Méthodes quantitatives en épidémiologie

Joseph Lellouch, Pierre Ducimetière, Denis Hémon, Monique Kaminski

SEMINAIRE TECHNOLOGIQUE INSERM

"METHODES QUANTITATIVES EN EPIDEMIOLOGIE"

J. LELLOUCH P. DUCIMETIERE

D. HEMON M. KAMINSKI

1988

TABLE DES MATIERES

SEMINAIRE TECHNOLOGIQUE INSERM

"METHODES QUANTITATIVES EN EPIDEMIOLOGIE"

J. LELLOUCH P. DUCIMETIERE

D. HEMON M. KAMINSKI

CHAPITRE I.

MESURE DE LA MORBIDITE ET DE LA MORTALITE

I. Mesures "brutes"

1. Prévalence et incidence

Nous supposerons nous intéresser à une pathologie donnée que nous appelerons "maladie" dans ce chapitre et les suivants et vouloir en mesurer l'"importance" dans une population particulière, c'est à dire quantifier la morbidité correspondante.

La mesure de morbidité la plus simple est la proportion de sujets atteints par la maladie à un moment donné ou "PREVALENCE"

$$P = \frac{N'}{N+N'} \qquad (1)$$

où N' est le nombre de sujets malades et N le nombre de sujets indemnes de la maladie étudiée.

Comme on s'en rend compte immédiatement cette mesure intègre deux aspects différents :

- l'apparition de nouveaux cas de la maladie dans la population,
- la durée de la maladie.

Si le second aspect peut présenter un intérêt essentiel du point de vue de la santé publique (quel est le nombre de sujets diabétiques en France aujourd'hui ?), le premier aspect correspond mieux à l'analyse de l'étiologie d'une maladie (dans cette population, l'apparition de nouveaux cas de diabète a-t-elle un niveau anormalement élevé ?).

On est donc conduit à proposer une nouvelle mesure de la morbidité, l'INCIDENCE, qui s'intéresse aux nouveaux cas de la maladie apparus dans la population pendant un intervalle de temps donné :

$$I = \frac{M}{N.T} \qquad (2)$$

où M est le nombre de nouveaux cas de la maladie apparus pendant l'intervalle de temps $[t, t+T[$

CHAPITRE I.

MESURE DE LA MORBIDITE ET DE LA MORTALITE

I. Mesures "brutes"

1. Prévalence et incidence

Nous supposerons nous intéresser à une pathologie donnée que nous appelerons "maladie" dans ce chapitre et les suivants et vouloir en mesurer l'"importance" dans une population particulière, c'est à dire quantifier la morbidité correspondante.

La mesure de morbidité la plus simple est la proportion de sujets atteints par la maladie à un moment donné ou "PREVALENCE"

$$P = \frac{N'}{N+N'} \qquad (1)$$

où N' est le nombre de sujets malades et N le nombre de sujets indemnes de la maladie étudiée.

Comme on s'en rend compte immédiatement cette mesure intègre deux aspects différents :

- l'apparition de nouveaux cas de la maladie dans la population,
- la durée de la maladie.

Si le second aspect peut présenter un intérêt essentiel du point de vue de la santé publique (quel est le nombre de sujets diabétiques en France aujourd'hui ?), le premier aspect correspond mieux à l'analyse de l'étiologie d'une maladie (dans cette population, l'apparition de nouveaux cas de diabète a-t-elle un niveau anormalement élevé ?).

On est donc conduit à proposer une nouvelle mesure de la morbidité, l'INCIDENCE, qui s'intéresse aux nouveaux cas de la maladie apparus dans la population pendant un intervalle de temps donné :

$$I = \frac{M}{N.T} \qquad (2)$$

où M est le nombre de nouveaux cas de la maladie apparus pendant l'intervalle de temps $[t, t+T[$

N est le nombre de sujets indemnes de la maladie au moment t
T est la durée d'observation de la population. (§)

Cette nouvelle mesure de la morbidité si elle présente sur la prévalence l'avantage d'éliminer l'influence de la durée de la maladie, n'en a pas moins l'inconvénient d'être influencée par les mouvements de sortie de la population que l'on désignera par le terme général de "soustraction au risque"*.

Ainsi, dans une étude concernant la mortalité pour une cause donnée d'une population travaillant actuellement dans une branche industrielle particulière, les déménagements, prises de retraite et décès pour une autre cause sont autant de "soustractions au risque" qui peuvent modifier l'incidence de la cause de décès étudiée.

Cette nouvelle difficulté dans le choix d'une mesure de morbidité concerne en fait seulement les maladies pouvant se déclarer à un moment quelconque dans le temps et ne concerne pas celles qui se produisent à un instant particulier.

Dans ce dernier cas, qui est par exemple celui de l'incidence d'une malformation particulière à la naissance, les sujets initialement inclus dans une étude sont finalement "totalement" inclus ou exclus de l'étude, situation qui permet d'estimer l'incidence de la maladie (si les sujets exclus le sont pour des raisons indépendantes de la maladie).

Si au contraire, la maladie peut se produire à un moment quelconque, comme par exemple la mortalité cardiovasculaire dans une population donnée, les sujets ne sont pas à proprement parler exclus de l'étude du fait des soustractions au risque mais plutôt exposés au risque pendant des périodes de temps variables.

(*)Expression condensée et souvent mal comprise qui signifie : "soustraction à la possibilité d'être décomptée comme un nouveau cas de la maladie dans la population" et ne signifie pas "soustraction à l'effet d'un facteur de risque"!

(§)*Remarque: On peut montrer que dans le cas d'une population "stationnaire" c'est à dire présentant des caractéristiques stables au cours du temps du point de vue :*

- *des entrées dans la population,*
- *de l'incidence de la maladie,*
- *de la durée de la maladie,*
- *des sorties de la population*

l'incidence et la prévalence sont liées par la relation :

$$P = \frac{I.d}{1+I.d}$$

expression qui prend la forme P = I.d (donnée par de nombreux ouvrages d'Epidémiologie) lorsque l'incidence de la maladie est faible.

Pour résoudre cette nouvelle difficulté on sera conduit à introduire la notion d'incidence instantanée c'est à dire à introduire une dimension temporelle donc la notion d'incidence et à étudier la morbidité conditionnellement à l'exposition au risque à un moment donné (§ 2). Compte tenu des reparques précédantes, cette nouvelle notion nous conduira à prendre en compte la durée d'exposition au risque en introduidant la notion de personne X année (§ 3).

2. Incidence instantanée

2.1. Définition de l'incidence instantanée

Pour définir la notion d'incidence instantanée nous considérerons une "cohorte" de sujets c'est à dire un ensemble de sujets entrant dans la population à un moment donné. Le temps sera compté à partir d'une origine qui est précisément le moment d'entrée dans la population. Ce temps pourra donc être également considéré comme le "recul" par rapport à l'entrée dans la population.

S'intéressant à un petit intervalle de temps $[t, t+\Delta t[$ de durée Δt on peut définir l'incidence instantanée par analogie avec l'incidence définie plus haut dans le cas d'une population

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{M[t, t+\Delta t[}{N(t) \cdot \Delta t} \qquad (3)$$

où $M[t, t+\Delta t[$ est le nombre de nouveaux cas de la maladie apparus dans la cohorte entre les moments t et $t+\Delta t$ et N(t)est le nombre de sujets non malades en début d'intervalle.

Si l'on suppose que chaque sujet peut être caractérisé par un délai T d'apparition de la maladie et que les délais varient d'un sujet à l'autre selon une distribution statistique de densité de probabilité f(t) et de fonction cumulative F(t), la densité de probabilité du délai d'apparition de la maladie chez les sujets indemnes jusqu'à l'instant t s'écrit :

$$f(t) / [1 - F(t)]$$

et on a :

$$M[t, t+\Delta t[= N(t) \cdot \frac{f(t)}{1 - f(t)} \cdot \Delta t \qquad (4)$$

soit :

$$\lambda(t) = \frac{f(t)}{1 - F(t)}$$

l'incidence instantanée est donc précisément égale à la densité du délai d'apparition de la maladie conditionnelle au fait d'avoir échappé à celle-ci jusqu'à l'instant t.

2.2. Quelques exemples

2.2.1. Incidence instantanée constante au cours du temps

Dans ce cas on a :

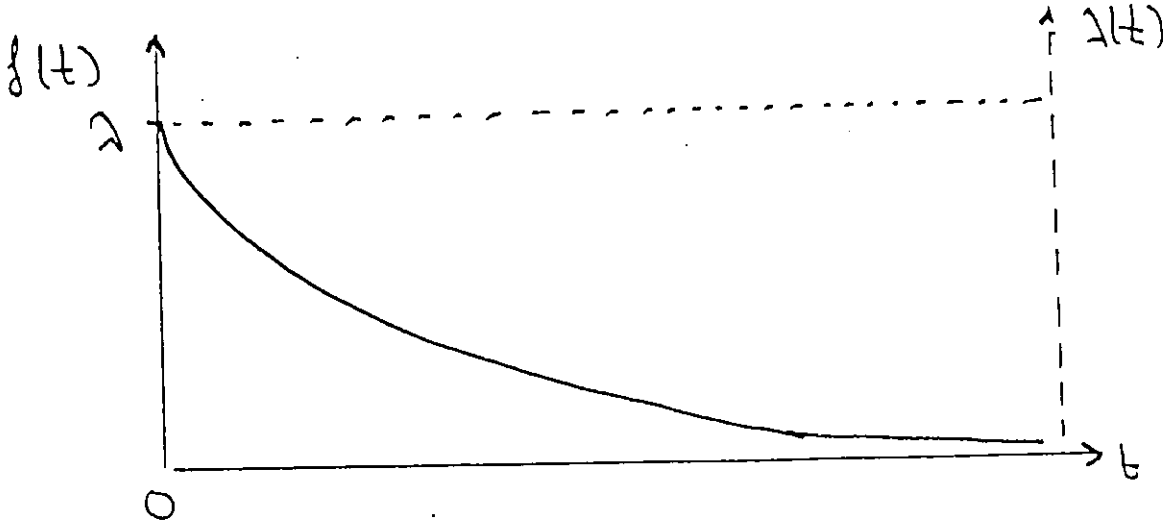
soit $$\frac{f(t)}{1-F(t)} = \lambda$$

ou $$-\frac{d\,\mathrm{Log}\{1-F(t)\}}{dt} = \lambda$$

$$\mathrm{Log}\{1-F(t)\} = -\lambda t$$

ou encore
$$\begin{cases} F(t) = 1 - \exp(-\lambda t) \\ f(t) = \lambda . \exp(-\lambda . t) \end{cases} \qquad (5)$$

Le seul cas où l'incidence instantanée est constante est celui où la distribution des délais d'apparition de la maladie a une forme exponentielle.



l'incidence instantanée est donc précisément égale à la densité du délai d'apparition de la maladie conditionnelle au fait d'avoir échappé à celle-ci jusqu'à l'instant t.

2.2. Quelques exemples

2.2.1. Incidence instantanée constante au cours du temps

Dans ce cas on a :

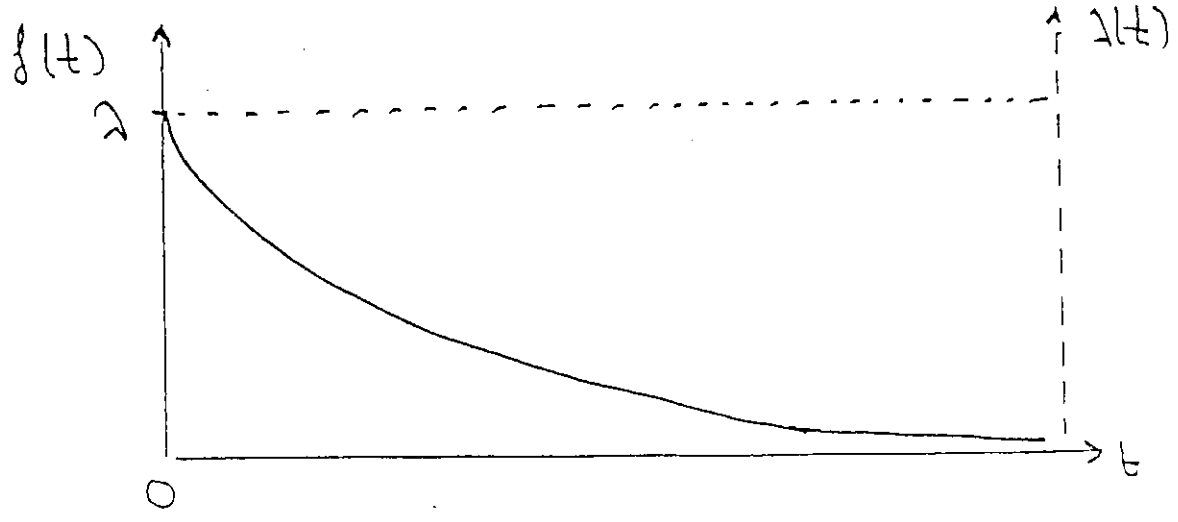
soit $$\frac{f(t)}{1-F(t)} = \lambda$$

ou $$-\frac{d\,\mathrm{Log}\{1-F(t)\}}{dt} = \lambda$$

$$\mathrm{Log}\{1-F(t)\} = -\lambda t$$

ou encore
$$\begin{cases} F(t) = 1 - \exp(-\lambda t) \\ f(t) = \lambda . \exp(-\lambda . t) \end{cases} \qquad (5)$$

Le seul cas où l'incidence instantanée est constante est celui où la distribution des délais d'apparition de la maladie a une forme exponentielle.



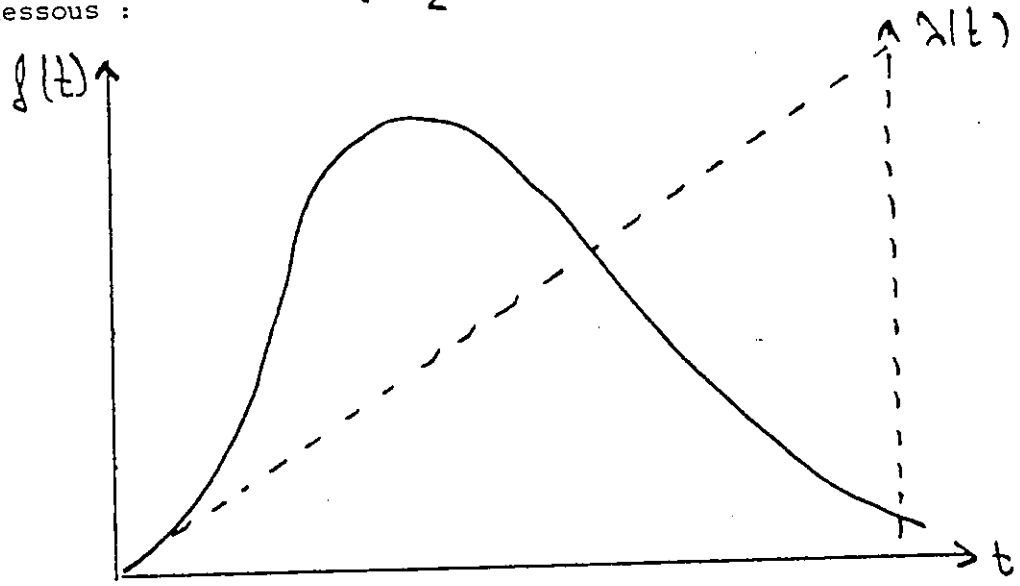
Cette distribution présente la particularité de correspondre à une population où les sujets ne "vieillissent pas" puisque la probabilité conditionnelle d'être atteint par la maladie est la même quel que soit le moment où l'on se place.

2.2.2. Incidence instantanée variable au cours du temps

Dans la pratique, l'incidence instantanée varie au cours du temps. Ainsi, dans une population où les sujets "vieillissent", l'incidence instantanée augmente au cours du temps. Par exemple :

$\lambda(t) = \alpha . t$

$f(t) = \alpha . t . \exp\left(-\alpha \frac{t^2}{2}\right)$ correspond à la densité des délais d'apparition : dont la forme est donnée dans le graphique ci-dessous :



2.2. Fonction de survie et incidence cumulée

A partir de l'incidence instantanée, il est facile de donner une expression de la "fonction de survie" qui donne la probabilité $S(t) = 1 - f(t)$ de ne pas être atteint par la maladie en fonction du délai depuis l'entrée dans la population, en effet :

$$\lambda(t) = f(t) / [1 - f(t)] = - \frac{d \operatorname{Log}[1 - f(t)]}{dt}$$

Soit :

$$\int_0^t \lambda(u)\, du = - \operatorname{Log}[1 - f(t)]$$

ou encore $$S(t) = 1 - f(t) = \exp\left\{ - \int_0^t \lambda(u)\,du \right\} \quad (6)$$

où la quantité $$\Lambda(t) = \int_0^t \lambda(u)\,du \quad (7)$$

somme des incidences instantanées de 0 à t est appelée "incidence cumulée".

L'expression (6) permettra comme on va le voir au paragraphe suivant d'estimer la fonction de survie quand il y a soustraction au risque, pourvu que l'on sache estimer l'incidence instantanée malgré ces soustractions. Celle-ci ne faisant intervenir que la probabilité conditionnelle d'apparition de la maladie, cette estimation sera possible.

2.4. Retour sur l'incidence définie au niveau d'une population

Nous avons initialement défini l'incidence dans une population où les sujets entraient et sortaient et présentaient donc des reculs variables. L'incidence pour l'intervalle de temps $[t, t+\Delta t[$ pouvait s'écrire en reprenant (2) :

$$I = \frac{M[t, t+\Delta t[}{N(t) \cdot \Delta t}$$

ou encore

$$I = \sum_u \frac{N(t/t-u)}{N(t)} \cdot \frac{M[t, t+\Delta t / t-u[}{N(t/t-u)}$$

où u est le délai d'entrée dans la population

et $N(t/t-u)$ et $M[t, t+\Delta t / t-u[$

sont des quantités analogues à $N(t)$ et $M[t, t+\Delta t[$ correspondant aux sujets entrés dans la population à l'instant t-u c'est à dire ayant un recul u à l'instant t.

L'incidence I s'écrit donc comme la moyenne des incidences instantanées pondéréepar la distribution des reculs

$$I = \sum_u w(u) \cdot \lambda(u) \quad (8)$$

avec $w(u) = \dfrac{N(t/t-u)}{N(t)}$

$$\lambda(u) = \frac{M[t, t+\Delta t / t-u[}{N(t/t-u).\Delta t}$$

Dans le cas où l'incidence instantanée est constante (cf. 2.2.1. ci-dessus) l'incidence I est égale à cette constante :

$$I = \sum_u w(u).\lambda = \left\{ \sum_u w(u) \right\}.\lambda = \lambda$$

Sinon la valeur de I dépend de la distribution des reculs et par conséquent, comme nous l'avons signalé plus haut (§ 1) de l'existence de soustraction au risque correspondant à la sortie des sujets de la population.

3. Estimation de l'incidence instantanée et de la fonction de survie

3.1. Incidence instantanée constante au cours du temps

3.1.1. Informations individuelles disponibles

On suppose ici qu'une étude longitudinale a permis de suivre un échantillon de n sujets pendant des intervalles de temps de durée t_1, t_2 ... t_n dont le début correspond à l'entrée dans la population et dont la fin correspond :

- à l'apparition de la maladie après des reculs t_1, t_2 ... t_n pour les M premiers sujets ;
- à la soustraction au risque après des reculs t_{M+1} ..... t_n pour les $n-M$ sujets suivants.

Si l'on suppose l'incidence instantanée constante, (cf. § 2.2.1. ci-dessus) :

-la probabilité d'apparition de la maladie au bout d'un temps t s'écrit :

$$f(t).\Delta t = \lambda.\exp(-\lambda.t).\Delta t$$

- la probabilité d'être non malade si une soustraction au risque intervient après un temps t s'écrit :

$$1 - f(t) = \exp(-\lambda.t)$$

la vraisemblance des observations s'écrit donc :

$$V = \prod_{i=1}^{M} \left\{ \lambda.\exp(-\lambda.t_i).\Delta t \right\} \times \prod_{i=M+1}^{n} \exp(-\lambda.t_i)$$

(où le symbole $\prod$ joue pour les produits le rôle du symbole $\sum$ pour les sommes).

soit

$$V = \lambda^M . \exp\left(-\lambda . \sum_{i=1}^{M} t_i\right) . \Delta t^M . \exp\left(-\lambda . \sum_{i=M+1}^{n} t_i\right)$$

ou encore

$$V = \lambda^M . \exp\left(-\lambda . PA\right) . \Delta t^M$$

où $PA = \sum_{i=1}^{M} t_i + \sum_{i=M+1}^{n} t_i$ est la somme des durées d'exposition au risque de tous les sujets de l'échantillon encore appelée "personnes années d'exposition".

On obtient alors facilement l'estimation du maximum de vraisemblance de l'incidence instantanée, en annulant la dérivée de log V.

$$\frac{d\,Log V}{d\lambda} = \frac{M}{\lambda} - PA = 0 \Leftrightarrow \lambda = \boxed{\hat{\lambda} = \frac{M}{PA}} \quad (9)$$

L'estimation du maximum de vraisemblance . de λ est ainsi simplement le rapport du nombre de cas de la maladie observés au nombre total de personnes X années d'exposition. On peut montrer par ailleurs que $var(\hat{\lambda}) = \lambda / PA$ la précision de l'estimation (9) étant donc fonction à la fois de l'incidence instantanée elle-même et du nombre de personnes X années observées.

