 où l'intervalle de classes choisi est de 10 mmHg.

Si on augmente le nombre de sujets en poursuivant la division des intervalles de classe indéfiniment, on atteint à la limite la courbe de fréquences illustrée à la figure A17-2. Cette dernière courbe est un modèle de distribution de fréquences, dans le sens qu'il idéalise la distribution de fréquences de tensions artérielles systoliques.

La forme de la courbe à la figure A17-2 explique qu'elle soit appelée courbe en cloche. Bien qu'elle se rencontre souvent, cette forme n'est pas exclusive. Certaines variables ont des distributions de fréquences qui épousent d'autres formes. C'est le cas notamment du poids à la naissance qui obéit plutôt à une courbe légèrement asymétrique à droite, ou de la durée de survie qui suit plutôt un modèle de distribution exponentielle comme à la figure Al 7-3.

Figure A17-1

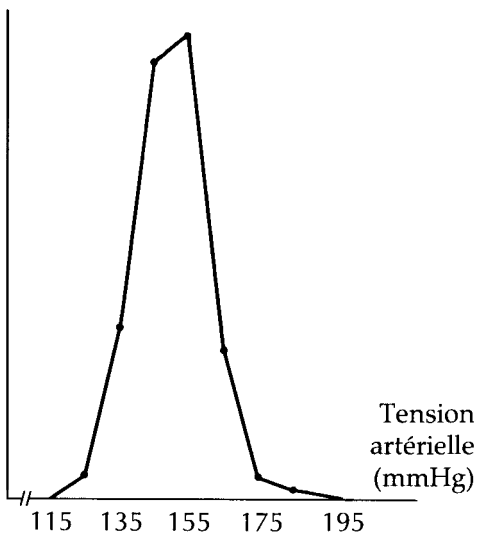
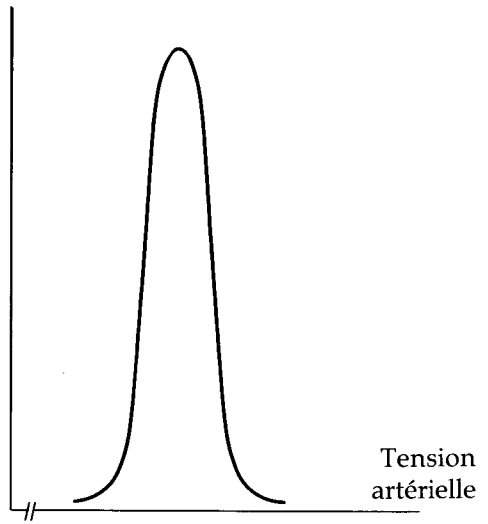


Figure A17-2

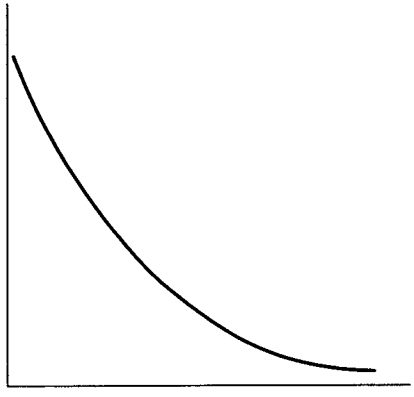


Dans cette annexe, nous porterons notre attention uniquement sur le modèle de la distribution en cloche (figure A17-2), communément appelé modèle de la distribution normale. Cette appellation n'est pourtant pas très heureuse, surtout dans les applications de la statistique aux sciences de la santé. En médecine, le terme normal a plusieurs sens. Il peut attester le bon état de santé d'un individu et s'oppose à anormal qui rappelle l'état de maladie. En statistique, le terme normal renvoie uniquement à un certain type de distribution. La tension artérielle d'un groupe de personnes peut suivre une distribution normale alors que certaines personnes du groupe ont une tension cliniquement anormale.

Pourquoi alors ce terme s'il est susceptible de créer une confusion? A une époque, on a cru, à tort, que les distributions de toutes les variables continues étaient gouvernées par la distribution en cloche. Elle apparaissait alors comme la distribution normale. Depuis, le terme est resté dans le langage.