A partie de (9) on peut également estimer la fonction de survie au temps t en utilisant les expressions (5) et (6) :

$$\hat{S}(t) = 1 - \hat{f}(t) = \exp\left(-\hat{\lambda} . t\right)$$

$$Log\,\hat{S}(t) = -\hat{\lambda} . t$$

3.1.2. Informations disponibles aux seules extrêmités de l'intervalle d'observations

On peut obtenir une expression approchée de (9) lorsque l'on ne connait pas les valeurs individuelles des durées d'expositions $t_1 \ldots, t_n$ mais seulement :

- le nombre n de sujets présents non malades au début de l'intervalle, $[0,T]$
- le nombre n' de sujets présents non malades à la fin de l'intervalle $[0,T]$
- le nombre M de cas de la maladie apparus au cours de l'intervalle $[0,T]$

En effet, si l'on suppose dans ce cas que l'apparition de la maladie et la soustraction au risque interviennent en moyenne au milieu de l'intervalle $[0,T]$ on peut calculer PA de façon approchée :

$$PA \simeq n' . T + (n - n') \frac{T}{2} = \left(\frac{n+n'}{2}\right) . T$$

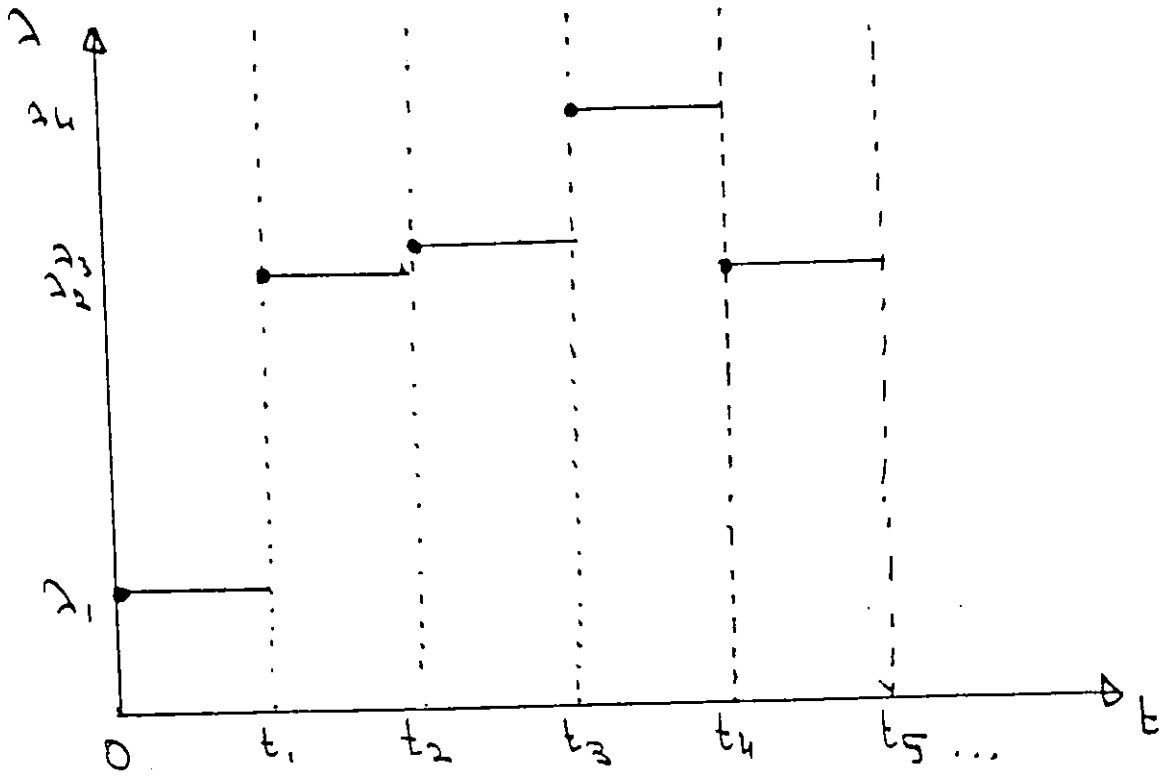
Le premier terme correspond aux n sujets exposés pendant tout l'intervalle et le second aux (n-n') sujets exposés en moyenne pendant une durée $\frac{T}{2}$ du fait de leur soustraction au risque. On a alors :

$$\hat{\lambda} = \frac{M}{\left(\frac{n+n'}{2}\right) . T} \qquad (10)$$

où $\frac{(n+n')}{2}$ est appelée "population exposée à la moitié de l'intervalle".

3.2. Incidence instantanée constante sur une série d'intervalles : méthode actuarielle

On supposera cette fois, comme l'indique le graphique ci-après, que l'incidence instantanée n'est pas constante sur tout l'intervalle observé mais seulement dans chacun des sous intervalles qui le composent.



On supposera ainsi :

que dans l'intervalle $[t_{i-i}, t_i[$ de durée $T_i = t_i - t_{i-i}$ l'incidence instantanée a la valeur λ_i

La vraisemblance des observations s'écrit en fait dans ce cas comme le produit des vraisemblances des observations faites (éventuellement sur les mêmes sujets) dans chacun des intervalles. Ainsi si M_i cas de maladies sont apparus dans l'intervalle $[t_{i-i}, t_i[$ en observant un total de PA_i personnes années d'exposition dans cet intervalle on a :

$$\hat{\lambda}_i = \frac{M_i}{PA_i} \qquad \text{var}(\hat{\lambda}_i) = \frac{\lambda_i}{PA_i}$$

Par ailleurs, pour estimer la fonction de survie, on peut écrire :

$$S(t_{k_2}) = \exp\left\{-\int_0^{t_{k_2}} \lambda(u)\,du\right\} = \exp\left\{-\sum_{i=1}^{k_2} \lambda_i \cdot T_i\right\}$$

pour obtenir l'estimation

$$\hat{S}(t_{k_2}) = \exp\left\{-\sum_{i=1}^{k_2} \hat{\lambda}_i \cdot T_i\right\} \qquad (11)$$

Si les valeurs $\hat{\lambda}_i . T_i$ sont "faibles"(*) cette expression peut s'écrire de façon approchée :

$$\hat{S}(t_k) = \prod_{i=1}^{k} (1 - \hat{\lambda}_i . T_i) \qquad (12)$$

estimateur dit de la méthode actuarielle.

Par ailleurs, les estimations fournies par chaque intervalle étant indépendantes on peut écrire :

$$\text{var Log } \hat{S}(t_k) = \text{var} \left\{ - \sum_{i=1}^{k} \hat{\lambda}_i . T_i \right\}$$

$$= \sum_{i=1}^{k} \frac{\lambda_i}{PA_i} . T_i^2$$

expression qui permet d'obtenir la variance de (12) connue sous le nom de formule de Greenwood :

$$\text{var } \hat{S}(t_k) = \{ S(t_k) \}^2 . \left\{ \sum_{i=1}^{k} \frac{\lambda_i}{PA_i} . T_i^2 \right\} \qquad (13)$$

3.3. Incidence instantanée quelconque

Si l'on découpe l'intervalle de temps où l'on a fait des observations en intervalles de plus en plus petits et que l'on utilise l'estimation selon la méthode actuarielle, on finit par atteindre la situation limite où les évènements "apparition d'un cas de maladie" et "soustraction au risque" ne se produisent qu'aux extrêmités des intervalles.

Si aucun cas de maladie n'apparait au cours d'un intervalle, l'estimation de l'incidence est nulle dans cet intervalle et la contribution à l'estimation (12) de la fonction de survie est égale à 1.

Si m cas de maladie apparaissent au temps t alors que n sujets non malades étaient présents immédiatement avant, la contribution correspondante à la fonction de survie est donnée par (12) :

$$(1 - \hat{\lambda}_i . T_i) = \left(1 - \frac{m}{n . T_i} . T_i \right) = \frac{n - m}{n}$$

(*) quand $x < 0{,}001$ on a $1 < \frac{\exp(-x)}{1-x} < 1{,}001$

On obtient l'estimateur de Kaplan-Meier de la fonction de survie :

$$\hat{S}(t) = \prod_{t_j < t} \frac{n'_j - m'_j}{n'_j}$$

La fonction de survie est constante entre les instants successifs d'apparition de cas de maladie et fait un "saut multiplicatif" de $(n'_j - m'_j)/n'_j$

à chaque instant t_j ou m_j cas apparaissent parmi n_j sujets non malades.

CHAPITRE II

MORBIDITE - MORTALITE

VARIATIONS AU COURS DU TEMPS

1 - ETUDE D'UN EXEMPLE

L'étude des variations de mortalité par cause spécifique en fonction du temps dans une zone géographique ... fait l'objet d'un grand nombre de travaux. Bien qu'en apparence purement descriptifs ils visent très généralement à fournir des interprétations susceptibles d'éclairer la connaissance épidémiologique. Bien entendu il peut s'agir aussi bien de données d'incidence dans la mesure où des registres de morbidité existent sur des périodes suffisamment longues. Nous parlerons essentiellement de mortalité.

Les données sont généralement présentées sous la forme de deux tableaux m_{ij} et n_{ij} donnant respectivement le nombre de morts observées dans la $j^{ème}$ époque et appartenant à la $i^{ème}$ tranche d'âge et le nombre total de personnes x années à risque correspondant.

On s'intéresse aux variations du taux $t_{ij} = m_{ij}/n_{ij}$. Bien entendu le but étant finalement de tenter une interprétation de ces variations, il convient d'éliminer au maximum celles dûes à des modifications des conditions de diagnostic et/ou de notification.

Prenons l'exemple de la mortalité par cancer du poumon en France de 1949 à 1978 dans le sexe masculin. Le tableau suivant donne le taux de mortalité/100.000 par an calculé pour les classes d'âge de 5 ans et des périodes de 5 ans (obtenu en divisant le nombre total de morts observé par l'estimation du nombre d'années x sujets à risque durant la période).

	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84
1949-1953	1.3	2.9	6.0	17.3	32.2	51.0	60.1	69.4	68.7	59.0	45.2
1954-1958	1.2	2.8	8.4	19.8	40.6	70.0	93.1	102.6	105.7	97.5	71.3
1959-1963	1.3	3.4	8.3	22.7	46.9	85.8	123.3	146.3	146.4	130.5	108.0
1964-1968	1.5	4.2	11.2	24.1	54.1	92.8	148.5	193.7	200.2	187.5	153.9
1969-1973	1.4	4.6	14.5	30.0	56.9	101.3	158.8	219.5	256.5	247.4	207.8
1974-1978	1.6	5.7	16.6	40.4	74.0	116.2	184.2	250.5	301.5	322.8	295.5

La représentation des variations conjointes de la mortalité en fonction de l'âge et de l'époque du décès peut être effectuée sous forme graphique : âge à époque constante ou époque à âge constant. Pour des raisons sur lesquelles nous reviendrons les taux de mortalité sont représentés par une échelle logarithmique.

Le premier graphique met en évidence deux faits : (Fig. 1)

- pour une époque de décès donné la mortalité présente un maximum à un âge d'autant plus élevé que l'époque est récente (de 67 à 77 ans)
- la mortalité à tout âge est d'autant plus élevée que l'époque est récente.

Le second graphique fournit la même information sous une forme différente mais met clairement en évidence que le taux de mortalité ne peut pas résulter simplement de l'addition d'un facteur "époque" à "l'effet propre" de l'âge (Fig. 2 et 2 bis).

Or la recherche d'une interprétation des variations de mortalité passe par l'emploi de modèles aussi simples que possible dont en particulier :

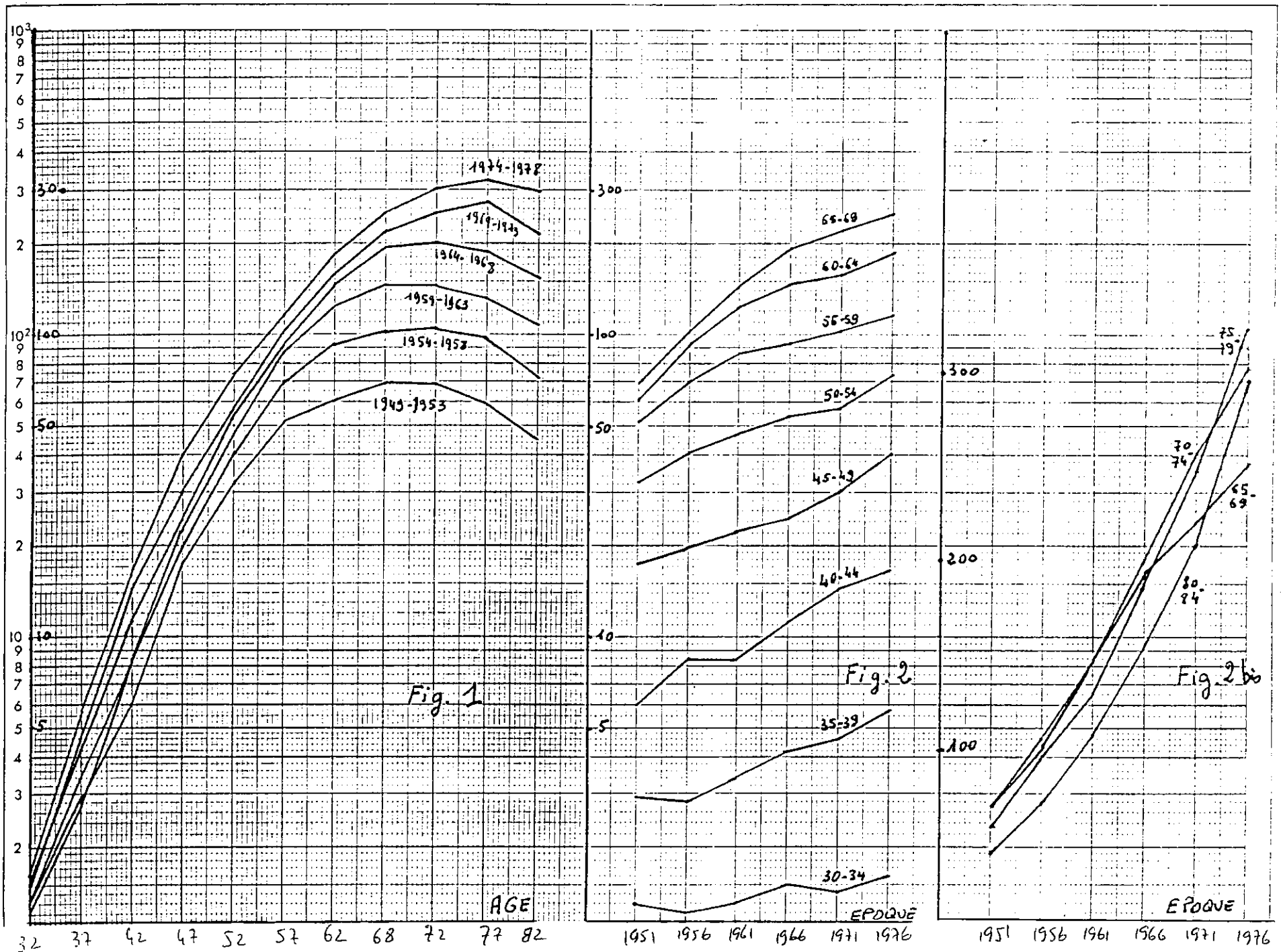
$$\log t_{ij} = a_i + b_j$$

où a_i représente la mortalité "biologique" ou "intrinsèque" à l'âge i et b_j le terme "correctif" indépendant de l'âge et propre à l'époque j.

Un tel modèle sera appelé dans la suite modèle multiplicatif puisqu'il peut s'écrire $t_{ij} = A_i B_j$, le terme B_j étant un terme multiplicatif de correction.

Sous forme complète il s'écrit $\log t_{ij} = a_i + b_j + \varepsilon_{ij}$ (modèle 1) où ε_{ij} représente une fluctuation aléatoire que l'on suppose normale de moyenne nulle et dont la variance σ^2_{ij} peut être approximée en considérant que le nombre observé de morts m_{ij} a une fluctuation de Poisson

$$\text{var } t_{ij} = \frac{\text{var } m_{ij}}{n_{ij}^2} \# \frac{m_{ij}}{n_{ij}^2} = \frac{t_{ij}}{n_{ij}}$$


1974-1978
1969-1973
1964-1968
1959-1963
1954-1958
1949-1953
AGE
Fig. 1
65-69
60-64
55-59
50-54
45-49
40-44
35-39
30-34
EPOQUE
Fig. 2
75-79
70-74
65-69
80-84
EPOQUE
Fig. 2 bis

$$\sigma^2_{ij} = \text{var}(\log t_{ij}) \simeq \frac{1}{t_{ij}}^2 \text{var } t_{ij} = \frac{1}{n_{ij} t_{ij}} = \frac{1}{m_{ij}}$$

L'estimation des paramètres a_i et b_j du modèle peut alors être obtenue par une méthode des moindres carrés pondérés, c'est-à-dire en minimisant :

$$S = \sum_{i,j} \omega_{ij} (\log t_{ij} - a_i - b_j)^2$$

où les poids ω_{ij} sont choisis égaux à l'inverse des variances soit :

$$S = \sum_{i,j} m_{ij} (\log t_{ij} - a_i - b_j)^2$$

La minimisation de S ne fournit cependant pas un ensemble unique de solutions pour les paramètres a_i et b_j, on montre qu'il est nécessaire pour cela de fixer une "contrainte" linéaire sur les paramètres que l'on fera porter sur les seuls termes b_j. Classiquement on choisit la contrainte :

$\sum_j b_j = 0$ mettant ainsi en évidence que le terme b_j s'écrit comme un terme correctif à la mortalité intrinsèque.

En pratique cette procédure n'est qu'un cas particulier d'une méthode plus générale (modèle linéaire généralisé) qui permet de fournir en plus de l'estimation des paramètres un test d'adéquation du modèle sous la forme d'un Chi-2. Le programme d'ordinateur GLIM par exemple est particulièrement adapté pour ces calculs.

Revenons à notre exemple, les résultats du modèle 1 sont représentés à la figure 3.

Le test du modèle conduit à un χ^2 de 639,7 pour 66-16 = 50 ddl qui est très hautement significatif. Comme on s'y attendait au vu des graphiques le modèle est rejeté.

Parmi les modèles alternatifs possibles qui doivent faire intervenir une "interaction" entre les effets "âge" et "époque" l'un d'entre eux est particulièrement important : il fait

intervenir le concept de "génération", terme devant être préféré à celui de "cohorte".

En effet le facteur temps que l'on cherche à interpréter, pourrait très bien ne pas être lié à la date du décès (époque) mais plutôt à la date de naissance des sujets (génération).

Par exemple un facteur étiologique du cancer du poumon pourrait être introduit dans l'environnement des sujets d'une même génération, c'est donc dans un graphique âge x génération et non pas âge x période que cet effet serait le mieux mis en évidence. Ce graphique peut être obtenu à partir d'un tableau analogue au tableau précédent mais pour lequel les lignes représenteraient des classes de 5 années de naissance au lieu de 5 années de décès. Il est facile de se rendre compte qu'en toute rigueur ce tableau nouveau ne peut être obtenu à partir du précédent mais qu'il est possible d'en avoir une estimation. En effet recherchons les dates de naissance des sujets d'une même case du tableau âge x époque

époque \ âge	a, a + 4	a + 5, a + 9
d, d + 4	d-a-5, d-a+4	d-a-10, d-a-1
d + 5, d + 9	d-a, d-a+9	d-a-5, d-a+4

On s'aperçoit d'une part que les dates de naissance des sujets d'une même case sont comprises dans un intervalle de 10 ans et non de 5 ans, que, si on appelle génération cet intervalle, les sujets apparaissant sur une même diagonale du tableau appartiennent à la même génération mais que les sujets d'une même génération peuvent appartenir à deux cases voisines du tableau. La reconstitution des dates de naissance à partir du tableau de départ conduit donc à des générations de 10 ans emboîtées les unes dans les autres, les "centres" des générations étant cependant distincts et séparés de 5 ans. En repérant chaque génération par sa date centrale, toute ambiguité est levée.

A partir d'un tableau I x J constitué de I classes d'âge au décès et de J époques de décès on déduit de cette manière un tableau I x (I + J - 1) comptant I + J - 1 générations dans lequel I (I - 1) cases sont vides (aucune observation de décès). Dans l'exemple, 11 + 6 - 1 = 16 générations dont les dates centrales vont de 1869 à 1944 sont constituées.

La représentation graphique des variations de mortalité peut là aussi être effectuée de deux manières : âge à génération constante (figure 4) ou génération à âge constant (figure 5). part, pour une même génération la mortalité est systématiquement croissante en fonction de l'âge et d'autre part sur un même graphique les courbes obtenues sont remarqueblement parallèles suggérant que sur le nouveau tableau le modèle

$$\log t_{ik} = a_i + c_k + \varepsilon_{ik} \quad \text{(modèle 2)}$$ pourrait bien s'ajuster aux données où i de 1 à I = 11 repère les classes d'âge, chacune de mortalité intrinsèque a_i et k de 1 à K = 16 repère les générations chacune introduisant le terme correctif c_k indépendant de l'âge.

De la manière dont les générations ont été construites, s'aperçoit que le modèle 2 peut s'écrire en termes de classes d'âge et d'époque de décès, en effet le taux t_{ik} est celui observé sur le tableau initial dans la case i, j = k + i - I soit ici j = k + i - 11.

Le modèle 2 peut donc être réécrit sous la forme :

$$\log t_{ij} = a_i + c_{j-i+11} + \varepsilon_{ij} \quad \text{(modèle 2')}$$ qui peut alors être comparé au modèle 1 : le terme correctif dû à l'époque j est supprimé et remplacé par un terme d'interaction représentant le terme correctif dû à la génération.

Le modèle 2 peut être ajusté aux données de la même façon que le modèle 1, les termes c_{j-i+11} étant soumis à une même contrainte linéaire indiquant par exemple que la moyenne des termes correctifs des générations est nulle :

$\sum_{k=1}^{16} c_k = 0$. Les résultats sont représentés à la figure 6.

L'effet intrinsèque de l'âge ne présente aucun maximum. Le test d'ajustement fournit un χ^2 de 46,9 pour 66 - 11 - 16 + 1 = 40 ddl, qui n'est pas significatif.

A partir d'un tableau I x J constitué de I classes d'âge au décès et de J époques de décès on déduit de cette manière un tableau I x (I + J - 1) comptant I + J - 1 générations dans lequel I (I - 1) cases sont vides (aucune observation de décès). Dans l'exemple, 11 + 6 - 1 = 16 générations dont les dates centrales vont de 1869 à 1944 sont constituées.

La représentation graphique des variations de mortalité peut là aussi être effectuée de deux manières : âge à génération constante (figure 4) ou génération à âge constant (figure 5). part, pour une même génération la mortalité est systématiquement croissante en fonction de l'âge et d'autre part sur un même graphique les courbes obtenues sont remarqueblement parallèles suggérant que sur le nouveau tableau le modèle

$\log t_{ik} = a_i + c_k + \varepsilon_{ik}$ (modèle 2) pourrait bien s'ajuster aux données où i de 1 à I = 11 repère les classes d'âge, chacune de mortalité intrinsèque a_i et k de 1 à K = 16 repère les générations chacune introduisant le terme correctif c_k indépendant de l'âge.

De la manière dont les générations ont été construites, s'aperçoit que le modèle 2 peut s'écrire en termes de classes d'âge et d'époque de décès, en effet le taux t_{ik} est celui observé sur le tableau initial dans la case i, j = k + i - I soit ici j = k + i - 11.

Le modèle 2 peut donc être réécrit sous la forme :

$\log t_{ij} = a_i + c_{j-i+11} + \varepsilon_{ij}$ (modèle 2') qui peut alors être comparé au modèle 1 : le terme correctif dû à l'époque j est supprimé et remplacé par un terme d'interaction représentant le terme correctif dû à la génération.

Le modèle 2 peut être ajusté aux données de la même façon que le modèle 1, les termes c_{j-i+11} étant soumis à une même contrainte linéaire indiquant par exemple que la moyenne des termes correctifs des générations est nulle :

$\sum_{k=1}^{16} c_k = 0$. Les résultats sont représentés à la figure 6.

L'effet intrinsèque de l'âge ne présente aucun maximum. Le test d'ajustement fournit un χ^2 de 46,9 pour 66 - 11 - 16 + 1 = 40 ddl, qui n'est pas significatif.

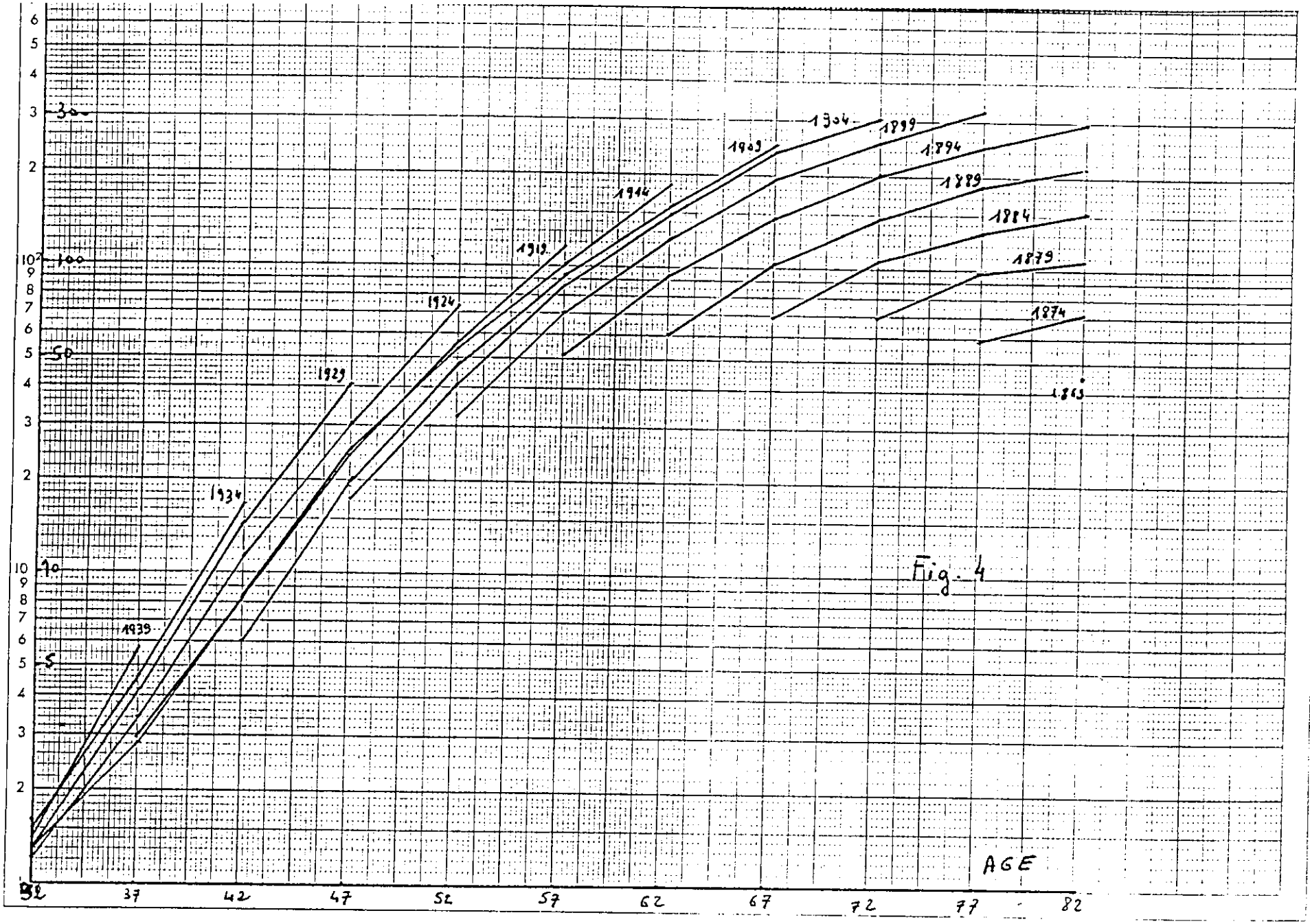

1944
1909
1904
1899
1894
1889
1884
1879
1874
1863
1919
1924
1929
1934
1939
AGE
32
37
42
47
52
57
62
67
72
77
82
30
100
50
10
5

Fig. 4

N° 1023

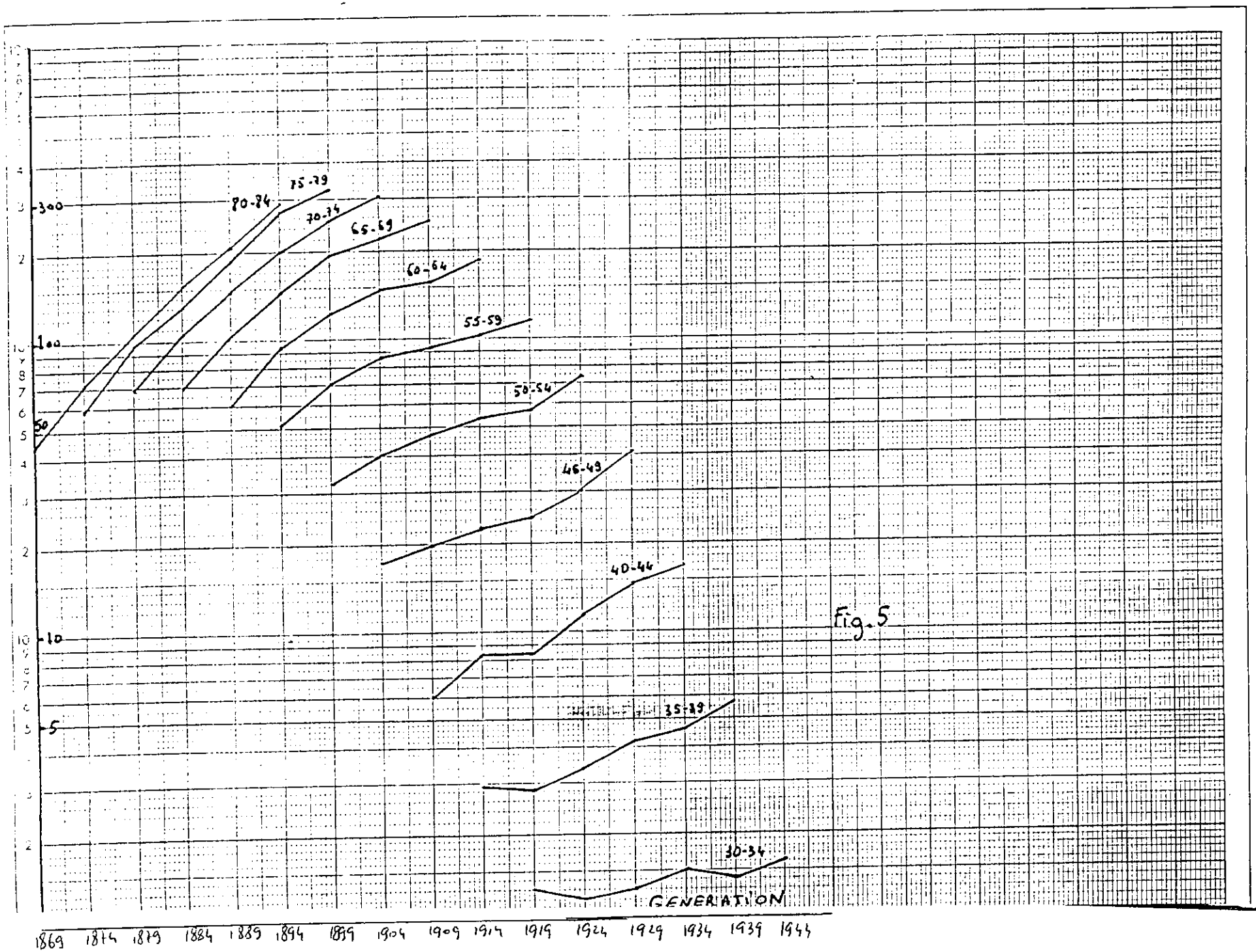

300
100
50
10
5
80-84
75-79
70-74
65-69
60-64
55-59
50-54
46-49
40-44
35-39
30-34
GENERATION
1869
1874
1879
1884
1889
1894
1899
1904
1909
1914
1919
1924
1929
1934
1939
1944

Fig. 5

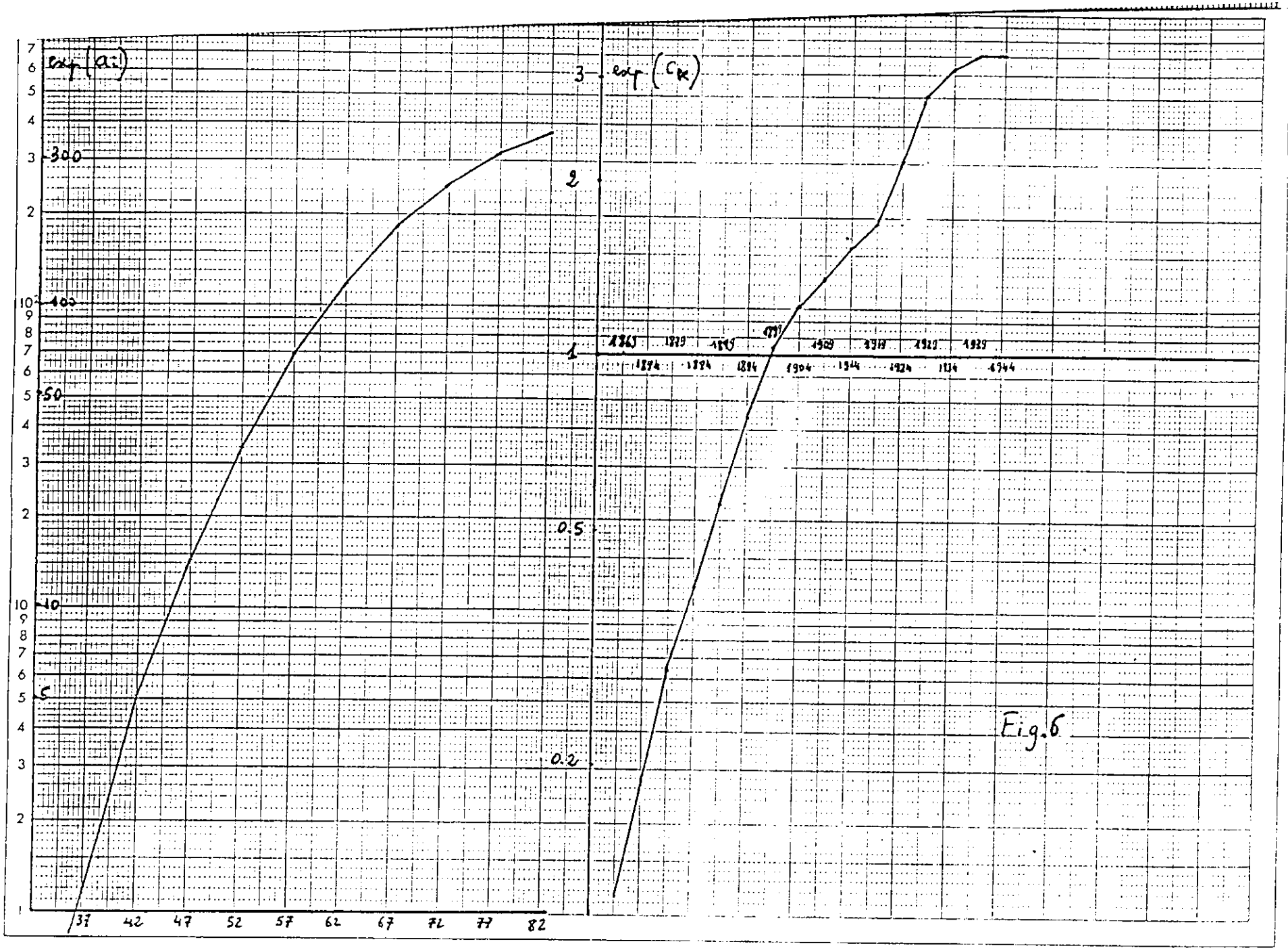


Fig. 6

N° 1023

2 - ETUDE GLOBALE

En pratique le problème se pose, à partir d'un tableau de mortalité:âge x époque, de détecter la présence, éventuellement simultanée, d'effets "période" et d'effets "génération". Avant d'étudier une généralisation des modèles 1 et 2' précédents, il est possible de raisonner graphiquement.

En définitive le tableau initial, nous l'avons vu, est susceptible d'être représenté sous la forme de quatre graphiques qui contiennent la même information.

Il est aisé de voir que les deux graphiques (âge à époque constante) et (âge à génération constante) ne constituent en réalité q'une seul et même graphique (figure 7). En reliant différemment les points du graphique, on passe immédiatement de l'un à l'autre, permettant ainsi de voir si l'effet âge se présente différemment selon que l'on raisonne à époque ou à génération constante : c'est ce qui se passait dans l'exemple précédent puisque cet effet présentait ou non un maximum selon le cas.

L'analyse des graphiques représentant l'effet époque ou l'effet génération à âge constant est par contre moins directe, on raisonnera sur un exemple théorique. La figure 8 suppose le problème résolu et représente les variations des 3 composantes de la mortalité dans le modèle complet

$$\log t_{ij} = a_i + b_j + c_{j-i+1} + \varepsilon_{ij} \quad \text{(modèle 3)}$$

qui comprerd à la fois un effet âge intrinsèque (a_i) un effet période(b_j) et un effet génération (c_k) dans un cas simple :

- effet linéaire de l'âge
- diminution brusque de l'effet période à partir de 1955
- augmentation brusque de l'effet génération à partir de la génération née en 1915.

Les figures 9 et 9 bis donnent les graphiques période à âge constant et génération à âge constant que l'on peut reconstituer. On s'aperçoit que les effets période et génération peuvent en effet être repérés dans chacun des graphiques : si "l'accident" effet période (respectivement génération) s'observe d'une manière synchrone sur le graphique période (respectivement génération),

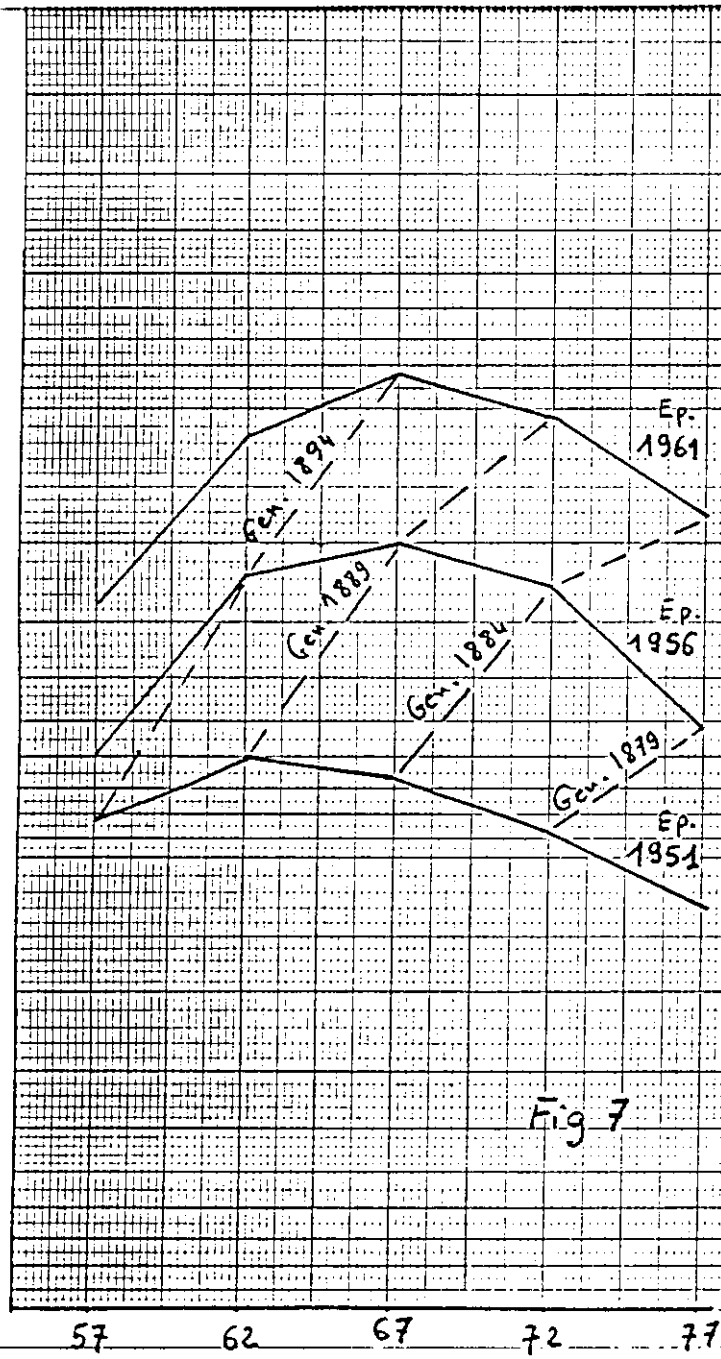

Ep. 1961
Gen. 1894
Gen. 1889
Ep. 1956
Gen. 1884
Gen. 1879
Ep. 1951
57
62
67
72
77

Fig 7

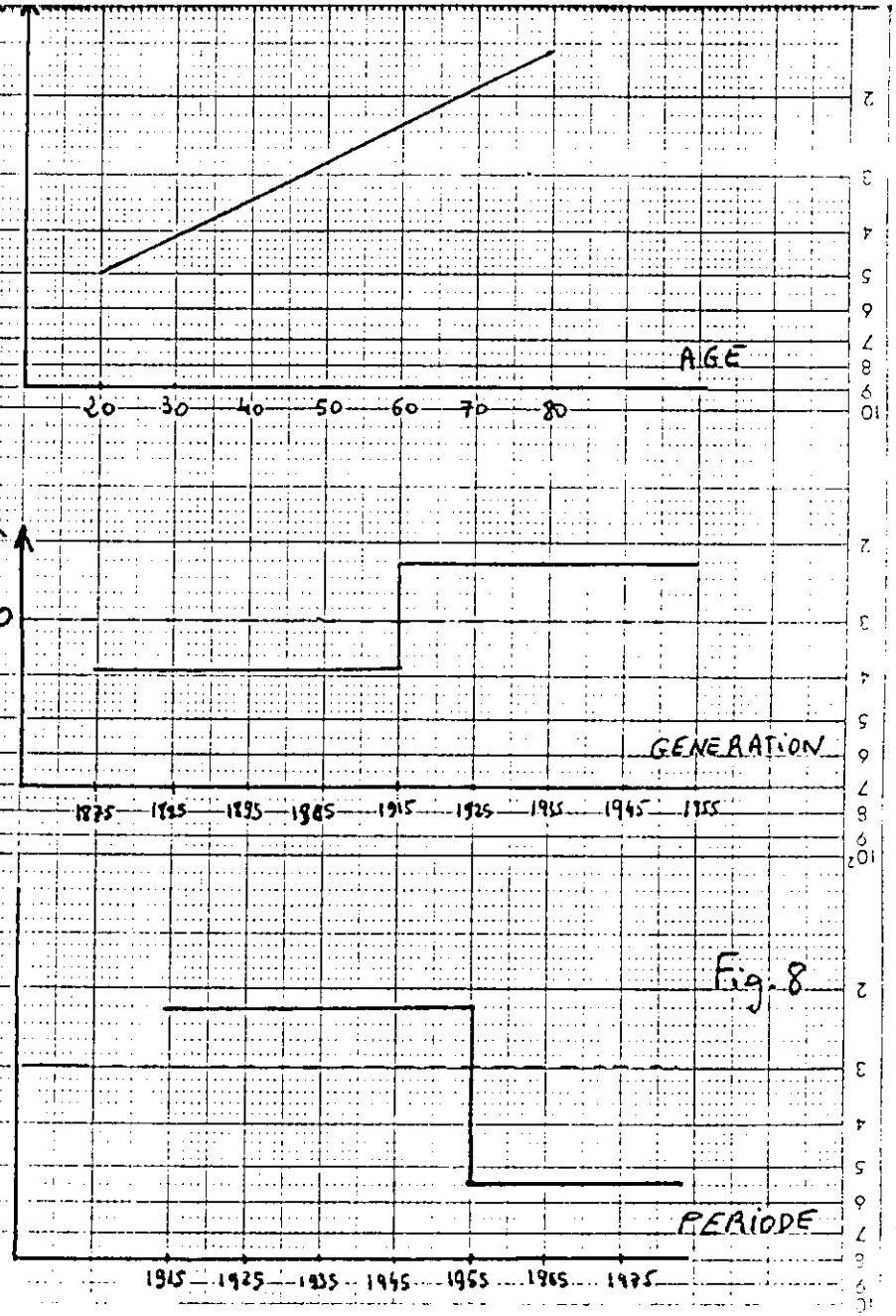

AGE
20
30
40
50
60
70
80
C_{β}
0
GENERATION
1875
1885
1895
1905
1915
1925
1935
1945
1955
b_j
0
PERIODE
1915
1925
1935
1945
1955
1965
1975

Fig. 8

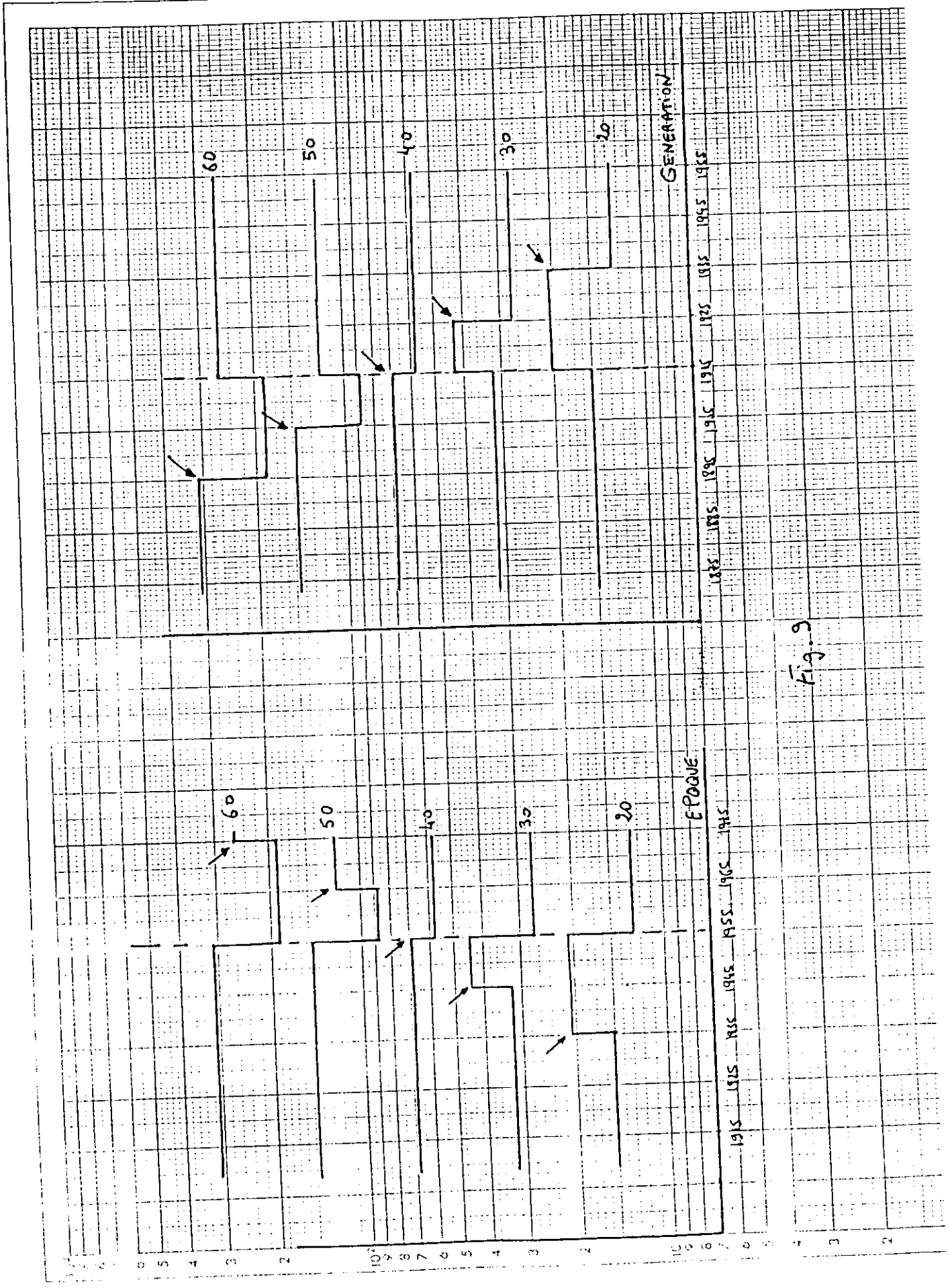

GENERATION
60
50
40
30
20
ÉPOQUE
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990

Fig. 6

l'effet génération s'observe décalé dans le temps en fonction des classes d'âge, l'accident étant visible à des époques d'autant plus anciennes que l'âge est bas. Il en est de même pour l'effet période dans le graphique génération mais cette fois-ci, l'accident est visible pour des générations d'autant plus récentes que l'âge est bas.