D'autres appellations circulent, comme la distribution de Laplace-Gauss, ou simplement de Gauss. Elles ont au moins l'inconvénient de ne pas rendre justice à de Moivre qui y apporta aussi sa contribution. Nous pourrions l'appeller la distribution campaniforme pour rappeler sa forme en cloche (du latin *campana* qui signifie cloche et *forma* qui veut dire forme); ou encore, campanuliforme pour rappeler la forme en cloche qu'elle partage avec la campanule.

Figure A17-3



Modèle de la distribution normale

La distribution normale est un certain modèle de distribution de fréquences pour des variables continues. Elle est caractérisée principalement par les trois propriétés suivantes:

- La distribution normale est unimodale.
- La distribution normale est symétrique par rapport à la moyenne. Cela signifie que, si la tension artérielle systolique dans une population d'adultes est distribuée symétriquement autour d'une tension moyenne de 150 mmHg, alors on doit théoriquement retrouver la même proportion de personnes entre, par exemple, 140 et 150 mmHg qu'entre 150 et 160.
- La distribution normale vérifie les proportions clés 68, 95 et 99 %. On doit théoriquement trouver 68,3 % (plus exactement 68,27 %) des valeurs d'une variable entre la moyenne moins un écart-type et la moyenne plus un écart-type. De la moyenne moins deux écarts-types à la moyenne plus deux écarts-types doivent se situer 95,5 % (plus précisément 95,45 %) des observations. A trois écarts-types de chaque côté de la moyenne, c'est presque la totalité des observations, soit 99,7 % (plus précisément 99,73 %). Pour la tension artérielle avec une moyenne de 150 mmHg et un écart-type de 10,9 mmHg, cela voudrait dire qu'il y a environ:

68,3 % des adultes avec une tension

comprise entre 139 et 161 mmHg;

95,5 % entre 128 et 172;

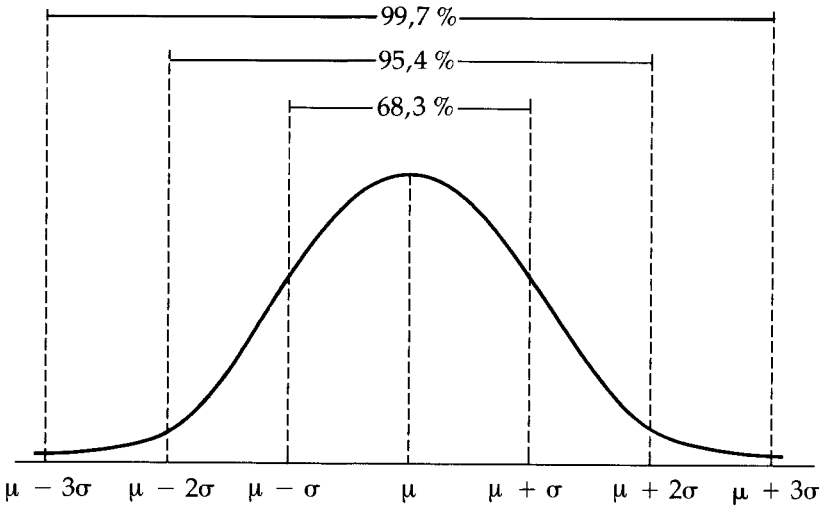
99,7% entre 117 et 183.

Une distribution normale doit comprendre ces trois propriétés. Celles-ci sont présentées à l'aide du graphique à la figure A17-4.