Cet exemple théorique montre la complémentarité des deux graphiques pour l'analyse des variations de mortalité en s'attachant à repérer les points caractéristiques des courbes (changements de pente, extrema ...) et leur concordance éventuelle sur l'un ou l'autre graphique. On conçoit qu'en l'absence de points caractéristiques sur les courbes des graphiques période ou génération à âge constant, l'identification d'éventuels effets devienne très hasardeuse. A la limite si les variations sont linéaires, il n'est théoriquement pas possible de distinguer un effet période et un effet génération, ainsi que le montre le raisonnement suivant : dans le modèle 3 une relation linéaire (à i constant) entre $\log t_{ij}$ et j peut s'interpréter également comme témoignant d'un effet période pur : $\log t_{ij} = a_i + b \times j$ ou d'un effet génération pur : $\log t_{ij} = a_i + c(i-1) + c(j-i+1) = a'_i + c \times k$ avec les notations précédentes.

Cette remarque a une conséquence importante : il n'est pas possible d'identifier sans ambiguité simultanément un effet période et un effet génération à partir des variations de mortalité, autrement dit les paramètres du modèle 3 ne peuvent pas être estimés de façon unique : à une solution constituée d'un ensemble de a_i, de b_j et de c_k, on peut substituer une solution équivalente en ajoutant une fonction linéaire arbitraire du temps aux a_i et c_k et en la retranchant des b_j.

Néanmoins il est possible d'analyser le modèle 3 par la méthode des moindres carrés définie plus haut.

- sur le plan de l'estimation des paramètres il convient en plus des deux contraintes :

$\sum_j b_j = 0$ et $\sum_k c_k = 0$ d'ajouter une contrainte supplémentaire qui lève l'indétermination précédente. Le programme GLIM fixe à zéro la dernière valeur des c_k.

- sur le plan du test d'adéquation du modèle, il n'existe cependant aucune difficulté, il est obtenu par un χ^2 dont le nombre de degrés de liberté s'obtient comme d'habitude par le nombre total d'observations (IJ) auquel on retranche le nombre total de paramètres auquel on rajoute le nombre de contraintes entre ces paramètres soit IJ - I - J - (I + J - 1) + 3 soit IJ - 2 (I + J) + 4. Il permet de tester si l'introduction d'un effet période dans un modèle qui contient déjà un effet génération est significative ou réciproquement.

Dans l'exemple de la mortalité par cancer du poumon, l'introduction de l'effet période au modèle avec effet génération fait passer le χ^2 d'ajustement (non significatif) valant 46,9 pour 40 ddl à la valeur de 23,2 pour 36 ddl ce qui représente un gain significatif (χ^2 = 23,7 pour 4 ddl). L'estimation des paramètre par le programme GLIM fournit une solution particulièrement non satisfaisante comme on pouvait le penser a priori par le choix de la contrainte $C_{16} = 0$ alors que l'analyse du seul effet génération avait montré que l'effet des générations les plus récentes est au contraire maximum.

Différents auteurs ont proposé des techniques pour tenter de dépasser cette difficulté et obtenir des solutions qu'ils souhaitent non ambigues (cf. C. OSMOND). Enfin des modèles plus particuliers ont été étudiés (cf. I.R. JAMES par exemple).

CHAPITRE II - REFERENCES RECENTES

- H.O. LANCASTER
  Cohort or generation methods. A priority from John BROWNLEE (1868-1927).
  Am. J. Epid. 1982, 115, 153-154.

- J.C. BARRETT
  A methods of mortality analysis : application to breast cancer.
  Rev. Epidem. et Santé Publ. 1978, 26, 419-425.

- N.M. VANDER HOFF
  Cohort analysis of lung cancer in the Netherlands
  Int. J. Epidem. 1979, 8, 41-47.

- M. SUSSER
  Periods effects, generations effects and age effects in peptic ulcer mortality
  J. Chron. Dis. 1982, 35, 29-40.

- C. OSMOND, M.J. GARDNER, E.D. ACHESON.
  Analysis of trends in cancer mortality in England and Wales during 1951-80 separating changes associated with period of birth and period of death.
  B.M.J. 1982, 284, 1005-1008.

- C. OSMOND, M.J. GARDNER
  Age, period and cohort models applied to cancer mortality rates
  Statistics in Medicine, 1982, 1, 245-259.

- I.R. JAMES, M.R. SEGAL
  On a method of mortality analysis incorporating age-year interactions with application to prostate cancer mortality
  Biometrics 1982, 38, 433-443.

CHAPITRE III.

MESURE DE LA MORBIDITE ET DE LA MORTALITE

II.Standardisation

2.1. Introduction : La standardisation pour quoi ?

Les mesures de morbidité introduites au chapitre précédent, prévalence incidence, incidence instantanée, correspondaient au simple "constat" d'une situation. Pour avancer dans l'analyse étiologique, l'épidémiologie procède comme on le sait de façon comparative :
"la mortalité pour cause de cancer de la vessie est-elle "trop" élevée chez les travailleurs du caoutchouc ?".

Vient alors immédiatement à l'esprit le problème du choix d'une population de référence et celui de la prise en compte des facteurs de risque susceptibles d'introduire une différence "artificielle" entre les populations étudiées et de référence.

La nécessité de définir des mesures de morbidité et de mortalité standardisées sur certains facteurs de risque tels que l'âge, le sexe, la catégorie socio-professionnelle afin d'en éliminer l'effet apparait ainsi comme un premier pas dans la direction de l'analyse étiologique.

Nous définirons dans un premier temps les méthodes de standardisation directe et indirecte qui sont celles que l'on utilise classiquement (§ 2.2.).

Nous comparerons ces deux méthodes du point de vue des informations sur lesquelles elles sont fondées, des quantités qu'elles estiment et de leur précision statistique (§ 2.3.).

Enfin, nous verrons comment on peut tester l'existence d'un écart significatif entre une mesure standardisée et une valeur de référence (§ 2.4.).

Afin d'adopter un langage plus concret, nous nous intéresserons dans ce chapitre au cas très fréquent de la standardisation de la mortalité sur l'âge. Les méthodes décrites pourront bien entendu être utilisées lorsque l'on souhaite standardiser des valeurs de mortalité ou de morbidité sur d'autres facteurs de risque (sexe, catégorie socio-professionnelle, etc...) pourvu que l'on dispose des informations nécessaires.

Par ailleurs, nous supposerons disposer d'observations sur un groupe de sujets que nous appellerons échantillon pour l'opposer à la population qui nous fournira des valeurs de référence concernant la distribution statistique des âges ou des mortalités par classe d'âges. Cette terminologie présente bien entendu un certain arbitraire, suivant les situations concrètes, l'"échantillon" pourra par exemple être constitué par l'ensemble des habitants d'une région, l'ensemble des salariés d'une entreprise, ou l'ensemble des sujets effectivement tirés au sort dans une population plus large.

La population de référence pourra elle aussi varier d'un cas particulier à l'autre : ensemble des habitants d'un pays, d'une région, cohorte fictive subissant des taux de mortalité par classe d'âges constatés à un moment donné sur des sujets de cohortes différentes d'une région.

2.2. Les deux méthodes de standardisation : méthode directe, méthode indirecte

Nous supposerons avoir observé dans un échantillon des taux de mortalité[(a)] t_i $i = 1,2,....I$ dans différentes classes d'âges. Par ailleurs, nous noterons $w_i^{(*)}$ $i = 1,2....I$ la distribution statistique des âges dans la population de référence[(b)].

Le taux de mortalité standardisé par la méthode directe est la moyenne des taux (t_i) de l'échantillon pondéré p par la distribution d'âge $(w_i^{(*)})$ dans la population de référence. Ce taux s'écrit donc :

$$t_{SD} = \sum_{i=1}^{I} w_i^* . t_i \qquad (1)$$

Dans la standardisation par la méthode indirecte, on dispose au contraire de valeurs des taux de mortalité par classe d'âges dans la population de référence t_i^* $i = 1,.....,I$ et on utilise la distribution des âges de l'échantillon $(n_i$ $i = 1,....,I)$[(c)] pour calculer un nombre de décès attendu :

$$\mathcal{E} = \sum_{i=1}^{I} n_i \times t_i^* \qquad (2)$$

(a) Nous ne préciserons par ici si la mesure de mortalité adoptée est une incidence instantanée ou non, par ailleurs ce qui sera développé à propos des incidences pourra en fait être repris dans le cas de prévalence.

(b) Remarque: Il peut s'agir d'une répartition de sujets à proprement parler avec $w_i = n_i / (\sum n_i)$ comme d'une répartition de personnes-années, avec $w_i = PA_i / (\sum PA_i)$.

(c) Remarque: n_i peut être un nombre de sujets ou de personnes x années.

qui tient compte de la structure d'âge de l'échantillon. On rapporte ensuite le nombre total de décès observés dans l'échantillon "O" à cette valeur pour définir le "Rapport Comparatif de Mortalité"(§).

$$SMR = 100 \times \frac{O}{E} \qquad (3)$$

(§)*Remarque: Noté SMR compte tenu de l'expression Standardized Mortality Ratio.*

2.3. Comparaison des deux méthodes de standardisation

Les informations nécessaires au calcul du taux de mortalité standardisé (*) sont plus détaillées que celles qui servent de base au calcul du rapport comparatif de mortalité puisque l'on doit connaître non seulement la répartition des sujets de l'échantillon par classe d'âges mais encore celle des décès. Dans la plupart des situations on dispose simultanément des deux informations, on peut toutefois imaginer des situations où tel n'est pas le cas. Ainsi, si l'on veut standardiser sur l'âge maternel à l'accouchement l'incidence d'un certain type de malformations à la naissance dans une région donnée, on est contraint d'utiliser la méthode indirecte si les seules informations statistiques dont on dispose sont le nombre total de naissances d'enfants atteints de cette malformation et la répartition de l'ensemble des naissances par classe d'âges maternels.

Pour comparer les quantités estimées par le taux de mortalité standardisée et le rapport comparatif de mortalité, nous nous ramènerons à deux quantités proportionnelles ayant la même "dimension", celle du rapport de deux taux :

- le rapport RD : $\frac{t_{SD}}{t^*}$ de la mortalité standardisée à la mortalité t^* de la population de référence ;

- le rapport RI : $\frac{O}{E}$ qui diffère des rapports comparatifs de mortalité par le seul facteur 100.

Le taux de mortalité "exact" estimé par le taux t_i de l'échantillon sera noté T_i et le rapport $R_i = T_i / t^*_i$ de ce taux à la valeur correspondante dans la population de référence est le risque relatif de mortalité dans la ième classe d'âge.

(*) on sous-entend en général "par la méthode directe"

En remarquant que :

$$t^* = \sum_i w^*_i . t^*_i \quad \text{et} \quad E(t_i) = T_i = RR_i . t^*_i$$

l'espérance de RD s'écrit :

(4) $$E(RD) = \frac{E[\sum_i w^*_i . t_i]}{t^*} = \frac{\sum_i w^*_i . t^*_i . RR_i}{\sum w^*_i . t^*_i}$$

Pour obtenir l'espérance de RI on notera que les nombres observés et attendus de décès de l'échantillon peuvent s'écrire :

$$O = \sum_i O_i \overset{(\S)}{=} \sum n_i . \frac{O_i}{n_i} = n . \sum_i w_i . t_i$$

$$E = \sum_i n_i . t^*_i = n . \sum_i w_i . t^*_i$$

et on a:

(5) $$E(RI) = \frac{E[O]}{E} = \frac{n . \sum_i w_i E(t_i)}{n . \sum_i w_i t^*_i} = \frac{\sum w_i t^*_i RR_i}{\sum w_i t^*_i}$$

(§) Remarque: où n_i est le dénominateur de t_i, c'est à dire un nombre de sujets ou de personnes années et $n = \sum n_i$

Si le risque relatif prend des valeurs identiques dans les différentes classes d'âges : $RR_i = RR$ les expressions (4) et (5) ci-dessus prennent des valeurs identiques et égales à cette valeur commune RR :

$$\left.\begin{aligned} E(RD) &= RR \\ E(RI) &= RR \end{aligned}\right\}$$

Les méthodes directes et indirectes estiment donc dans ce cas essentiellement la même quantité qui est le risque relatif "âge spécifique" de l'échantillon par rapport à la population de référence.

Dans le cas contraire, on dit qu'il existe une "interaction"(§) entre l'âge et le facteur de risque distinguant l'échantillon de la population de référence. L'utilisation d'un unique indice mesurant l'écart de la mortalité dans l'échantillon par rapport à la population de référence n'a plus alors de sens clair sinon celui d'un risque relatif "moyen". La standardisation par la méthode directe est alors la seule qui élimine à proprement parler l'effet de la structure d'âge de l'échantillon, on constate en effet en comparant les expressions (4) et (5) :

- que l'espérance de RD est une moyenne pondérée des risques relatifs âge-spécifiques où les poids ($\omega_i^* t_i^*$) ne dépendent pas de l'échantillon mais de la seule population de référence,

- que l'espérance de RI estime également une moyenne pondérée des risques relatifs âge-spécifiques mais que les poids ($\omega_i t_i^*$) dépendent en fait de la structure d'âge de l'échantillon.

S'agissant des précisions statistiques des deux estimateurs RD et RI, on peut montrer que la variance d'échantillonnage de RD est toujours supérieure ou égale à celle de RI, l'égalité étant uniquement atteinte lorsque la répartition d'âge de l'échantillon est identique à celle de la population de référence.

(§)Remarque: cette notion sera définie plus précisément au chapitre 5.

La standardisation par la méthode indirecte est donc plus précise. On comprend d'ailleurs intuitivement que la pondération des mortalités par classes d'âge de la méthode directe peut conduire à une très mauvaise précision statistique s'il existe une forte disparité entre les structures d'âge de l'échantillon et de la population de référence : dans une classe d'âge donnée, une mortalité établie sur un très faible nombre de sujets peut en effet recevoir un poids important si cette classe d'âge est fréquente dans la population. A l'inverse, dans la méthode indirecte, le nombre observé de décès, $O = \sum o_i = \sum_i n_i \times t_i$ donne à chaque estimation t_i un poids qui est inversement proportionnel à sa variance.

La méthode indirecte est donc plus précise que la méthode directe et ne requiert pas de connaître la répartition des décès par classe d'âge ; elle présente pour seul inconvénient de ne pas supprimer totalement l'influence de la structure d'âge de l'échantillon s'il existe une interaction.

2.4. Comparaison de la mortalité observée à une valeur de référence

Celle-ci peut être effectuée en utilisant un χ^2 à 1 ddl :

$$\chi^2_1 = \frac{(O-E)^2}{E}$$

où les quantités O et E nombre observés et attendus de décès ont été définis au § 2.1. ci-dessus.

CHAPITRE IV

RELATION ENTRE LA MALADIE ET L'EXPOSITION A UN FACTEUR

Nous dirons très généralement qu'il existe une relation entre l'exposition à un facteur et la maladie si cette exposition est associée à ("entraîne") une augmentation de la probabilité d'être malade. Nous supposons implicitement que la maladie "apparaît" sous la forme d'un événement en tout ou rien et ceci sera commun à l'ensemble des chapitres ultérieurs. D'autre part, nous supposons que l'exposition du facteur ne joue pas un rôle protecteur mais tout ce que nous dirons s'applique également dans ce cas. Si cette définition théorique de l'existence d'une relation n'est guère ambiguë, la mesure de cette relation va dépendre du choix d'un modèle qui lui même peut être lié à la nature de l'étude entreprise Bien que le raisonnement soit finalement identique quel que soit le type d'exposition étudiée (qualitatif, quantitatif) nous traiterons successivement les différents cas.

1. L'EXPOSITION AU FACTEUR EST QUALITATIVE

Prenons l'exemple de la recherche d'une relation entre l'exposition de femmes enceintes à un toxique et l'existence d'anomalies congénitales chez les enfants. Il est possible d'imaginer simplement une étude permettant d'étudier l'existence d'une relation,en effet étant donné un groupe de femmes enceintes, la présence ou non d'anomalies congénitales chez les enfants peut être établie dès leur naissance et sa probabilité est estimée par le pourcentage d'enfants porteurs d'anomalies. Il suffira d'estimer cette probabilité pour un groupe de femmes exposées et non exposées pour répondre à la question.

Dans cet exemple, le temps en quelque sorte n'intervient pas puisque dès qu'une mère est recrutée dans l'étude, on sait "définitivement" (éventuellement à quelques mois près) si son enfant est "malade" ou non.

Nous allons étudier dans un premier temps cette situation.

../...

1. 1 - Les événements sont observés à "date fixe"

Nous appellerons P_1 et P_o les probabilités que l'enfant soit atteint selon que la mère était ou non exposée. L'existence d'une relation implique alors $P_1 > P_o$. L'intensité de cette relation est d'autant plus grande que P_1 est grand par rapport à P_o.

Le modèle additif exprime la relation par la différence $P_1 - P_o = \Delta$ appelée "risque en excès" (excess risk)

Le modèle multiplicatif implique quant à lui que l'exposition au facteur multiplie la probabilité de la maladie par un coefficient $R = \frac{P_1}{P_o}$ appelé "risque relatif"

Cependant bien d'autres choix peuvent être faits pour exprimer la relation, le plus important pour la suite, correspond également à un modèle multiplicatif :

$\psi = \frac{P_1}{1-P_1} \Big/ \frac{P_o}{1-P_o}$ l'"odds ratio" représente le rapport des "chances" relatives d'être malade ou ou non malade selon l'exposition. Le choix entre ces diverses mesures ne peut être guidé que par des considérations propres aux problème étudié : par exemple il peut exister des modèles "explicatifs" de la relation (ce cas est hélas rare) qui pourraient impliquer une forme particulière de la mesure. Plus habituellement le choix est fait empiriquement : la stabilité de la mesure lorsque la relation est étudiée dans des populations différentes ou dans une même population en fonction de caractéristiques diverses des sujets est l'argument essentiel. Dans la plupart des cas, les modèles multiplicatifs vérifient beaucoup mieux ce critère que le modèle additif et leur emploi est de ce fait très général.

1. 1.1 - Nature de l'étude

Le recrutement des sujets dans l'exemple étudié ici peut s'effectuer de plusieurs manières et classiquement on distingue trois cas :

. On recrute un échantillon représentatif de femmes enceintes (ou de mères récentes) d'une population que l'on classe selon l'exposition et selon que leur enfant porte ou non une anomalie (type I). On a les notations suivantes pour les probabilités qu'une mère de la population soit dans chacune des catégories.

../...

	E^+	E^-	
M^+	$p\ P_1$	$(1-p)\ P_0$	P
M^-	$p(1-P_1)$	$(1-p)(1-P_0)$	$1-P$
	p	$1-p$	

. On recrute un échantillon de femmes exposées et non exposées (type II)

	E^+	E^-
M^+	P_1	P_0
M^-	$1-P_1$	$1-P_0$
	1	1

Dans chacun de ces cas l'étude permet d'atteindre les quantités P_1 et P_0 donc permet le choix de n'importe quelle mesure de la relation entre l'exposition et la maladie.

. On recrute un échantillon d'enfants atteints et non atteints et les mères sont classées en fonction de leur exposition (type III ou étude "cas-témoin").

	E^+	E^-	
M^+	p_1	$1-p_1$	1
M^-	p_0	$1-p_0$	1

L'étude ne permet pas d'avoir accès aux quantités P_1 et P_0 sans information complémentaire (par exemple la fréquence P de la maladie ou p de l'exposition). Cependant, une mesure particulière de la relation peut être obtenue à partir de p_1 et p_0 : il s'agit de "l'odds ratio" en effet lorsque l'information est complète on a :

$$p_1 = \frac{p\ P_1}{P} \quad \text{et} \quad p_0 = \frac{p\ (1-P_1)}{1-P}$$

et

$$\frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}} = \frac{p\ P_1}{P-p\ P_1} \Big/ \frac{p\ (1-P_1)}{1-P-p+p\ P_1} = \frac{P_1}{1-P_1} \Big/ \frac{P-pP_1}{1-P-p+pP_1} = \frac{P_1}{1-P_1} \Big/ \frac{P_0}{1-P_0} = \Psi$$

../...

On peut remarquer à cette occasion que lorsque P_1 et P_0 sont"petits" c'est à dire lorsque la maladie est rare, "l'odds ratio" est approximativement égal au risque relatif :

$$\Psi = R + \frac{R-1}{R} P_1 + \ldots$$

On obtient donc le résultat habituel : dans une étude de type III l'odds ratio peut être estimé et si la maladie est rare c'est une approximation du risque relatif (ce dernier résultat est cependant indépendant de la nature de l'étude).

1. 1.2 - Autres expressions de la relation exposition-maladie

Les mesures de la relation que nous venons de voir expriment "l'intensité du lien étiologique" entre le facteur et la maladie. Bien d'autres mesures ont été proposées dont certaines visent à calculer la proportion de cas "dûs" à la maladie.

Nous appellerons "fraction étiologique" la proportion de cas chez les exposés qui seraient évités si l'exposition était supprimée ... et si la relation était causale !

$$FE = \frac{p\,P_1 - p\,P_0}{p\,P_1} = 1 - \frac{1}{R}$$

il ne s'agit donc pas d'une mesure nouvelle en elle-même.

Par contre, la proportion de cas de l'ensemble de la population qui dans les mêmes conditions seraient évités a un sens très concret sur le plan de la santé publique : elle est appelée "risque attribuable".

$$RA = \frac{p\,P_1 - p\,P_0}{p\,P_1 + (1-p)\,P_0} = \frac{p\,(R-1)}{p\,R + 1-p} = \frac{P_1 - p}{1-p}$$

RA ne peut donc être obtenu que dans les études de type I sans autre information. Dans les études de type III et si la maladie est rare dans la population, on peut approximer FE et RA par :

$$FE \;\#\!\#\; 1 - \frac{1}{\Psi} \qquad \text{et} \qquad RA \;\#\!\#\; \frac{P_1 - P_0}{1 - P_0}$$

1. 1.3 - Ordre de grandeur des différentes approximations

Le tableau donne cet ordre de grandeur dans le cas d'un risque relatif "moyen" (R = 2) et "fort" (R = 10) en fonction de la fréquence P de la maladie et celle p de l'exposition au facteur.

Alors que RA' = $\frac{p_1 - p_0}{1 - p_0}$ est toujours de l'ordre de grandeur de RA dans les limites des valeurs des paramètres que l'on a fixées, Ψ devient nettement trop grand par rapport à R (≥ 10%) dès que la fréquence de la maladie atteint 0.10 et que celle du facteur est inférieure à 0.10 pour R = 2 ou quelle que soit celle du facteur pour R = 10.

R = 2

p =		P = 10^{-3}	P = 10^{-2}	P = 10^{-1}
0.10 (RA= 0.09)	Ψ	2.00	2.02	2.22
	RA'	0.09	0.09	0.10
0.50 (RA= 0.33)	Ψ	2.00	2.01	2.15
	RA'	0.33	0.34	0.36
0.80 (RA= 0.44)	Ψ	2.00	2.01	2.13
	RA'	0.45	0.45	0.47

R = 10

p =		P = 10^{-3}	P = 10^{-2}	P = 10^{-1}
0.10 (RA= 0.47	Ψ	10.05	10.50	20.00
	RA'	0.47	0.48	0.50
0.50 (RA= 0.82)	Ψ	10.02	10.17	12.00
	RA'	0.82	0.82	0.83
0.80 (RA= 0.87)	Ψ	10.01	10.11	11.25
	RA'	0.88	0.88	0.89

../...

1. 1.4 - Cas d'un facteur qualitatif à plusieurs classes :

Il peut s'agir d'un facteur authentiquement qualitatif (type de toxique de l'exposition par exemple) ou quantitatif en réalité mais traité en classes (degré d'exposition par exemple). Dans les deux cas une classe est choisie comme référence (classe 0) et la mesure de la relation de chaque classe d'exposition par rapport à la classe 0 peut être calculée. C'est l'ensembl des k-1 mesures obtenues qui exprime la relation cherchée.

Si on réunit les k-1 classes d'exposition que l'on compare à la classe 0 (ce qui revient à rendre dichotomique l'exposition) la mesure de la relation que l'on obtient ainsi n'a plus la qualité de mesure "étiologique" car elle dépend de la distribution des différentes classes d'exposition ainsi que le montrent les résultats suivants :

. Soit $\pi_0, \pi_1 \ldots \pi_{k-1}$ les fréquences d'exposition des classes du facteur dans la population (type I).

alors Δ = risque en excès de toutes les classes d'exposition par rapport à la classe 0.

$$\Delta = \sum_{i=0}^{k-1} \pi_i \Delta_i$$ où Δ_i est le risque en excès de la classe i/ classe 0.

de même $$R = \frac{\sum_{i=1}^{k-1} \pi_i R_i}{1-\pi_0}$$ où Ri est le risque relatif de la classe i/ classe 0.

. Si on choisit l'odds ratio (en particulier si étude du type III), l'odds ratio après réunion des classes s'écrit en fonction de celui de chaque classe de la même façon que R mais $\pi'_0 \ldots \pi'_{k-1}$ sont alors les fréquences respectives des classes d'exposition dans le groupe non-malade

$$\Psi = \frac{\sum_{i=1}^{k-1} \pi'_i \psi_i}{1-\pi'_0}$$

. Si la mesure choisie est le risque attribuable, après réunion des classes d'exposition, il est bien entendu la somme des risques attribuables de chaque classe.

$$RA = \sum_{i=1}^{k-1} RA_i$$

../...

Exemple numérique

Etude cas témoins sur la relation entre le cancer des voies aériennes supérieures et le tabagisme.

	Malades	Témoins	Ψi	$\pi i'$	RAi' *
non fumeurs	26	83	-	0.190	-
1-19 c/j	66	97	2.2	0.217	0.07
20-39 c/j	248	197	4.1	0.441	0.39
⩾ 40 c/j	143	68	6.9	0.152	0.25
tous fumeurs	457	362	$\Psi = \frac{2.2 \times 0.217 + ..}{1-0.19}$ =	4.1	0.71

* Ce dernier calcul suppose la maladie rare ce qui est le cas ici.

1. 2 - Les évènements sont observés lors de la "surveillance d'une cohorte"

La notion de "probabilité d'être malade" qui permet de définir la relation exposition-maladie doit être précisée puisque cette probabilité dépend de la durée d'observation des sujets. C'est $\lambda(t)$ incidence instantanée qui est alor le paramètre d'intérêt. Les études les plus classiques pour lequel le recrutement des sujets est effectué par la surveillance d'une cohorte sont les études prospectives de type I (étude prospective sur les cardiopathies ischémiques par exemple) mais ce ne sont pas les seules.

1. 2.1 - Choix d'un modèle

On rappelle que l'incidence instantanée d'une maladie lors de la surveillance d'une cohorte s'écrit :

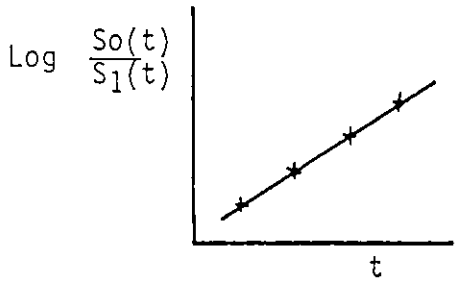
$$\lambda(t) = \frac{-\, dS(t)}{S(t)\, dt}$$ où $S(t)$ est la probabilité de ne pas être encore atteint au temps t de surveillance (fonction survie dans le cas d'une étude de mortalité). Rappelons que cette fonction S(t) s'écrit :

$$\exp - \int_0^t \lambda(t)dt$$

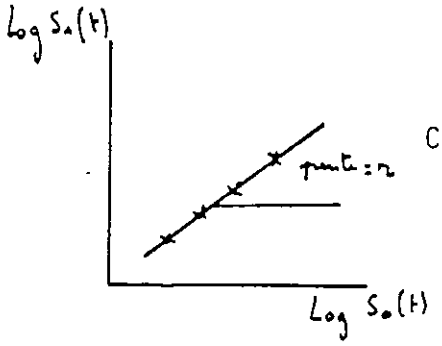
../...

Il y a relation entre l'exposition et la maladie si $\lambda_1(t)$ (exposés) $>\lambda_0(t)$ (non exposés).

Le choix d'une mesure de la relation est dans ce cas essentiellement guidé par le fait qu'il est souhaitable qu'elle soit indépendante du temps. Par exemple la quantité $\delta = \lambda_1(t) - \lambda_0(t)$ peut être choisie (modèle additif) si elle demeure constante en fonction de t, on pourra l'appeler le risque instantané en excès. Dans ce cas les courbes de "survie" des exposés et non exposés doivent être telles que :



Plus généralement, c'est la quantité $r = \frac{\lambda_1(t)}{\lambda_0(t)}$ qui demeure constant conduisant à l'emploi du risque relatif instantané r. Les courbes de survie vérifient alors :



Ce modèle s'appelle modèle de COX.

1. 2.2 - Retour au cas précédent

Lorsque la surveillance s'est déroulée durant T années pour tous le sujets de l'étude, on peut se ramener au cas du paragraphe 1.1 en considérant la probabilité d'être malade en T années comme le paramètre d'intérêt, soit P_1 dans le groupe exposé et P_0 dans le groupe non exposé. Dans le cas du modèle multiplicatif ci-dessus de paramètre r, on s'aperçoit que r ne s'écrit pas simplement en fonction de :

$$R = \frac{P_1}{P_0} \qquad \text{mais} \quad r = \frac{\text{Log}(1-P_1)}{\text{Log}(1-P_0)}$$

Le risque relatif après T années (R) n'est donc une approximation du risque relatif instantané r que si P_1 et P_0 sont petits, c'est à dire si l'incidence de la maladie est faible ou la durée de surveillance petite. Mais dans ce cas l'odds ratio Ψ calculé dans les mêmes conditions, étant voisin de R, est aussi une approximation de r. Plus précisément, il est aisé de voir que $R<r<\Psi$.

1. 2.3 - Ordre de grandeur des approximations

Il est donné au tableau suivant en fonction de la fréquence de la maladie dans le groupe non exposé après T années (P_0) et le risque relatif instantané r.

r =	P_0 =	10^{-3}	10^{-1}	3.10^{-1}
2	R	1.99	1.90	1.70
	Ψ	2.01	2.11	2.43
5	R	4.90	4.10	2.77
	Ψ	5.10	6.24	11.55
10	R	9.56	6.51	3.24
	Ψ	10.47	16.81	80.27

1. 2.4 - Nature de l'étude

. Le cas classique est celui rappelé en introduction de ce chapitre : une populati "non malade" est suivie et les évènements sont identifiés et leur date connue ainsi que l'éventuelle exposition au facteur pour tous les sujets (type I). Tout aussi classique est l'étude de type II constituée de la surveillance, simultanée en principe, d'une cohorte de sujets exposés et d'une cohorte de sujets non exposés.(type II)

L'information fournie permet d'étudier la mesure de la relation la plus satisfaisante entre l'exposition et la maladie c'est à dire le risque relatif instantané r.

../...

En pratique cependant le problème de calculer r se pose, ce sera fait dans un chapitre ultérieur (modèles de survie). Une estimation simple (qui n'est pas en général la meilleure) est néanmoins classique : en effet les quantités Log S(t) peuvent se mettre sous la forme :

$$\int_0^t \lambda(u)\,du$$ incidence "cumulée"

Si tous les sujets sont suivis pendant un temps T, on peut écrire :

$$r = \frac{\log S_1(T)}{\log S_2(T)} = \frac{\int_0^T \lambda_1(u)\,du}{\int_0^T \lambda_0(u)\,du} = \frac{\frac{1}{T}\int_0^T \lambda_1(u)\,du}{\frac{1}{T}\int_0^T \lambda_0(u)\,du}$$

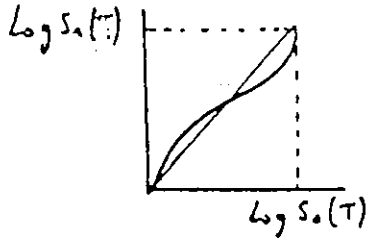
où on reconnait l'incidence "moyenne" durant T dans les groupes exposés et non exposés. On obtient une estimation de r en estimant ces incidences moyennes :

$$\frac{m}{T(n-m)+\sum_m t_i}$$

où le dénominateur est la somme des personnes x années des sujets de l'étude "à risque"

Comme on l'a montré dans un chapitre précédent, c'est l'estimation du maximum de vraisemblance dans le cas où dans chacun des deux groupes la survie est supposée exponentielle

Remarquons que cette méthode d'estimation de r se représente simplement :



. Le second cas est celui d'une étude de type I et II dans laquelle les malades sont identifiés au temps T sans connaissance des dates individuelles. On ne connait donc que les paramètres P_1 et P_0 (pourcentages de malades dans les deux groupes). Le risque relatif R ou l'odds ratio Ψ fournissent néanmoins une approximation de r si P_1 et P_0 sont faibles.

../...

. Enfin il peut se produire que l'on ne connaisse l'exposition que pour les cas observés à l'instant T et pour un échantillon des non malades (cas d'une enquête cas-témoin réconstituée à partir d'une étude cohorte). L'odds ratio Ψ est une approximation de r dans les mêmes conditions que précédemment.

1.3 - Les évènements sont observés dans une population en "état stationnaire".

Dans une étude cas-témoin (type III) le recrutement des malades et des témoins n'est pas en général éffectué dans une cohorte aisément identifiable mais plutôt dans une "grande population" constituée de malades et de non malades dont les flux d'entrée et de sortie peuvent être considérés comme constants durant la période de recrutement. On distinguera deux cas selon que les cas sont recrutés lors de leur apparition dans la population ("cas incidents") ou s'il s'agit de cas existants ("cas prévalents"). Dans chacun de ces cas, un raisonnement simple indique que l'odds ratio de l'étude cas-témoin pourrait, sous certaines conditions, représenter le risque relatif instantané de maladie.

1. 3.1 - Recueil de "cas incidents"

Un bon exemple de cette situation est la planification d'une étude cas-témoins effectuée à partir d'un "registre" de la maladie étudiée dans une même zone géographique, par exemple un registre d'infarctus du myocarde dans un département. Les cas de l'étude sont recrutés lors de leur apparition dans la population supposée stationnaire par exemple des hommes de 40 à 65 ans habitant le département. Remarquons qu'une définition relativement précise de la "grande population" peut ici être donnée.

Appelons I l'incidence de la maladie dans la population. Remarquons que I ne répond pas en toute rigueur à la définition que nous avons donnée de l'incidence instantanée i dans une étude de cohortes puisque la population se renouvelle au cours du temps et ne se dirige pas vers "l'extinction" comme pour une cohorte fixée au départ de l'étude. Il s'agit tout simplement du rapport entre le nombre de nouveaux cas observés par unité de temps au nombre N de sujets de la population susceptibles de devenir malades (effectif total de la population auquel on a soustrait le nombre de cas déjà malades). L'hypothèse de "stationnarité" conduit à considérer que I et N sont des constantes au cours de l'étude. Par définition le recrutement de M malades s'effectuera durant un temps t tel que :

$$M = I\ N\ t$$

Appelons p0 la proportion de sujets exposés dans la population de N sujets susceptibles de devenir malades, (c'est aussi la proportion d'exposés chez les sujets témoins de l'étude), I1 et I0 respectivement l'incidence des sujets

../...

exposés et non exposés et $\rho = I1/I0$ le risque relatif ainsi défini.

En écrivant que les M malades se recrutent d'une part chez les exposés et chez les non exposés, on a :

$$M = p0 \quad I1 \; N \; t + (1-p0) \; I0 \; N \; t$$

On en déduit la proportion d'exposés chez les malades :

$$p1 = p0 \quad \rho \; / \; (p0 \quad \rho + 1-p0)$$

puis :

$$\rho = (p1/(1-p1)/(p0/(1-p0)$$

On reconnait au deuxième membre de cette formule l'odds-ratio OR de l'étude cas-témoins. Dans un tel modèle l'odds ratio est donc égal au risque relatif dans la population supposée stationnaire. Il reste maintenant à étudier dans quelle mesure le risque relatif ρ représente le risque relatif instantané r tel qu'il aurait été obtenu par une étude de cohortes.

Appelons a et b les limites d'âge définissant la population, i (u) l'incidence instantanée de la maladie à l'âge u que l'on suppose indépendante du temps t et $\pi(u)$ la distribution de l'âge dans la population (densité de probabilité) distribution qui ne dépend pas du temps t à cause de l'hypothèse de stationnarité.

L'incidence I par unité de temps s'écrit :

$$\int_0^1 dt \int_a^b \pi(u) \; i(u) \; du = \int_a^b \pi(u) \; i(u) \; du$$

Le risque relatif instantané s'écrit :

$$r = \frac{i_1(u)}{i_0(u)} \quad \text{supposé indépendant de u (modèle de COX)}$$

On a donc :

$$\rho = \frac{I_1}{I_0} = \frac{r \int_a^b \pi_1(u) \quad i_0(u) \; du}{\int_a^b \pi_0(u) \quad i_0(u) \; du}$$

../...

Le risque relatif ρ est égal au risque relatif instantané r si la distribution de l'âge dans la tranche (a,b) est la même chez les exposés et les non exposés. On obtient donc le résultat suivant : l'odds ratio "ajusté par âge" de l'étude cas-témoins est égal au risque relatif instantané et bien entendu si l'exposition au facteur est indépendante de l'âge, ceci est vrai pour l'odds ratio lui-même.

1. 3.2 - Recrutement de "cas prévalents"

Si on souhaite étudier par une étude cas-témoins la relation entre le "rhumatisme" et l'exposition à un facteur, le modèle de recrutement précédent ne peut évidemment s'appliquer : le moment où débute la maladie est difficilement identifiable et la durée de la maladie est très longue. Par contre, l'échantillon de malades sera aisément constitué de sujets atteints dans la population au moment du recrutement.

Il est néanmoins possible à partir de la prévalence de la maladie de définir une incidence ~~instantanée~~ si on suppose que la population est stationnaire au sens où on l'a défini plus haut, en effet on a alors la relation :

$$P = \frac{I\,\bar{d}}{1 + I\,\bar{d}} \quad \text{où } \bar{d} \text{ est la durée moyenne de la maladie}$$

En appliquant ce résultat aux exposés et non exposés on a :

$$\frac{P_1}{1-P_1} = I_1\,\bar{d}_1 \quad \text{et} \quad \frac{P_0}{1-P_0} = I_0\,\bar{d}_0$$

Si on suppose que malades exposés et malades non exposés ont en moyenne la même durée de maladie :

$$\rho = \frac{I_1}{I_0} = \frac{P_1}{1-P_1} \Big/ \frac{P_0}{1-P_0} = \frac{P_1}{1-P_1} \Big/ \frac{P_0}{1-P_0} = \Psi$$

avec les notations habituelles.

Ceci montre que dans une étude cas témoins ayant un recrutement de "cas prévalents", l'odds ratio Ψ est égal au risque relatif instantané sous l'hypothèse d'indépendance entre l'exposition et la durée de la maladie.

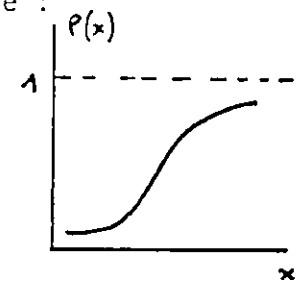
../...

2. L'EXPOSITION AU FACTEUR EST QUANTITATIVE

Nous avons dit en introduction que les développements faits dans le cas où l'exposition au facteur est qualitatif s'appliquait intégralement au cas quantitatif. Nous ne ferons donc dans ce chapitre que reprendre ce qui a été dit plus haut en introduisant de nouvelles notations.

2. 1 - Les évènements sont observés à "date fixe"

Appelons P(x) la probabilité d'être malade si l'exposition est x (une dose par exemple). Il y a relation entre la maladie et l'exposition si P(x) dépend de x et sauf cas particulier pour lequel il existerait un modèle explicatif qui stipulerait une relation autre, P(x) sera considéré comme une fonction croissante de x donc de la forme :



Définir la relation entre la maladie et l'exposition x revient à modéliser P(x).

Parmi les modèles les plus simples, on distinguera :

. $P(x) = ax + b$ modèle appelé additif (1)

. $\text{Log } P(x) = ax + b$ modèle multiplicatif (2)

. $\text{Log } \frac{P(x)}{1-P(x)} = ax + b$ autre modèle multiplicatif appelé modèle logistique (3)

qui peut également s'écrire :

$$P(x) = \frac{k \exp ax}{1 + k \exp ax} \quad \text{avec } k = \exp b$$

Remarquons que sur le seul plan mathématique, le modèle logistique a l'avantage de permettre à l'exposition x de prendre une valeur absolument quelconque.

Si, en réalité, l'exposition est dichotomique (x = o ou 1) on retrouve les modèles étudiés précédemment comme cas particuliers :

en effet Δ = risque en excès = $P_1 - P_0 = a$ (1)

R = risque relatif = $\frac{P_1}{P_0} = \exp(a+b) / \exp b = e^a$ (2)

Ψ = odds ratio = $\frac{P_1}{1-P_1} \Big/ \frac{P_0}{1-P_0} = e^a$ (3)

Là encore, la possibilité de connaître les paramètres a et b des modèles ci-dessus dépend de la nature de l'étude.

. Dans une étude de type I ou II, on peut atteindre directement la probabilité d'être malade si l'exposition est x soit P(x), on peut donc estimer les paramètres de n'importe quel modèle.

. Par contre, dans une étude cas témoin (type III) ceci n'est pas possible. En effet l'information ne permet que de connaître les distributions de x chez les malades $f_1(x)$ et chez les témoins $f_0(x)$. Appelons f(x) la distribution du facteur x dans l'ensemble de la population; en appliquant le théorème de Bayes on peut écrire :

$$f_1(x) = \frac{P(x)\ f(x)}{P} \quad \text{et} \quad f_0(x) = \frac{(1-P(x))\ f(x)}{1-P}$$

où P, probabilité inconditionnelle d'être malade (fréquence dans la population) doit vérifier : $P = \int P(x)\ f(x)\ dx$

Des relations précédentes on tire : $\frac{P(x)}{1-P(x)} \Big/ \frac{P}{1-P} = \frac{f_1(x)}{f_0(x)}$

En conclusion, dans une étude de type III, la connaissance de $f_1(x)$ et de $f_0(x)$ permet de connaître le rapport $\frac{P(x)}{1-P(x)}$ à une constante multiplicative près. (P fréquence dans la population n'est pas connue). Seul le modèle logistique (3) peut donc être appliqué à cette étude et seul le paramètre a peut être estimé sans information complémentaire. (on retrouve le fait que lorsque le facteur est dichotomique, seul l'odds ratio peut être atteint). Remarquons que le modèle logistique, si cela était nécessaire, pourrait s'écrire plus généralement :

$\text{Log}\ \frac{P(x)}{1-P(x)} = \varphi(x)$ où φ serait par exemple un polynôme de degré supérieur à 1.

../...

2. 2 - Les évènements sont observés lors de la "surveillance d'une cohorte"

C'est l'incidence instantanée $\lambda(t, x)$ qu'il convient de modéliser de la façon la plus simple et la plus réaliste.

Le modèle de COX stipule que le risque relatif instantané $\frac{\lambda\ (t, x)}{\lambda\ (t, x_0)}$ n'est une fonction que de la différence d'exposition x-xo (en particulier il ne dépend pas de t) : $\frac{\lambda\ (t, x)}{\lambda\ (t, x_0)} = r\ (x - xo)$ il généralise donc le cas particulier de l'exposition dichotomique avec x = o ou 1.

La modélisation la plus courante de la fonction r(x -xo) s'écrit $\exp^{a}(x - xo)$ qui indique que le risque relatif instantané pour deux niveaux d'exposition est une fonction exponentielle de la différence de ces niveaux.