Note: Le symbole μ. désigne la moyenne de la variable X considérée et ri son écart-type. La différence entre *m* et μ. repose sur le fait que *m* est la moyenne calculée à partir d'un nombre limité de valeurs observées de la variable, tandis que μ. est la moyenne vraie, celle que l'on obtiendrait en prenant en considération toutes les observations possibles. Une différence de même nature existe entre s et σ. L'aire sous la courbe entre μ. — σ et μ + σ représente 68,27 % de l'aire totale. La portion entre μ. — 2σ et μ. + 2σ en représente 95,45 %, alors qu'elle atteint 99,73 % entre μ. — 3σ et μ. + 3σ. Les points de changement de courbure se produisent à un écart-type de la moyenne. Ils sont appelés les points d'inflexion. La courbe est unimodale, symétrique et s'approche de plus en plus de l'axe horizontal à mesure que l'on s'éloigne de la moyenne. A trois écarts-types, elle touche presque l'axe horizontal. Pour une distribution normale, il y a égalité entre le mode, la médiane et la moyenne arithmétique. Ajoutons au passage que, pour la majorité des modèles de distribution rencontrés en pratique, 60 à 80 % des valeurs de la variable tombent entre la moyenne moins un écart-type et la moyenne plus un écart-type. A deux écarts-types, on y trouve généralement 90 % des valeurs.

La distribution normale est entièrement déterminée par sa moyenne μ et son écart-type σ. Deux distributions normales qui ont la même moyenne et le même écart-type ont des courbes confondues. Deux distributions normales qui ont des moyennes différentes et un même écart-type sont représentées par des courbes identiques mais localisées différemment sur l'axe horizontal (figure A17-5). Enfin, deux distributions normales qui ont des écarts-types différents et une même moyenne sont représentées par des courbes qui ont la

Figure A17-4



même position sur l'axe horizontal; mais l'une est plus évasée que l'autre (figure A17-6).

Table de la distribution normale centrée réduite et calcul de probabilités

Considérons une variable X continue de moyenne μ. et d'écart-type σ. Supposons qu'elle soit adéquatement décrite par le modèle de la distribution normale, représentée à la figure A17-7.

On peut désirer connaître différentes probabilités, comme la probabilité que la variable X prenne une valeur comprise entre x_1 et x_2. Cette probabilité, dénotée Prob $(x, < X < x_2)$, correspond à l'aire sous la courbe normale entre les valeurs x_1 et x_2, c'est-à-dire à la partie ombragée à la figure A17-8.

Qu'un calcul de probabilités en rapport avec une variable X continue, en particulier normale, revienne à un calcul de surfaces ne doit pas surprendre. Il faut se rappeler que l'aire des rectangles qui composent un histogramme des fréquences (relatives) mesure la fréquence (relative). Il s'ensuit que toute portion de l'aire totale sous un polygone des fréquences relatives mesure une fréquence relative. Pour une courbe de distribution, c'est une probabilité.

Quelle est la probabilité d'observer chez un individu une tension artérielle systolique inférieure à 135 mmHg si celui-ci est choisi au hasard dans une population dont la distribution

Figure A17-5

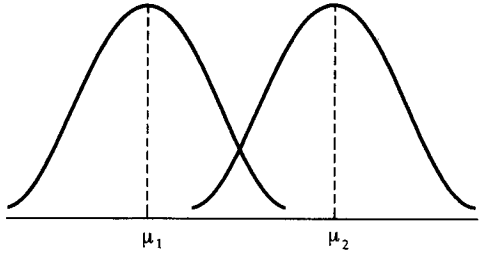


Figure A17-6

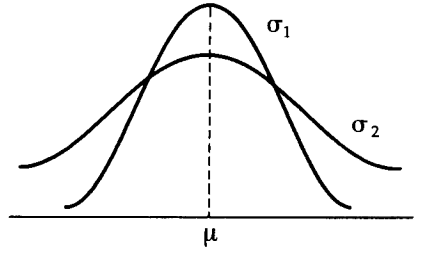


Figure A17-7

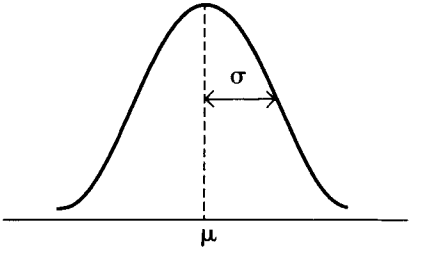
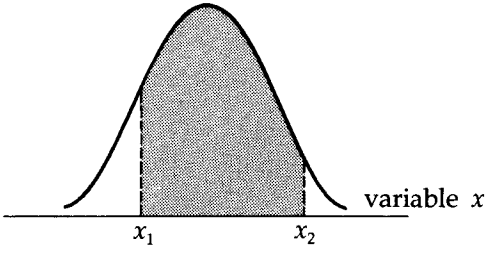