Ceci entraîne que l'incidence instantané $\lambda(t, x) = \lambda(t, xo) \exp(-axo)ex$

$$\boxed{\lambda(t, x) = \varphi(t) \exp ax}$$

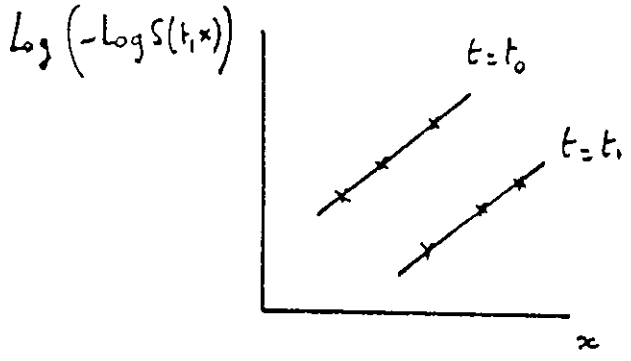
qui est la forme usuelle du modèle de COX.

Ecrivons la probabilité S(t,x) d'être encore indemne à l'instant t lorsque l'exposition est x :

$$S(t,x) = \exp - \int_0^t \lambda(u,x)\,du = \exp\left(- \exp ax \int_0^u \varphi(u)\,du\right)$$

$$S(t,x) = \left[\alpha(t)\right]^{\exp ax}$$

Cette relation implique le graphique suivant qui résume les conditions d'application du modèle de COX :



../...

Ecrivons maintenant la probabilité d'être malade à un instant T fixé que nous appellerons P(x) :

$$\text{Log } (1-P(x)) = k \exp ax$$

Si quel que soit x, la maladie est peu fréquente alors :

$$\text{Log } P(x) \# ax + b \text{ (modèle 2)}$$

et aussi bien :

$$\text{Log } \frac{P(x)}{1-P(x)} \# ax + b \text{ (modèle logistique 3)}$$

c'est à dire que lorsque la maladie est peu fréquente, le modèle de COX pour les incidences instantanées revient pratiquement à appliquer le modèle logistique pour la fréquence de la maladie à l'époque T.

2. 3 - Les évènements sont observés dans une population en "état stationnaire"

Par analogie avec le cas qualitatif, on distingue les cas de recueil "incident" ou "prévalent".

Dans le premier cas, on obtient :

$\frac{P(x)}{1-P(x)} = k \frac{f_1(x)}{f(x)} = k' \exp ax$ où $f(x)$ est la distribution de x dans la population générale, c'est à dire que le modèle logistique appliqué à cette étude de type III particulière où les témoins sont issus de la population générale fournit le même paramètre a que celui qui modèlise l'incidence instantanée $\lambda(x) = h \exp ax$ dans la population.

Dans le second cas, on obtient :

$\frac{P(x)}{1-P(x)} = k \frac{f_1(x)}{f_o(x)} = k' \bar{d}x \exp ax$ où $f_o(x)$ est cette fois la distribution de x chez les non malades (témoins)

Là encore, si $\bar{d}x$ durée moyenne de la maladie ne dépend pas de x, le modèle logistique appliqué à l'étude de type III fournit le paramètre a qui modélise l'incidence instantanée $\lambda(x) = h \exp ax$ dans la population.

CHAPITRE V

RELATION ENTRE L'EXPOSITION A PLUSIEURS FACTEURS ET LA MALADIE

QUELQUES NOTIONS SUR L'INTERACTION

Dans toute étude étiologique on est en pratique amené à étudier la relation entre la maladie et plus d'un facteur. Plusieurs situations se présentent : soit d'emblée on s'interroge simultanément sur le rôle étiologique de ces facteurs, soit un facteur étiologique est déjà bien établi et l'on souhaite étudier le rôle "propre" d'un nouveau facteur, soit même il se peut qu'un facteur non étiologique soit lié par exemple aux conditions de diagnostic de la maladie et étant également lié à l'exposition au facteur étudié, il peut introduire un biais dans l'étude. Ces questions seront développées dans le chapitre traitant des "facteurs de confusion", pour l'instant nous supposerons être dans la situation où l'on étudie simultanément deux facteurs étiologiques en relation avec la maladie.

Plaçons nous pour fixer les idées dans le cas où l'exposition aux facteurs soit dans les deux cas dichotomique et que la nature de l'étude permette de connaître la "fréquence" de la maladie que l'on appellera P_{11}, P_{10}, P_{01}, P_{00} chez les sujets exposés simultanément aux deux facteurs, à l'un d'entre eux et chez les non exposés. La généralisation aux autres cas est directe.

DEFINITION -

Il n'y a pas d'interaction entre les deux facteurs si la mesure de l'association d'un facteur avec la maladie est la même quel que soit le niveau de l'autre facteur. La conditions de non interaction s'écrit différemment selon la mesure utilisée.

Dans le modèle additif, elle s'écrit par exemple $P_{11} - P_{01} = P_{10} - P_{00}$ soit $P_{11} = P_{10} + P_{01} - P_{00}$. Si on prend le groupe non exposé à aucun des facteurs pour référence, cette condition peut s'écrire en termes de risques relatifs :

$$\boxed{R_{11} = R_{01} + R_{10} - 1} \quad (1)$$

Dans le cas de modèles multiplicatifs en choisissant le risque relatif comme mesure, on obtient la condition :

$$\frac{P_{11}}{P_{00}} = \frac{P_{10}}{P_{00}} \; \frac{P_{01}}{P_{00}} \qquad \text{soit} \qquad \boxed{R_{11} = R_{10} \; R_{01}} \quad (2)$$

../...

De même, si l'odds ratio est la mesure, la condition s'écrit :

$$\boxed{\Psi_{11} = \Psi_{10}\ \Psi_{01}} \quad ③$$

REMARQUES -

Sauf dans un cas trivial (R_{10} ou R_{01} = 1) il n'est pas possible que deux parmi les trois conditions ci-dessus soient simultanément vérifiées. En conséquence, alors que l'absence de relation entre l'exposition et la maladie pour une mesure implique qu'elle est vraie pour les autres, il n'y a absence d'interaction que pour une mesure donnée.

Néanmoins, si les fréquences P de la maladie sont faibles, les conditions ② et ③ reviennent approximativement au même.

EXEMPLE -

Dans une étude cas témoin sur l'étiologie des cancers des cavités orales, on étudie simultanément l'effet de la consommation d'alcool et de tabac (ROTHMANN K.J. et KELLER, 1972).

Un tableau simplifié des résultats est présenté ci-dessous :

		non fumeurs		fumeurs	
		cas	témoins	cas	témoins
Consommation d'alcool	nulle ou modérée	21	77	220	263
	forte	5	8	237	99

Les odds ratios $\left[\Psi = \frac{p1}{\overline{1-p1}} \Big/ \frac{p0}{\overline{1-p0}}\right]$ mesurant la relation tabac x maladie à "alcool constant" valent respectivement :

		non fumeurs	fumeurs
Consommation d'alcool	nulle ou modérée	1	3.06
	forte	1	3.83

../...

Les deux odds ratios ont des valeurs proches, d'autant plus que le deuxième est assez imprécis (faibles effectifs de non fumeurs à consommation forte d'alcool). Ces données n'impliquent donc pas une interaction ("multiplicative") entre les deux facteurs.

On aurait aussi bien pu raisonner sur la relation alcool x maladie à "tabac constant" et on aurait obtenu les odds ratios suivants et la même conclusion.

	non fumeurs	fumeurs
nulle ou modérée	1	1
forte	2.29	2.86

Cependant, à titre d'exercice, on peut rechercher s'il existe une interaction "additive" entre les deux facteurs. En effet, en assimilant les odds ratios à des risques relatifs, l'absence d'interaction "additive" devrait conduire à :

$$\Psi_{11} \neq \Psi_{10} + \Psi_{01} - 1$$

Ψ_{11} s'obtient à partir du tableau $\begin{pmatrix} 21 & 77 \\ 237 & 99 \end{pmatrix}$ $\Psi_{11} = 8.77$

Ψ_{01} et Ψ_{10} mesurant la relation "propre" d'un facteur, on a pour chacun d'eux deux estimations que l'on a trouvées proches. On verra dans un chapitre ultérieur la possibilité d'en obtenir une estimation commune.

L'estimation de MANTEL-HAENSZEL donne :

Ψ_{10} (relation alcool x maladie) = 2,83

Ψ_{01} (relation tabac x maladie) = 3,17

On retrouve l'absence (approximative) d'interaction "multiplicative" car :

$\Psi_{11} = 8.77 \neq \Psi_{10} \times \Psi_{01} = 8.97$

Par contre, il existe une forte interaction "additive" puisque :

$\Psi_{10} + \Psi_{01} - 1 = 5 < 8.77$

../...

CHAPITRE VI

ESTIMATION ET TESTS DANS LES DIVERS TYPES D'ENQUETES

On a introduit au Chapitre IV différents paramètres permettant de mesurer dans une population l'association entre l'exposition à un facteur et l'existence (ou l'occurence) d'une maladie. Ce chapitre a pour but de montrer comment on peut, à partir d'un échantillon :

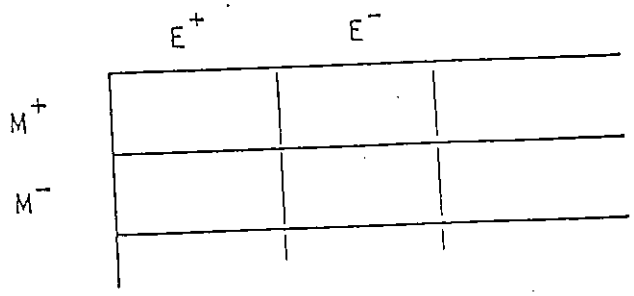
- tester l'existence d'une association
- estimer un paramètre de mesure de cette association.

On a vu que la nature des paramètres mesurables dépend du type d'étude : on peut mesurer le odds ratio (OR) dans les trois types d'études alors qu'on ne peut mesurer le risque relatif (RR) que dans les études de type échantillon représentatif ou exposés-témoins.

(Pour ce qui concerne plus particulièrement les enquêtes cas-témoins, plusieurs méthodes sont développées dans le livre de BRESLOW et DAY, Chapitre IV).

1 - NOTATIONS ET RAPPELS

1-1. Dans la population d'étude



$P_1 = P\ (M^+/E^+)$

$P_o = P\ (M^+/E^-)$

$p_1 = P\ (E^+/M^+)$

$p_o = P\ (E^+/M^-)$

.../...

- Dans les études de type échantillon représentatif ou exposés-témoins :

$$RR = P_1 / P_o \quad , \quad OR = \frac{P_1}{1-P_1} \Big/ \frac{P_o}{1-P_o}$$

- Dans les études de type cas-témoins :

$$OR = \frac{P_1}{1-P_1} \Big/ \frac{P_o}{1-P_o}$$

1-2. Dans un échantillon de taille n, les observations se répartissent suivant le tableau :

	E^+	E^-	
M^+	a	b	n_1
M^-	c	d	n_o
	m_1	m_o	n

n_1 = nombre de malades

n_o = nombre de non-malades

m_1 = nombre d'exposés

m_o = nombre de non-exposés

$n_1 + n_o = m_1 + m_o = n$

2 - TEST DE L'ASSOCIATION ENTRE L'EXPOSITION AU FACTEUR ET LA MALADIE

L'hypothèse nulle d'absence d'association peut s'exprimer de manière équivalente sous les formes suivantes :

1) $P_1 = P_o$

2) $p_1 = p_o$

3) $RR = 1$

4) $OR = 1$

et on arrive au même test quel que soit le type d'étude.

.../...

2-1. Test exact

Le "meilleur" test, c'est-à-dire non biaisé*, uniformément le plus puissant** est un test conditionnel aux marges comme on le verra au Chapitre XIII. On va donc construire le test en supposant connus les effectifs de sujets exposés et non-exposés, et de sujets malades et non-malades.

Dans un premier temps, on conditionne sur m_1 et m_o. Si on appelle A (resp. B) la variable aléatoire : nombre de malades chez les exposés (resp. les non-exposés), A et B suivent deux lois binomiales indépendantes avec les paramètres :

$$A : \mathcal{B}(m_1, P_1)$$

$$B : \mathcal{B}(m_o, P_o)$$

A et B résument toute l'information du tableau et :

$$P(A = a, B = b/m_1, m_o) = C^a_{m_1} P_1^a (1-P_1)^{m_1-a} C^b_{m_o} P_o^b (1-P_o)^{m_o-b} \quad (1)$$

Si on suppose de plus que n_1 et n_o sont connus, alors A résume toute l'information du tableau puisque

$$B = n_1 - A$$

et l'on obtient

$$P(A = a/m_1, m_o, n_1, n_o)$$

à partir de (1) en tenant compte des différentes combinaisons de a et b compatibles avec la condition :

$$a + b = n_1$$

$$P(A = a/m_1, m_o, n_1, n_o) = \frac{C^a_{m_1} P_1^a (1-P_1)^{m_1-a} C^{n_1-a}_{m_o} P_o^{n_1-a} (1-P_o)^{m_o-n_1+a}}{\sum_{x=z_1}^{z_2} C^x_{m_1} P_1^x (1-P_1)^{m_1-x} C^{n_1-x}_{m_o} P_o^{n_1-x} (1-P_o)^{m_o-n_1+x}} \quad (2)$$

où $z_1 = \max(o, n_1-m_o)$

$z_2 = \min(m_1, n_1)$

Si on introduit $\Psi = OR$, paramètre symétrique en P_1 et P_o ou p_1 et p_o (2) s'écrit :

---

* un test est non biaisé si toutes les hypothèses qui constituent l'alternative ont une plus grande probabilité d'être rejettées que H_o (LEHMANN, Ch. IV)

** un test est uniformément le plus puissant si sa puissance est maximum contre toutes les hypothèses qui constituent l'alternative (LEHMANN, Ch. III)

(3) $$P(A=a/m_1,m_o,n_1,n_o;\psi) = \frac{C_{m_1}^{a}\, C_{m_o}^{n_1-a}\, \psi^a}{\sum_{x=z_1}^{z_2} C_{m_1}^{x}\, C_{m_o}^{n_1-x}\, \psi^x}$$

(loi hypergéométrique non centrée)
et il ne reste que ψ comme paramètre.

Remarque : Si, au lieu de commencer par conditionner sur m_1 et m_o, on commence par conditionner sur n_1 et n_o, en raisonnant sur les lois des variables A et C, nombres d'exposés chez les malades et chez les témoins :

$$A : \mathcal{B}(n_1, p_1)$$
$$C : \mathcal{B}(n_o, p_o)$$

on arrive à la formule suivante :

(4) $$P(A = a/m_1,m_o,n_1,n_o ;\psi) = \frac{C_{n_1}^{a}\, C_{n_o}^{m_1-a}\, \psi^a}{\sum_{x=z_1}^{z_2} C_{n_1}^{x}\, C_{n_o}^{m_1-x}\, \psi^x}$$

qui est la même que (3).

On a donc bien le même test quel que soit le type d'étude, ce qui est logique puisque ψ est un paramètre symétrique en P_1, P_o et p_1, p_o.

Sous H_o : $\psi = 1$, (4) s'écrit

$$P(A = a/m_1,m_o,n_1,n_o;\psi=1) = \frac{C_{n_1}^{a}\, C_{n_o}^{m_1-a}}{\sum_{x} C_{n_1}^{x}\, C_{n_o}^{m_1-x}}$$

(5) $$= \frac{C_{n_1}^{a}\, C_{n_o}^{m_1-a}}{C_{n_1+n_o}^{m_1}}$$

c'est-à-dire une loi hypergéométrique.

Si on note :

$$P_1 = P(A \leq a/m_1,m_o,n_1,n_o;\psi=1)$$

(6) $$P_1 = \sum_{x=\sup(o,n_1-m_o)}^{a} \frac{C_{n_1}^{x}\, C_{n_o}^{m_1-x}}{C_{n}^{m_1}}$$

.../...

De même :

$$P_S = P(A \geq a / m_1, m_o, n_1, n_o ; \Psi = 1)$$

(7)
$$P_S = \sum_{x=a}^{\inf(n_1, m_1)} \frac{C_{n_1}^{x} \; C_{n_o}^{m_1 - x}}{C_n^{m_1}}$$

C'est le test exact de Fisher.

. Test unilatéral : $H_o : \Psi = 1$ contre $H_1 : \Psi > 1$

on rejette H_o au risque α si $P_S < \alpha$

. Test unilatéral : $H_o : \Psi = 1$ contre $H_1 : \Psi < 1$

on rejette H_o au risque α si $P_I < \alpha$

. Test bilatéral : $H_o : \Psi = 1$ contre $H_1 : \Psi \neq 1$

une solution approchée est obtenue en rejettant H_o au risque α si $2 \inf (P_I, P_S) < \alpha$

Exemple : Diethylstilbestrol (DES) pendant la grossesse et adenocarcinome du vagin chez les filles (HERBST)

	DES Oui	DES Non	
Cas	7	1	8
Témoin	0	32	32
	7	33	40

Test unilatéral $H_o : \Psi = 1$ contre $H_1 \; \Psi > 1$

Ici : $a = 7$ et $\inf. (n_1, m_1) = 7$

P_s se réduit à :

$$P_s = \frac{C_8^7 \; C_{32}^0}{C_{40}^7}$$

$$= \frac{8!}{7! \times 1!} \times \Big/ \frac{40!}{7! \; 33!}$$

$$= \frac{8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2}{34 \times 35 \times 36 \times 37 \times 38 \times 39 \times 40} = 4{,}3 \; 10^{-7}$$

Il y a eu significativement plus ($p < 10^{-6}$) d'exposition au DES chez les mères des jeunes filles ayant un adenocarcinome du vagin que chez les mères de témoins.

...

2-2. Test approché

Le test exact n'est faisable pratiquement que pour des petits échantillons. Dans le cas général, on suppose que A, conditionnellement aux marges m_1, m_o, n_1, n_o suit une loi normale de moyenne $E(A ; \Psi)$ et de variance $V(A ; \Psi)$. Sous l'hypothèse $H_o : \Psi = 1$, on a les valeurs suivantes de la moyenne et de la variance :

(S) $$E(A ; \Psi = 1) = \frac{m_1 n_1}{n}$$

$$V(A ; \Psi = 1) = \frac{m_1 n_1 m_o n_o}{n^2 (n-1)}$$

(9) $$\sim \frac{m_1 n_1 m_o n_o}{n^3}$$

Ces dernières formules sont celles de la moyenne et de la variance d'une loi hypergéométrique.

On peut alors calculer P_I et P_S pour une loi normale dont les paramètres sont estimés par (8) et (9). Cela revient à calculer la quantité :

(10) $$\frac{(a - E(A ; \Psi = 1))^2}{V(A ; \Psi = 1)}$$

qui suit une loi du χ^2 à 1 d.d.l.

En fait (10) s'écrit à partir des formules (8) et (9)

(11) $$\frac{(ad - bc)^2 \times n}{m_1 m_o n_1 n_o}$$

qui est le test du χ^2 habituel pour un tableau 2 x 2, avec les conditions habituelles sur les effectifs théoriques (conditions pour que A suive une loi normale).

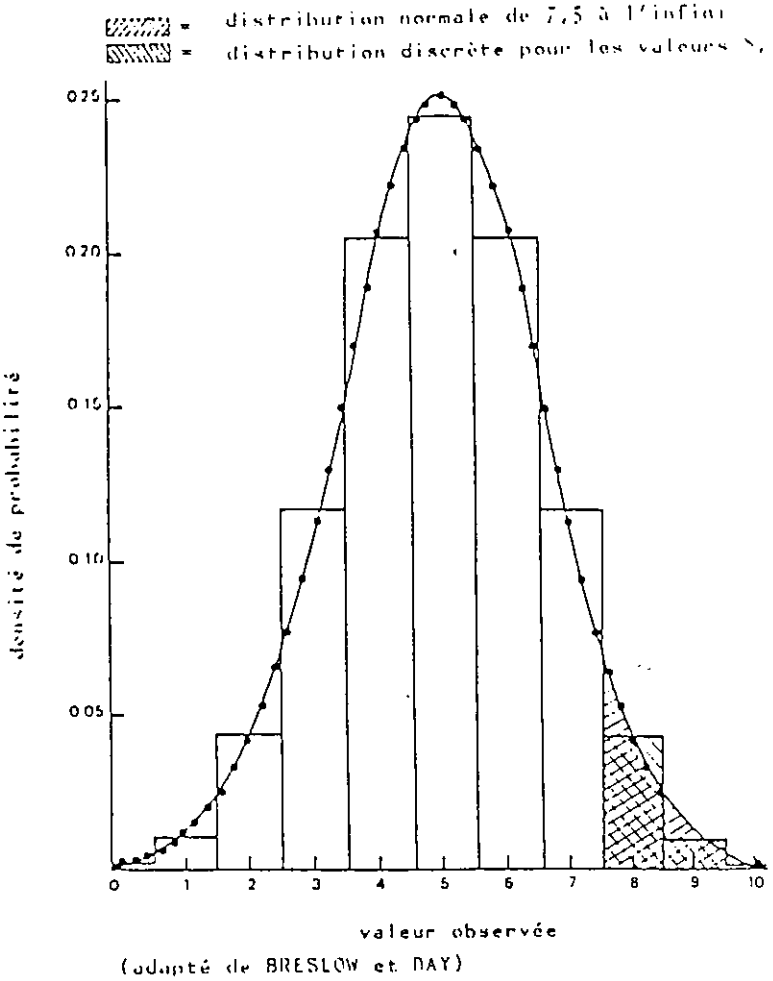
Remarque : au lieu de la formule (11), on trouve parfois une formule du χ^2 corrigé, ou χ^2 de Yates.

(12) $$\frac{(|ad - bc| - \frac{n}{2})^2 \times n}{m_1 m_o n_1 n_o}$$

.../...

On introduit une correction de continuité $(\frac{n}{2})$ pour approcher une distribution discrète par une distribution continue

Figure : Approximation normale d'une distribution de probabilité discrète.



(adapté de BRESLOW et DAY)

Exemples

1. Association entre consommation d'alcool pendant la grossesse et hypotrophie de l'enfant : enquête prospective. Paris 1962-69 (données de KAMINSKI et al.)

- Consommation totale d'alcool estimée en cl de vin par jour en 2 classes selon que inférieure ou égale à 40 cl par jour, ou supérieure à ce seuil.
- Hypotrophie : poids de naissance inférieur au 3ème percentile compte tenu du terme

Enfant hypotrophique	Consommation d'alcool >40 cl	≤40 cl	Total
Oui	20	189	209
Non	392	7.372	7.764
Total	412	7.561	7.973
P	4,9 %	2,5 %	

Test d'association consommation d'alcool x hypotrophie

$$\chi^2 = \frac{(20 \times 7372 - 392 \times 189)^2 \times 7973}{412 \times 7561 \times 209 \times 7764}$$

$$= 8,49$$

χ^2 à 1 d.d.l : test significatif à $p < 0,01$

2. Etude cas témoins de l'association entre mononucléose infectieuse et antécédent d'amygdalectomie chez des étudiants de 18 à 24 ans (données de MILLER)

	Amygdalectomie Oui	Non	Total	p
Cas	40	145	185	22 %
Témoins	235	420	655	36 %
Total	275	565	840	

Test d'association

$$\chi^2 = \frac{(40 \times 420 - 145 \times 235)^2 \times 840}{275 \times 565 \times 185 \times 655} = 13,3$$

χ^2 à 1 ddl : test significatif à $p < 0,001$

.../...

3 - ESTIMATION DU RISQUE RELATIF

Dans les études de type échantillon représentatif, ou exposés-témoins, on peut estimer le risque relatif comme mesure d'association (voir Chapitre IV).

3-1. Estimation ponctuelle du risque relatif

Les probabilités de maladie chez les sujets exposés et non-exposés sont estimées (estimateurs du maximum de vraisemblance) par :

$$\hat{P}_1 = {}^{a}/_{m_1}$$

$$\hat{P}_o = {}^{b}/_{m_o}$$

et le risque relatif par :

$$\widehat{RR} = \hat{P}_1/\hat{P}_o = \frac{a}{m_1} \Big/ \frac{b}{m_o} \qquad (13)$$

3-2. Intervalle de confiance pour le risque relatif (KATZ et al.)

On suppose que Log $\widehat{RR}$ suit une loi normale de moyenne estimée par Log $\left(\frac{a}{m_1} / \frac{b}{m_o}\right)$. Log $\widehat{RR}$ converge plus rapidement vers une loi normale que $\widehat{RR}$.

La transformation logarithmique a en outre l'avantage de faire jouer un rôle symétrique à P_1 et P_o.

3-2-1. Estimation de la variance de Log RR

$$\text{Log } \widehat{RR} = \text{Log } \hat{P}_1 - \text{Log } \hat{P}_o$$

$$\text{var Log } \widehat{RR} = \text{var Log } \hat{P}_1 + \text{var Log } \hat{P}_o$$

$$\text{var Log } \hat{P}_1 = \text{var } \hat{P}_1 \cdot (\text{Log } P)'^{2}_{P_1} = \frac{P_1(1-P_1)}{m_1} \times \frac{1}{P_1^{\,2}}$$

$$= \frac{1 - P_1}{m_1 P_1}$$

que l'on estime par :

$$\frac{c}{m_1 a}$$

.../...

D'où, var Log $\widehat{RR}$ est estimée par

(14) $$\frac{c}{m_1\ a} + \frac{d}{m_o\ b}$$

3-2-2. Intervalle de confiance pour RR

Intervalle de confiance au risque α pour Log RR :

(15) $$\text{Log}\left(\frac{a}{m_1} \Big/ \frac{b}{m_o}\right) \pm \mathcal{E}_\alpha \sqrt{\frac{c}{m_1\ a} + \frac{d}{m_o\ b}}$$

Intervalle de confiance au risque α pour RR : en prenant l'exponentielle des bornes de l'intervalle de confiance pour Log RR.

3-2-3. Méthode de Miettinen pour estimer l'intervalle de confiance de RR (MIETTINEN 1974)

Méthode utilisant la valeur du χ^2 du test d'association

(16) $$(R\widehat{R}_1,\ R\widehat{R}_S) : \quad R\widehat{R}\left(1 \pm \mathcal{E}_\alpha \Big/ \sqrt{\chi^2}\right)$$

Miettinen a proposé cette méthode en supposant que le χ^2 habituel du tableau 2 x 2 et la quantité $\frac{\text{Log}^2\ \widehat{RR}}{\text{var Log } \widehat{RR}}$ sont deux manières équivalentes de tester l'hypothèse nulle : RR = 1 et que par conséquent l'égalité :

$$\frac{\text{Log}^2\ \widehat{RR}}{\text{var Log } \widehat{RR}} = \chi^2$$

est à peu près vérifiée.

En résolvant cette égalité pour en tirer une estimation de la variance de Log $\widehat{RR}$, on obtient :

$$\text{var Log } \widehat{RR} = \frac{\text{Log}^2\ \widehat{RR}}{\chi^2}$$

d'où la formule (16)

Cette méthode approximative ne donne de résultats valables que si RR est voisin de 1.

Remarque : si on peut calculer un intervalle de confiance pour RR, on peut en déduire un intervalle de confiance pour la fraction étiologique FE, puisque $FE = 1 - \frac{1}{RR}$

.../...

Exemple :

Consommation d'alcool x hypotrophie (exemple 1 du paragraphe 2)

a) $\hat{RR} = {}^{20}/_{412} \Big/ {}^{189}/_{7561} = 1,94$

b) Log $\hat{RR} = 0,66$

$$\text{var Log } \hat{RR} = \frac{c}{m_1\ a} + \frac{d}{m_o\ b} = \frac{392}{412 \times 20} + \frac{7372}{7561 \times 189}$$

$$= 0,053$$

Intervalle de confiance pour Log RR :

$$\text{Log RR} : 0,66 \pm 2\sqrt{0,053}$$

$$: (0,20 \ ; \ 1,12)$$

en prenant l'exponentielle, intervalle de confiance pour RR

$$RR : (1,22 \ ; \ 3,06)$$

c) méthode de Miettinen : intervalle de confiance pour RR :

$$1,94^{\left(1 \pm \frac{2}{\sqrt{8,49}}\right)} = 1,94^{(1 \pm 0,686)}$$

$$: (1,23 \ ; \ 3,06)$$

4 - ESTIMATION DU ODDS-RATIO

4-1. Estimation ponctuelle du odds-ratio

Si l'on revient au modèle exact introduit en 2-1, l'estimateur du maximum de vraisemblance de Ψ, conditionnellement aux marges, est la quantité $\hat{\Psi}$ mv cond. qui maximise la vraisemblance conditionnelle des observations, c'est-à-dire la formule (4). On peut montrer que $\hat{\Psi}$ mv cond. est solution de l'équation

(18) $$a = E\ (A/m_1,\ m_o,\ n_1,\ n_o\ ;\ \Psi)$$

C'est-à-dire que $\hat{\Psi}$ mv est la valeur du paramètre Ψ pour laquelle l'espérance conditionnelle de A est égale à la valeur observée pour le nombre de malades exposés.

.../...

Cette équation polynomiale de degré élevé est compliquée à résoudre quand n est grand. Une solution approchée de (18) est

(19) $$\hat{\Psi} = \frac{ad}{bc}$$

Remarque : on peut aussi raisonner comme pour le risque relatif, en considérant que :

$$\Psi = \frac{P_1}{1-P_1} \Big/ \frac{P_o}{1-P_o} \quad \left(\text{ou } \frac{p_1}{1-p_1} \Big/ \frac{p_o}{1-p_o}\right)$$

et estimer P_1 et P_o (ou p_1 et p_o) à partir des observations selon le type d'étude ; par exemple pour des études de type échantillon représentatif ou cas-témoins, les estimations du maximum de vraisemblance de p_1 et p_o sont :

$$\hat{p}_1 = \frac{a}{n_1}$$

$$\hat{p}_o = \frac{c}{n_o}$$

et on retrouve

$$\hat{\Psi} = \frac{a/n_1}{b/n_1} \Big/ \frac{c/n_o}{d/n_o} = \frac{ad}{bc}$$

Si un des effectifs a, b, c ou d est nul, la formule (19) n'est pas utilisable, et il faut résoudre l'équation (18).

4-2. Intervalle de confiance pour le odds-ratio

4-2-1. Modèle exact

L'intervalle de confiance pour Ψ au risque α est compris entre les bornes $(\hat{\Psi}_I, \hat{\Psi}_S)$ telles que

(20) $$\begin{cases} P\ (A \leq a/m_1,\ m_o,\ n_1,\ n_o\ ;\hat{\Psi}_I) = \frac{\alpha}{2} \\ P\ (A \geq a/m_1,\ m_o,\ n_1,\ n_o\ ;\hat{\Psi}_S) = \frac{\alpha}{2} \end{cases}$$

On peut théoriquement calculer ces probabilités à partir de (4) mais comme pour l'estimateur $\hat{\Psi}$ mv cond., on est amené à résoudre des équations polynomiales de degré élevé.

.../...

4-2-2. Limites de Cornfield (CORNFIELD)

Si on fait, comme en 2 - 2, l'hypothèse que A suit, conditionnellement aux marges, une loi normale de moyenne $E(A;\Psi)$ et variance $V(A;\Psi)$, (20) se ramène, avec une correction de continuité, à :

$$(21)\quad \begin{cases} a - E(A;\hat{\Psi}_1) - \frac{1}{2} = \xi_\alpha \sqrt{\text{var}(A;\hat{\Psi}_1)} \\ a - E(A;\hat{\Psi}_S) + \frac{1}{2} = -\xi_\alpha \sqrt{\text{var}(A;\hat{\Psi}_S)} \end{cases}$$

$E(A;\Psi)$ est solution de l'équation :

$$(22)\quad \frac{E(A;\Psi) \times \left[n_o - m_1 + E(A;\Psi)\right]}{\left[n_1 - E(A;\Psi)\right]\left[m_1 - E(A;\Psi)\right]} = \Psi$$

et

$$(23)\quad V(A;\Psi) = \left[\frac{1}{E(A;\Psi)} + \frac{1}{n_1 - E(A;\Psi)} + \frac{1}{m_1 - E(A;\Psi)} + \frac{1}{n_o - m_1 + E(A;\Psi)}\right]^{-1}$$

Remarque : pour $\Psi = 1$, on retrouve (8) et (9).

On ne peut résoudre (21) que par des méthodes iteratives ; mais les limites de Cornfield constituent la meilleure approximation de l'intervalle de confiance si l'approximation normale de la loi de A est justifiée.

4-2-3. Méthode des logits (WOOLF)

Définition : $\text{Logit } p = \text{Log } \frac{p}{1-p}$

alors $\log \Psi = \text{Logit } p_1 - \text{Logit } p_o$

Cette transformation logarithmique permet d'avoir une distribution plus proche de la loi normale et à l'avantage de faire jouer un rôle symétrique à p_1 et p_o (ou P_1 et P_o).

On suppose donc que $\log \hat{\Psi}$ suit une loi normale dont la moyenne peut être estimée par $\text{Log } \frac{ad}{bc}$

.../...

Estimation de la variance de Log $\hat{\Psi}$

$$\text{Log } \hat{\Psi} = \text{Logit } \hat{p}_1 - \text{Logit } \hat{p}_o$$

$$\text{Var Logit } \hat{p}_1 = \text{var Log } \frac{\hat{p}_1}{1 - \hat{p}_1}$$

$$= \text{Var}(\hat{p}_1) \times \left(\text{Log } \frac{p}{1-p}\right)'^2_{p_1}$$

$$= \frac{p_1(1-p_1)}{n_1} \times \left(\frac{1}{p_1(1-p_1)}\right)^2$$

$$= \frac{1}{n_1 p_1 (1-p_1)} = \frac{1}{n_1 p_1} + \frac{1}{n_1(1-p_1)}$$

qu'on peut estimer par $\frac{1}{a} + \frac{1}{b}$

On peut donc estimer Var Log $\hat{\Psi}$ par

(24) $$\text{Var Log } \hat{\Psi} = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

D'où l'intervalle de confiance au risque α pour Log Ψ

(25) $$(\text{Log } \hat{\Psi}_I \ , \ \text{Log } \hat{\Psi}_S) : \text{Log } \frac{ad}{bc} \pm \varepsilon_\alpha \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

L'intervalle de confiance pour Ψ est obtenu par transformation exponentielle des bornes ainsi calculées.

Cette méthode a l'inconvénient de fournir un intervalle de confiance trop étroit si les effectifs sont faibles.

4-2-4. Méthode de Miettinen (MIETTINEN 1976)

Méthode utilisant la valeur du χ^2 du test d'association (voir explication au paragraphe 3-2-3.)

(26) $$(\hat{\Psi}_I, \hat{\Psi}_S) : \frac{ad}{bc}^{\left(1 \pm \varepsilon_\alpha \ / \sqrt{\chi^2}\right)}$$

Cette méthode approximative ne donne des résultats valables que si Ψ est voisin de 1. Sinon, elle fournit selon les cas des intervalles trop larges ou des intervalles trop étroits.

.../...

Des études ont comparé les différentes méthodes d'estimation par intervalle du odds-ratio : elles ont montré que c'est la méthode de Cornfield qui fournit les meilleurs résultats, et ont décrit les erreurs dues aux deux autres méthodes (GART et THOMAS). En pratique, toutefois, ce sont ces deux méthodes qui sont les plus utilisées.

Exemple : Mononucléose infectieuse et antécédent d'amygdalectomie (exemple 2 du paragraphe 2)

1) estimation du odds-ratio :

$$\hat{\Psi} = \frac{40 \times 420}{145 \times 235} = 0,49$$

2) intervalle de confiance pour $\text{Log}\,\Psi$

a) méthode des logits

$$(\text{Log}\,\hat{\Psi}_I, \text{Log}\,\hat{\Psi}_S) : \text{Log}\,\hat{\Psi} \pm 2\sqrt{\text{var Log}\,\hat{\Psi}}$$

$$\text{Log}\,\hat{\Psi} = -0,713$$

$$\text{Var Log}\,\hat{\Psi} = \frac{1}{40} + \frac{1}{420} + \frac{1}{145} + \frac{1}{235} = 0,0385$$

$$(\text{Log}\,\hat{\Psi}_I, \text{Log}\,\hat{\Psi}_S) : -0,713 \pm 2\sqrt{0,0385}$$

$$(-1,106 ; -0,320)$$

D'où pour $\hat{\Psi}$:

$$(\hat{\Psi}_I ; \hat{\Psi}_S) = (0,33 ; 0,73)$$

b) méthode de Miettinen

$$(\hat{\Psi}_I, \hat{\Psi}_S) = \hat{\Psi}\left(1 \pm \varepsilon_\alpha / \sqrt{X^2}\right)$$

or $X^2 = 13,3$

$$(\hat{\Psi}_I, \hat{\Psi}_S) = 0,49\left(1 \pm 2/\sqrt{13,3}\right)$$

$$= (0,49^{(1,548)}, 0,49^{(0,452)})$$

$$(0,33 ; 0,72)$$

.../...

5 - FACTEUR QUALITATIF A PLUSIEURS NIVEAUX D'EXPOSITION

5-1. Notations

	Niveau d'exposition							
	0	1	2	...	i	...	k	
M^+	a_o	a_1	a_2		a_i		a_k	n_1
M^-	c_o	c_1	c_2		c_i		c_k	n_o
	m_o	m_1	m_2		m_i		m_k	n

$$P_i = P(M^{\pm}/E_i) \qquad i = 0,1 \ldots k$$

$$P_{1i} = P(E_i/M^{+}) \qquad i = 0,1 \ldots k$$

$$P_{oi} = P(E_i/M^{-}) \qquad i = 0,1 \ldots k$$

On suppose qu'il y a k + 1 niveaux d'exposition et que le niveau 0 est celui choisi comme référence.

- Pour une étude de type échantillon représentatif ou exposés-témoins

$$RR_i = \frac{P_i}{P_o} \qquad \Psi_i = OR_i = \frac{P_i}{1-P_i} \Big/ \frac{P_o}{1-P_o}$$

- Pour une étude de type échantillon représentatif ou cas-témoins

$$\Psi_i = OR_i = \frac{P_{1i}}{1-P_{1i}} \Big/ \frac{P_{oi}}{1-P_{oi}}$$

5-2. Test de l'association entre l'exposition au facteur et la maladie

Hypothèse nulle d'absence d'association :

1) Tous les P_i sont égaux
2) Les distributions de p_{1i} et p_{oi} sont identiques
3) Tous les RR_i sont égaux à 1
4) Tous les OR_i sont égaux à 1

.../...

5-2-1. Test du χ^2

Si l'hypothèse alternative est qu'il existe une association entre facteur et maladie, sans précision sur la forme de cette association, le test utilisé est le test du χ^2 sur un tableau $2 \times (k + 1)$:

Si e_i est le nombre attendu de malades au niveau d'exposition i :

$$e_i = \frac{n_1 \, m_i}{n}$$

$$n \left(\frac{1}{n_1} + \frac{1}{n_o}\right) \sum_{i=0}^{k} \frac{(a_i - e_i)^2}{m_i} \qquad (27)$$

suit une loi du χ^2 à k ddl.

5-2-2. Test de tendance du risque selon le niveau d'exposition ("effet-dose") (ARMITAGE)

Si l'hypothèse alternative à l'absence d'association est l'existence d'un effet d'autant plus important que le niveau d'exposition est plus élevé, plus exactement que la fréquence de la maladie P_i croît linéairement avec le niveau d'exposition, le test de tendance d'Armitage est plus puissant que le test du χ^2 habituel.

Si x_i est le niveau d'exposition dans la classe i (x_i peut être simplement égal à i si les classes sont ordonnées), on étudie la régression de $(a_i - e_i)$ sur x_i :

$$\frac{n^3 \left\{ \sum_{i=0}^{k} x_i \, (a_i - e_i) \right\}^2}{n_1 \, n_o \left\{ n \sum_{o}^{k} x_i^2 \, m_i - \left(\sum_{o}^{k} x_i \, m_i \right)^2 \right\}}$$

suit une loi du χ^2 à 1 ddl.

Ce test n'a pas d'interprétation dans les études cas-témoins.

5-3. Estimation du risque relatif

pour chaque niveau comparé au niveau 0 :

$$\widehat{RR}_i = \frac{a_i}{m_i} \Big/ \frac{a_o}{m_o}$$

dont la variance est estimée comme dans le cas du facteur à 2 classes (voir 3.2)

.../...

5-4. Estimation du odds-ratio

Pour chaque niveau comparé au niveau 0 :

$$\hat{\Psi}_i = \frac{a_i \, c_o}{c_i \, a_o}$$

Pour l'intervalle de confiance, voir 4.2.

Exemple : consommation d'alcool pendant la grossesse et hypotrophie = étude de l'effet de la quantité d'alcool.

Consommation d'alcool en cl/vin/jour

Enfant hypotrophique	0	1-20	21-40	41-59	60 +	Total
Oui	95	63	32	13	7	210
Non	3853	2465	1136	315	81	7850
Total	3948	2528	1168	328	88	8060

P	2,4%	2,5%	2,7%	4,0%	8,0%
$RR/_0$	1	1,04	1,14	1,65	3,31
e_i	102,9	65,9	30,4	8,5	2,3
x_i	1	2	3	4	5

1) Test du χ^2

$$\chi^2 = 8060 \left(\frac{1}{210} + \frac{1}{7850}\right) \left[\frac{(102,9 - 95)^2}{3948} + \frac{(65,9 - 63)^2}{2528} + \frac{(32 - 30,4)^2}{1168} + \frac{(13 - 8,5)^2}{328} + \frac{(7 - 2,3)^2}{88}\right]$$

$$= 13,17 \qquad p < 0,05 \quad (4 \text{ ddl})$$

.../...

2) Test de tendance linéaire

$$\chi^2 = \frac{8060^3 \left[(95-102,9)+2(63-65,9)+3(32-30,4)+4(13-8,5)+5(7-2,3)\right]^2}{7850\times210\times\left[8060(3948+4\times2528+9\times1168+16\times328+25\times88) - (3948+2\times2528+3\times1168+4\times328+5\times88)^2\right]}$$

$$= \frac{8060^3 \times 32,6^2}{7850 \times 210 \times \left[8060 \times 32020 - 14260^2\right]}$$

$$= 6,17 \qquad p < 0,05 \qquad (1 \text{ ddl})$$

L'hypothèse d'une augmentation linéaire des taux d'hypotrophie avec la consommation d'alcool est donc compatible avec les données.

CHAPITRE V - REFERENCES

METHODOLOGIE :

- ARMITAGE P.
  Statistical methods in medical research
  Blackwell, Oxford, 1977, 4è édition, Chapitre 12

- BRESLOW N.E., DAY N.E.
  Statistical methods in Cancer Research
  Volume 1. The analysis of case-control studies
  IARC Scienfific Publications n° 32
  IARC, Lyon, 1980. Chapitre 4.

- CORNFIELD J.
  A statistical problem arising from retrospective studies
  In : NEYMAN J ed. Proceedings of the 3rd Berkeley Symposium IV,
  University of California Press Berkeley, 1956.

- GART J.J., THOMAS D.G.
  The performance of three approximate confidence limit methods of the odds ratio
  Am. J. Epidemiol., 1982, 115, 453-470.

- KATZ D., BAPTISTA J., AZEN S.P., PIKE M.C.
  Obtaining confidence intervals for the risk ratio in cohort studies
  Biometrics, 1978, 34, 469-474.

- LEHMANN E.L.
  Testing statistical hypothesis
  Wiley, New York, 1970, 5è édition, Chapitres 3 et 4.

- MIETTINEN O.
  Simple interval estimation of risk ratio
  Am. J. Epidemiol. 1974, 100, 515-516.

- MIETTINEN O.
  Estimability and estimation in case referrent studies
  Am. J. Epidemiol. 1976, 103, 226-235.

- WOOLF B.
  On estimating the relationship between blood group and disease
  Am. Human Gen. 1955, 19, 251-253.

EXEMPLES ADAPTES DE :

- HERBST A.L., ULFELDER H., POSKANCER D.C.
  Adenocarcinoma of the vagina : association of maternal stilbestrol therapy with tumor appearance in young women.
  N. Engl. J. Med, 1971, 284, 878-881.

- KAMINSKI M., RUMEAU-ROUQUETTE C., SCHWARTZ D.
  Consommation d'alcool chez les femmes enceintes et issue de la grossesse
  Rev. Epidem. et Santé Publ. 1976, 24, 27-40

- MILLER R.G.
  Combining 2 x 2 contingency tables.
  in : Biostatistics casebook, MILLER RG, EFRON B., BROWN BW, MOSES LE.
  Wiley, New York, 1980, pp 73-83.
  (exemple sur la mononucléose infectieuse).

CHAPITRE VII

" FACTEURS DE CONFUSION "

Supposons qu'une étude vise à rechercher si l'exposition à un toxique professionnel augmente la fréquence de troubles de la glycorégulation. Dans cette étude, des sujets exposés et non exposés, par exemple d'une grande entreprise, sont comparés et aucune différence nette n'apparait entre ces deux groupes quant à la fréquence de l'hyperglycémie. On peut imaginer que les sujets exposés sont vraisemblablement plus jeunes que les non exposés et comme la fréquence de l'hyperglycémie augmente avec l'âge, l'effet du toxique peut être masqué. Si tel est le cas, nous dirons que l'âge est un facteur de confusion pour l'étude du toxique. La réflexion sur l'existence de facteurs de confusion dans une étude étiologique est essentielle car elle met en cause la possibilité de biais dans la relation entre la maladie et le facteur d'intérêt. Nous verrons également que la puissance de l'étude pour mettre en évidence cette relation peut également en dépendre.