Figure A17-8



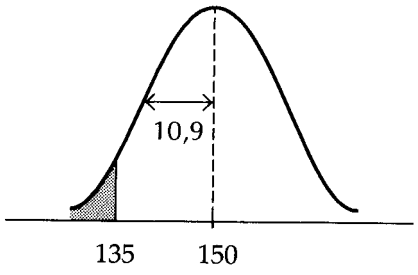
des tensions suit le modèle normal avec une moyenne de 150 et un écart-type de 10,9 mmHg? Elle correspond à l'aire ombragée à la figure A17-9 dont la mesure est faite à partir de procédés mathématiques.

Ces procédés permettent de constituer une table que l'on peut, par la suite, consulter à volonté pour calculer différentes aires, donc des probabilités. Cette table toutefois ne convient pas à une variable de moyenne différente de 150 ou d'écart-type différent de 10,9. Si la localisation ou la dispersion de la courbe à la figure A17-9 change, la partie ombragée change aussi. Comme il y a autant de courbes normales que de valeurs pour la moyenne et l'écart-type, il faudrait disposer d'un nombre illimité de tables pour faire face à toutes les situations possibles. Il est heureusement possible de pallier cette difficulté en ramenant toutes les distributions normales, quels que soient leur moyenne et leur écart-type, à une distribution normale de référence, de moyenne 0 et d'écart-type 1. C'est la *distribution normale centrée réduite.*

Pour passer d'une variable X normale à une variable normale Z centrée réduite, il suffit simplement d'utiliser la transformation linéaire:

$$Z = \frac{X - \mu}{\sigma}.$$

Figure A17-9



La variable Z a évidemment une moyenne égale à 0 et un écart-type égal à 1.

Pour l'exemple de la tension artérielle X, l'équation de transformation est:

$$Z = \frac{X - 150}{10,9}.$$

Nous étions intéressés par la probabilité d'une tension artérielle inférieure à 135, soit Prob $(X < 135)$. Cette probabilité est équivalente à Prob $(Z < -1,38)$ puisque, en vertu de la transformation précédente, $z = -1,38$ lorsque $x = 135$ $(-1,38 = \frac{135 - 150}{10,9})$. Désormais, une seule table suffira pour calculer toutes les probabilités relatives à une variable distribuée normalement. II s'agit de la table de la distribution normale centrée réduite présentée en A17-1. Elle donne l'aire de l'une ou l'autre des parties ombragées à la figure A17-11, les deux aires étant égales à cause de la symétrie de la courbe normale.

De la table A17-1, on lit que:

$$\text{Prob } (Z > 1,38) = 0,0838$$

En raison de la symétrie de la courbe normale, on a:

$$\text{Prob}(Z < -1,38) = \text{Prob}(Z > 1,38) = 0,0838.$$

Finalement, en vertu de la correspondance entre $Z < -1,38$ et $X < 135$, on a :

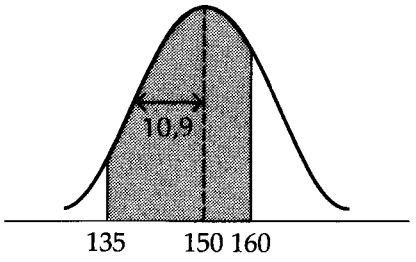
$$\text{Prob } (X < 135) = 0,0838$$
$$\approx 0,084.$$

On peut dire, qu'en moyenne, il faut s'attendre à trouver, dans un groupe de 1000 personnes,

84 avec une tension inférieure à 135 mmHg.

Pour terminer, la probabilité que la tension artérielle X prenne une valeur entre 135 et 160 mmHg, correspond à la surface ombragée à la figure A17-10.

Figure A17-10



Notons que l'aire de la partie ombragée est évidemment égale à l'aire totale moins l'aire de la partie non-ombragée ; alors :

$$\text{Prob}\,(135 < X < 160) = 1 - \text{Prob}\,(X \leq 135) - \text{Prob}\,(X \geq 160).$$

Utilisant la transformation $Z = \dfrac{X - 150}{10,9}$, on a :

$z = -1,38$ lorsque $x = 135$ et, $z = 0,92$ lorsque $x = 160$.

d'où

Prob (135 < X < 160)

= 1 — Prob $(X \leq 135)$ — Prob $(X \leq 160)$

= 1 — Prob $(Z \leq -1,38)$ — Prob $(Z \geq 0,92)$

= 1 — 0,0838 — 0,1788

= 0,7374

≈ 0,737

Sur 1000 personnes, on doit s'attendre à en trouver 737 avec une tension artérielle systolique comprise entre 135 et 150 mm Hg.

Figure A17-11

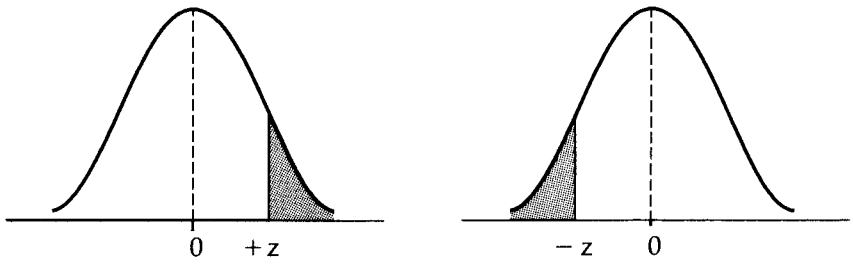


Table A17-1 (Table de la distribution normale centrée réduite)

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1771	0,1685	0,1660	0,1635	0,1611
1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0067	0,0064
2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
3,0	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010

La première colonne indique la première décimale de z, la première rangée donne la deuxième décimale. Les valeurs à quatre décimales sont les probabilités: Prob (Z> z). On peut vérifier à la figure A17-11.

Lexique anglais-français

A

Accuracy: justesse (exactitude)
Actuarial method : méthode actuarielle
Adjusted measure: mesure ajustée
Adjustment: ajustement
Age effect: effet de l'âge
Agreement: accord
Alternative hypothesis: contre-hypothèse
Ambispective study: étude ambispective
Antagonism: antagonisme
Arithmetic mean : moyenne arithmétique
Association: association
Attack rate : taux d'attaque
Attributable risk: risque attribuable

B

Bias: biais
Binary variable: variable binaire

C

Case-control study: étude cas-témoins
Categorical variable: variable qualitative
Causality: causalité
Censored data : données censurées
Central tendency: tendance centrale
Classification scale: échelle de classification
Clinical trial: essai clinique
Coefficient of variation: coefficient de variation
Coherence: cohérence
Cohort effect : effet de cohorte
Cohort study: étude de cohorte
Conditional probability: probabilité conditionnelle
Confidence interval: intervalle de confiance
Confounding: confusion
Confounding factor: facteur confondant, de confusion
Continuous variable: variable continue
Correlation : corrélation
Correlation measure: mesure de corrélation
Covariance: covariance
Cross-sectional study: étude transversale
Crude measure: mesure brute
Crude rate : taux brut
Cumulative incidence: incidence cumulative
Cutting point: seuil de positivité

D

Death : décès
Death rate : taux de décès
Decile: décile
Descriptive study: étude descriptive
Detection : détection
Dichotomous variable: variable dichotomique
Differential misclassification: erreur de classification différentielle
Discrete variable: variable discrète
Disease: maladie
Dispersion: dispersion
Dynamic population: population dynamique

E

Epidemiology: épidémiologie

Estimation: estimation
Etiologic fraction : fraction étiologique
Etiologic study: étude à visée étiologique
Event: événement
Experimental study: étude expérimentale
External validity: validité externe

F

Frequency distribution: distribution de fréquences
Frequency polygon: polygone de fréquences
Frequency table: tableau de fréquences

G

Geometric mean: moyenne géométrique

H

Healthy worker effect: effet de bonne santé
Histogram: histogramme
Hypothesis: hypothèse

I

llness: maladie
Impact: impact
Incidence: incidence
Incidence density: densité d'incidence
Incidence rate: taux d'incidence
Independent events: événements indépendants
Information: information
Interaction: interaction
Internal validity: validité interne
Interquartile range: intervalle semi-interquartile
Interval scale: échelle par intervalle
Intervention study: étude d'intervention
Intraclass correlation coefficient: coefficient de corrélation intra-classe

L

Lethality (fatality): létalité
Life expectancy : espérance de vie
Linear correlation coefficient: coefficient de corrélation linéaire
Longitudinal study: étude longitudinale
Lost to follow-up: perdu-au-suivi, perdu de vue,

M

Matching: assortiment
Mean: moyenne
Measure of agreement: mesure d'accord
Measure of association: mesure d'association
Measure of central tendency: mesure de tendance centrale
Measure of dispersion: mesure de dispersion
Measure of frequency: mesure de fréquence
Measure of impact: mesure d'impact
Measure of interaction: mesure d'interaction
Median: médiane
Misclassification: erreur de classement
Mode : mode
Modification: modification
Modifying factor: facteur modifiant
Mortality rate: taux de mortalité
Mutually exclusive events: événements mutuellement exclusifs

N

Nominal scale: échelle nominale
Nondifferential misclassification: erreur de classification non-différentielle

Normal distribution: distribution normale
Normality: normalité

Null hypothesis: hypothèse nulle

O

Observational study: étude d'observation
Occupational health: santé au travail Odds: cote

Odds ratio: rapport des (de) cotes
Ordinal scale: échelle ordinale

P

Pairing: appariement

Percentiles: centiles

Period effect: effet de période
Person-time: personne-temps

Person-time at risk: personne-temps à risque
Person-year: personne-année Population: population

Population at risk: population à risque
Precision: précision

Predictive value: valeur prédictive

Prevalence: prévalence, prévalence relative
Prevalence ratio: rapport de prévalences relatives

Prevented fraction: fraction prévenue, évitable

Preventive trial: essai préventif
Probability: probabilité

Probability distribution: distribution de probabilité

Proportion : proportion

Prospective study: étude prospective
P-value: valeur-*p* (valeur de *p*), degré de signification

Q

Qualitative variable: variable qualitative
Quantitative variable: variable quantitative

R

Random error: erreur aléatoire

Random sample: échantillon aléatoire
Randomization: randomisation

Range: étendue

Rate : taux

Ratio: rapport, ratio

Ratio scale: échelle proportionnelle
Relative risk: risque relatif

Retrospective study: étude rétrospective

Risk difference: différence des (de) risques

Risk ratio: rapport de risques

S

Sample: échantillon

Sample size: taille d'échantillon

Sampling distribution: distribution d'échantillonnage Scale: échelle

Scatter diagram: diagramme de dispersion
Selection: sélection

Sensitivity: sensibilité

Significance level: seuil (niveau) de signification

Specific measure: mesure spécifique
Specific rate: taux spécifique

Specificity: spécificité Standard deviation: écart-type

Standard error: erreur-type

Standard population: population-type

Standardized measure: mesure standardisée
Steady state: situation d'équilibre
Strata: strate
Stratification : stratification
Strength of association: force de l'association
Survival function: fonction de survie
Survival probability: probabilité de survie
Survival table: table de survie
Synergy : synergie
Systematic error: erreur systématique

T

Trend: tendance

U

Unconditional association: association non-conditionnelle

V

Validity: validité
Variable: variable
Variance: variance

W

Weight: poids
Weighted average: somme pondérée
Withdrawal: exclu-vivant

Index

T

V

AGMV
MARQUIS
Québec, Canada
1998