Prenons un autre exemple : le tabagisme est un facteur très important de l'étiologie de la bronchite chronique; dans une étude concernant tout autre facteur éventuel (pollution par exemple), le facteur tabac peut jouer potentiellement un rôle de facteur de confusion car il suffirait pour cela, ainsi que nous le verrons, qu'il soit également lié à l'exposition. Il est donc nécessaire que la notion de tabagisme soit connue pour les sujets de l'étude (exposés/non exposés ou cas/témoins) et peut être même il en sera tenu compte dès la planification de l'étude.

Dans les exemples ci-dessus, les relations âge x maladie et tabagisme x maladie sont considérées comme "causales" et connues a priori. Plus difficiles à détecter sont les facteurs de confusion associés non pas à la maladie elle-même ("biologiquement") mais aux conditions de son diagnostic dans l'étude particulière entreprise : par exemple lorsqu'on compare des malades hospitalisés à des témoins de la population générale, la répartition des classes sociales est vraisemblablement différente dans les deux groupes et se révèlerait ainsi un facteur de confusion pour l'étude d'un facteur qui lui serait associé.

../...

Finalement le problème posé peut être défini de la façon suivante : alors que l'on s'intéresse au rôle étiologique d'un facteur E, on cherche à connaître la répercussion que peut avoir le fait que E soit lié à un facteur C lui-même associé "causalement" (biologiquement ou non) à la maladie, sur la relation entre E et la maladie. Soit φ_p la mesure de la relation entre le facteur E et la maladie observée dans l'étude. Soit φ la mesure de cette relation pour les sujets ayant le même niveau du facteur C. Nous supposons donc que φ reste identique quel que soit le niveau du facteur C : cette condition est celle d'une absence d'interaction entre E et C pour la mesure φ. Ceci montre bien que l'on ne s'intéresse pas simultanément aux deux facteurs E et C et donc à leur interaction mais que l'on cherche seulement à éliminer le biais que C pourrait introduire dans l'étude de E. C'est seulement dans le cas où il n'y a pas interaction entre les facteurs que des résultats généraux peuvent être obtenus. Soit φ' la mesure de la relation entre le facteur C et la maladie pour les sujets ayant le même niveau du facteur E. L'hypothèse d'absence d'interaction entre E et C pour la mesure φ indique que φ' ne dépend pas du niveau de E. Nous dirons que C est facteur de confusion pour E si $\varphi_r \neq \varphi$. On imagine aisément que l'effet de confusion puisse être positif $\varphi_r > \varphi$ ou négatif $\varphi_r < \varphi$. Nous distinguerons plusieurs cas selon la nature qualitative ou non de l'exposition aux facteurs E et C et selon la nature de la mesure .

1. L'EXPOSITION AUX FACTEURS E ET C EST QUALITATIVE A DEUX CLASSES

Les notations suivantes sont utilisées et sont valables quel que soit le type de l'étude :

	C^+		C^-	
	E^+	E^-	E^+	E^-
M^+	a_1	b_1	a_2	b_2
M^-	c_1	d_1	c_2	d_2
	n_1	N_1	n_2	N_2

où a_1 ..... sont les effectifs observés.

../...

1. 1 - Modèle additif

On peut écrire :

$$\Delta_p = \frac{a_1 + a_2}{n_1 + n_2} - \frac{b_1 + b_2}{N_1 + N_2}$$

$$\Delta = \frac{a_1}{n_1} - \frac{b_1}{N_1} = \frac{a_2}{n_2} - \frac{b_2}{N_2}$$

$$\Delta' \quad \frac{a_1}{n_1} - \frac{a_2}{n_2} = \frac{b_1}{N_1} - \frac{b_2}{N_2}$$

Un calcul direct montre que :

$$\boxed{\Delta_p = \Delta + \Delta' (p_1 - p_2)}$$

où $p_1 = \frac{n_1}{n_1 + n_2}$ = Prob (C^+/E^+)

et $p_2 = \frac{N_1}{N_1 + N_2}$ = Prob (C^+/E^-)

On en déduit que pour que C soit de confusion, il faut qu'à la fois :

. $\Delta' \neq 0$ c'est à dire que C soit lié à la maladie par classe de E.

. $p_1 \triangleq p_2$ c'est à dire que C soit lié à E dans l'ensemble de la population (comprenant à la fois les malades et non malades).

1. 2 - Modèle multiplicatif (risques relatifs)

On peut écrire :

$$R_p = \frac{a_1 + a_2}{n_1 + n_2} \cdot \frac{N_1 + N_2}{b_1 + b_2}$$

$$R = \frac{a_1 \; N_1}{b_1 \; n_1} = \frac{a_2 \; N_2}{b_2 \; n_2}$$

$$R' = \frac{a_1 \; n_2}{a_2 \; n_1} = \frac{b_1 \; N_2}{b_2 \; N_1}$$

../...

On en déduit :

$$a_1 + a_2 = R \left(\frac{a_1 b_1}{N_1} + \frac{n_2 b_2}{N_2} \right)$$

$$b_2 = \frac{N_2 b_1}{b_2 R'}$$

d'où :

$$Rp = R \frac{N_1 + N_2}{n_1 + n_2} \frac{\frac{n_1 b_1}{N_1} + \frac{n_2 b_1}{N_1 R'}}{b_1 + \frac{N_2 b_1}{N_1 R'}} = R \frac{N_1 + N_2}{n_1 + n_2} \frac{n_1 R' + n_2}{N_1 R' + N_2}$$

soit :

$$R_p = R \frac{\frac{n_1}{n_1 + n_2} R' + \frac{n_2}{n_1 + n_2}}{\frac{N_1}{N_1 + N_2} R' + \frac{N_2}{N_1 + N_2}}$$

Avec les notations précédentes, on obtient :

$$\boxed{R_p = R \frac{p_1 R' + 1 - p_1}{p_2 R' + 1 - p_2}}$$

Comme dans le modèle additif pour que C soit facteur de confusion, il faut qu'à la fois :

. $R' \neq 1$ c'est à dire que C soit lié à la maladie par classe de E

. $p_1 \neq p_2$ c'est à dire que C soit lié à E dans l'ensemble de la population.

La quantité $\frac{R_p}{R}$, qui exprime le degré de confusion entraîné par C pour le risque lié à E, prend le nom de "rapport de confusion" (confounding risk ratio).

L'effet de confusion est positif si $R'>1$ et $p_1>p_2$ ou $R'<1$ et $p_1<p_2$ et négatif dans les autres cas.

Exemple :

On s'intéresse à la relation entre classe sociale (I ou II) et l'infarctus du myocarde dans une étude prospective. Les sujets appartiennent soit

à la classe d'âge 40-49 ans, soit 50-59 ans. Dans chaque classe sociale, le risque relatif de la 2ème classe d'âge est 2,1 par rapport à la première. Chez les sujets les plus jeunes, 30 % appartiennent à la classe II et 60 % chez les plus âgés.

On observe dans l'ensemble de la population un risque relatif de 1 entre les classes sociales.

Le rapport de confusion entraîné par l'âge sur la relation entre classes sociales et maladie s'écrit :

$$\frac{2,1 \times 0,6 + 0,4}{2,1 \times 0,3 + 0,7} = 1,3$$

Après élimination de l'effet de l'âge, le risque relatif entre classes sociales n'est plus que :

$$\frac{1,1}{1,3} = 0,85$$

1. 3 - Modèle multiplicatif (odds ratio)

On écrit :

$$\Psi_p = \frac{(a_1 + a_2)(d_1 + d_2)}{(b_1 + b_2)(c_1 + c_2)}$$

$$\Psi = \frac{a_1\, d_1}{b_1\, c_1} = \frac{a_2\, d_2}{b_2\, c_2}$$

$$\Psi' = \frac{a_1\, c_2}{a_2\, c_1} = \frac{b_1\, d_2}{b_2\, d_1}$$

On en déduit :

$$d_1 + d_2 = \Psi\left(\frac{b_1\, c_1}{a_1} + \frac{b_2\, c_2}{a_2}\right) = \Psi\left(\frac{b_1\, c_1}{a_1} + \frac{b_2\, a_2\, c_1\, \Psi'}{a_2\, a_1}\right)$$

$$= \Psi\ \frac{c_1}{a_1}\left(b_1 + b_2\, \Psi'\right)$$

$$\text{et } c_1 + c_2 = c_1 + \frac{c_1\, a_2}{a_1}\Psi' = \frac{c_1}{a_1}\left(a_1 + a_2\, \Psi'\right)$$

../...

d'où :

$$\Psi_p = \Psi \frac{a_1 + a_2}{b_1 + b_2} \frac{b_1 + b_2 \Psi'}{a_1 + a_2 \Psi'} = \Psi \frac{\frac{b_2}{b_1 + b_2} \Psi' + \frac{b_1}{b_1 + b_2}}{\frac{a_2}{a_2 + b_2} \Psi' + \frac{a_1}{a_1 + a_2}}$$

soit :

$$\Psi_p = \Psi \frac{\pi_1 \Psi' + 1 - \pi_1}{\pi_2 \Psi' + 1 - \pi_2}$$

où π_1 = Prob (C^-/E^- et M^+)
π_2 = Prob (C^-/E^+ et M^+)

On obtient une relation semblable au cas précédent (risque relatif) mais en plus de la conditions $\Psi' \neq 1$ pour qu'il y ait confusion, il faut que $\pi_1 \neq \pi_2$ c'est à dire que chez les malades E et C soient liés. Un même calcul montre que la formule ci-dessus est également vraie si π_1 = Prob (C^+/E^+ et M^-) et π_2 = Prob (C^+/E^- et M^-).

Il y a donc confusion si, à la fois, C est lié à la maladie par classe de E et si C est lié à E conditionnellement au groupe malade et au groupe témoin. La condition d'absence d'interaction entre E et C posée au départ implique que si C et E sont liés dans le groupe malade, ils le sont aussi dans le groupe témoin et réciproquement.

1. 4 - Conditions pour qu'une relation soit entièrement "expliquée" par un facteur de confusion.

Raisonnons en prenant les notations des risques relatifs :

$$\text{Si } R = 1 \text{ alors } R_p = \frac{R' p_1 + 1 - p_1}{R' p_2 + 1 - p_2} \text{ ou } R' - 1 = \frac{R_p - 1}{p_1 - p_2 R_p}$$

R' et R_p étant supérieurs à 1, cette relation implique :

$p_1 - p_2 R_p > 0$ soit $\frac{p_1}{p_2} > R_p$

$p_1 - p_2 R_p < 1$ soit $R' > R_p$

Il est donc nécessaire que le facteur de confusion présente dans chaque classe de E un risque relatif (R') au moins aussi grand que celui (R_p) observé pour E dans l'ensemble de la population et qu'il soit au moins R_p fois plus fréquent en présence de E qu'en son absence.

../...

Par exemple si on retient un risque relatif de cancer du poumon de 9 pour le tabagisme, pour que cette relation soit entièrement explicable par un facteur de confusion, ce dernier devrait présenter un risque relatif au moins 9 chez les fumeurs et les non fumeurs et devrait être 9 fois plus fréquent, au moins, chez les fumeurs que chez les non fumeurs.

Ceci montre que plus un risque relatif est élevé, plus il est plausible qu'il ne puisse pas être expliqué par un facteur de confusion et donc qu'il soit possiblement causal.

2. LE FACTEUR DE CONFUSION EST QUALITATIF A PLUSIEURS CLASSES

Les formules obtenues généralisent celles ci-dessus sans difficulté. Ecrivons celle correspondant au choix de l'odds ratio.

Soit Ψ_p l'odds ratio pour E dans l'ensemble de la population, Ψ l'odds ratio pour E dans chaque classe C_i du facteur C (supposé constant quel que soit i) et Ψ_i l'odds ratio de la classe C_i par rapport à la classe C_0 dans chaque classe de E. Avec des notations évidentes, on a pour tout i :

$$\Psi = \frac{a_i\, d_i}{b_i\, c_i} \quad \text{et} \quad \Psi_i = \frac{b_i\, d_0}{b_0\, d_i}$$

d'où $$\Psi_r = \frac{\sum a_i \sum d_i}{\sum b_i \sum c_i} = \Psi \frac{\sum d_i}{\sum c_i} \frac{\sum \frac{b_0 c_i}{d_0} \Psi_i}{\sum \frac{b_0 d_i}{d_0} \Psi_i}$$

soit $$\Psi_r = \Psi \frac{\sum d_i \sum c_i \Psi_i}{\sum c_i \sum d_i \Psi_i} = \Psi \frac{\sum c_i \Psi_i / \sum c_i}{\sum d_i \Psi_i / \sum d_i}$$

ou $$\boxed{\Psi_r = \Psi \frac{\sum_{i=0}^{k} \Psi_i \pi_{1i}}{\sum_{i=0}^{k} \Psi_i \pi_{2i}}}$$ avec $\pi_{1i} = \text{Prob}\left(C_i / E^+ \text{ et } M^-\right)$

$\pi_{2i} = \text{Prob}\left(C_i / E^- \text{ et } M^-\right)$

../...

Le rapport de confusion s'écrit $\frac{\sum \Psi_i \, p_{1i}}{\sum \Psi_i \, p_{2i}}$. Il apparait une différence importante avec le cas où le facteur C n'avait que deux classes : si les conditions $\Psi_i = 1$ et $p_{1i} = p_{2i}$ quel que soit i sont suffisantes pour qu'il n'y ait pas confusion, elles ne sont pas nécessaires ainsi qu'en témoigne l'exemple suivant :

Le facteur C a 3 classes et les odds ratios par classe du facteur E (dichotomique) sont respectivement 1 (classe 0), 2 et 3. L'association de C et de E chez les témoins est décrite au tableau suivant :

	C_0	C_1	C_2
Ψ_i	1	2	3
p_{1i}	0.5	0.3	0.2
p_{2i}	0.4	0.5	0.1

Le facteur C n'entraîne pas de confusion dans la relation de E avec la maladie car le rapport de confusion s'écrit :

$$\frac{0.5 \times 1 + 0.3 \times 2 + 0.2 \times 3}{0.4 \times 1 + 0.5 \times 2 + 0.1 \times 3} = 1$$

3. AUTRES CAS

Des développements du même ordre peuvent être obtenus. Par exemple si le facteur x est quantitatif et le facteur y de confusion est également quantitatif, le modèle logistique s'écrit :

Logit P = $\lambda_p x + \mu$ si P est la probabilité de la maladie. Lorsque l'on tient compte du facteur de confusion supposé sans interaction avec x :

Logit P = $\lambda x + \lambda' y + \mu'$

Si la régression de y/x est supposée linéaire, on a alors :

$$\boxed{\lambda_p = \lambda + a_{y/x} \, \lambda'}$$ où $a_{y/x}$ est la pente de cette régression.

../...

4. STRATIFICATION ET AJUSTEMENT

Nous venons de voir les principes généraux qui permettent de tenir compte, dans la mesure de la relation, de l'effet de confusion créé par une autre variable C : nous avons effectué un ajustement. Dans le cas où le ou les facteurs de confusion sont connus avant la réalisation de l'étude on peut se demander si dans sa planification il n'est pas possible d'en tenir compte afin que la mesure de la relation soit débarrassée du biais qu'ils entraînent. Pour cela il convient de distinguer différents cas selon la nature de l'étude entreprise. On n'envisagera pas le cas de l'étude de type I qui par définition recueille l'information sur l'ensemble d'une population.

4. 1 - Etudes de type II

Il s'agit de comparer un groupe de sujets exposés à E (E^+) à un groupe non exposé (E^-), la mesure de la relation étant le risque relatif. Dans le recrutement des sujets, il est toujours possible de faire en sorte que la proportion de sujets C^+ chez les E^+ soit la même que chez les E^-. Nous réalisons ainsi une stratification sur le facteur de confusion C.

Si on reprend les notations du paragraphe 1.2 ci-dessus, la stratification entraîne la condition $p_1 = p_2$ qui suffit à assurer l'égalité $R_p = R$ et donc à débarrasser le risque relatif de l'effet de confusion. Par exemple, dans une étude prospective sur la bronchite chronique en relation avec un facteur d'environnement on stratifiera les sujets sur leur consommation de tabac, c'est à dire que l'on prendra un pourcentage fixe d'exposés et de non exposés dans chaque strate constituée d'une classe de tabagisme.

Dans une telle étude, il est possible d'étudier également l'effet du facteur C par le risque relatif R' et évidemment discuter de son interaction éventuelle avec le facteur E.

Remarquons que l'appariement est un cas particulier de la stratification, il permet "d'équilibrer" les deux groupes E^+ et E^- pour un ensemble de facteurs de confusion ce qui en pratique est difficile à réaliser par stratification en classes.

../...

4. 2 - Etudes de type III

Dans chaque groupe, malades ou témoins, la proportion d'exposés à E est aléatoire, on ne peut donc réaliser la condition $\pi_1 = \pi_2$ (paragraphe 1.3) puisque ces proportions sont conditionnelles à chaque groupe.

La seule "standardisation" possible est de prendre la même proportion de C^+ dans le groupe malade et dans le groupe témoin. Il est important de remarquer que cette standardisation ne réalise pas l'autre condition $\Psi' = 1$ suffisante pour la disparition de l'effet de confusion puisque l'absence de relation facteur C x maladie devrait être conditionnelle à l'exposition au facteur E. Dans ce cas, le biais sur Ψ_p ne peut être totalement supprimé. Cependant, Ψ_p ne peut plus être quelconque (a priori) par rapport à Ψ en effet on montre que si :

. $\Psi > 1$ alors $1 < \Psi_p < \Psi$

. $\Psi < 1$ alors $\Psi < \Psi_p < 1$

C'est à dire que l'effet de confusion résiduel conduit à diminuer l'intensité de la relation facteur x maladie sans en changer le sens.

Exemple :

	C^+		C^-			
	E^+	E^-	E^+	E^-	E^+	E^-
M^+	20	30	40	10	60	40
M^-	20	80	60	40	80	120
	$\Psi = 2.67$		$\Psi = 2.67$		$\Psi_p = 2.56$	

Il conviendra donc de tenir compte de la stratification dans l'analyse (faire l'ajustement) pour obtenir une estimation non biaisée de Ψ.

5. QUAND FAUT-IL TENIR COMPTE D'UN FACTEUR DE CONFUSION ?

Les développements précédents pourraient donner l'impression qu'il convient de multiplier le nombre de facteurs "possiblement de confusion" dont il faut tenir compte dans l'analyse, voire dans la planification, afin d'éviter

../...

que l'un ou plusieurs d'entre eux n'introduise un biais important dans l'étude du facteur d'intérêt.

Rappelons la condition pour que C soit effectivement un facteur de confusion : C est lié "causalement" à la maladie, C est lié au facteur E (sans préciser le type de liaison). La situation la plus courante est celle où on tiendrait compte d'un facteur C alors qu'il n'est pas lié à la maladie indépendamment de E.

Tenir compte de C n'enlève aucun biais dans la relation E x maladie puisqu'il n'est pas facteur de confusion, mais ont montre que la précision de l'estimation de la relation E x M est affectée par le calcul d'ajustement et la puissance de l'étude est diminuée (voir chapitre plus loin).

Par exemple, ce n'est pas parce que les fumeurs ont souvent les doigts jaunis que ce dernier facteur doive être considéré comme de confusion pour la relation tabac x cancer du poumon. On imagine aisément la perte de puissance que cela entraînerait puisque dans chacun des deux groupes "doigts jaunis ou non" la variabilité de la consommation de tabac est considérablement réduite (gros fumeurs dans le premier, petits ou non fumeurs dans le second).

Cet exemple est bien entendu caricatural et en pratique il n'est pas toujours aisé de définir les facteurs à prendre en compte dans la planification et l'analyse.

D'autres situations peuvent se produire pour lesquelles il convient de s'interroger avant de tenir compte d'un facteur C, c'est la signification biologique des facteurs étudiés qui permet de guider le choix.

Par exemple, supposons que E soit une "cause" du facteur C et que l'on ait un modèle simple du type :

E (obésité) ⟶ C (hypertension artérielle) ⟶ M (infarctus)

Considérer l'hypertension comme un facteur de confusion pour la relation obésité x infarctus conduit à considérer qu'il n'y a pas de relation entre l'obésité et l'infarctus alors qu'en réalité ce facteur lui est lié par une "chaine causale".

../...

Un autre exemple de ce type est lorsque les facteurs C et E sont deux aspects du même phénomène : si l'hypertension artérielle est cause d'infarctus la mesure de la pression diastolique ne peut être considérée comme un facteur de confusion pour la mesure de la pression systolique car où serait le biais qu'il permettrait de supprimer ? Dans tous ces cas, on parle de "sur ajustement" (dans l'analyse) ou de "sur appariement" (dans la planification), la prise en compte du facteur C détruit la relation cherchée entre E et M.

CHAPITRE VIII

LES BIAIS DANS LES ENQUETES EPIDEMIOLOGIQUES

EVALUATION DE LEURS EFFETS

1 - LE PROBLEME

Ce chapitre concerne les biais qui, dans une enquête épidémiologique, peuvent conduire à une estimation erronée de la relation entre la maladie et un facteur de risque. Il ne traite pas des biais portant sur la mesure de la seule maladie (ou des seuls facteurs de risque). Dans ces derniers cas la solution consiste en l'étude de groupes "représentatifs" de la population, à définir précisément, que l'on étudie ou même en l'enregistrement systématique de faits concernant des populations entières (exemple des registres).

La littérature relative à ce problème est très importante. Dans la référence [1], plus d'une cinquantaine (!) de biais possibles sont listés. La plupart des articles restent cependant "descriptifs". Le but de ce chapitre fortement inspiré de [2] est de présenter une approche "quantitative" du problème.

2 - LES SOURCES DE BIAIS

On peut distinguer 3 sources possibles de biais :

- les facteurs de confusion, traités au chapitre précédent.
- les biais de sélection qui entrainent que la mesure de la relation maladie x facteur sur la population d'étude n'est pas égale à la mesure dans la population à laquelle on s'intéresse (population "cible"). Un cas particulier très important est celui

.../...

d'un mauvais choix des témoins (non malades dans les enquêtes cas-témoins, non exposés dans les enquêtes cohortes).

- les biais dusà des erreurs de classification portant sur la maladie ou le facteur d'exposition.

3 - LES BIAIS DE SELECTION

3-1. Le modèle

La table I représente la population "cible", la table II la population réelle (celle qui est en fait étudiée) d'où proviennent les échantillons sur lesquels porte l'analyse (Table III). Ces échantillons sont représentatifs de la population réelle. Il peut y avoir biais parce que la population réelle est différente de la population "cible".

	E^+	E^-
M^+	A	B
M^-	C	D

(I) Population cible

$\longrightarrow$

	E^+	E^-
M^+	$A'=\alpha A$	$B'=\beta B$
M^-	$C'=\gamma C$	$D'=\delta D$

(II) Population réelle

$\longrightarrow$

	E^+	E^-
M^+	a	b
M^-	c	c

(III) Echantillon

α, β, γ, δ, désignent les probabilités qu'un individu de la population cible appartienne à la population réelle, selon son état (M^+ ou M^-) et son exposition $[E^+$ ou $E^-]$.

Les vraies mesures de l'effet sont :

$$\Psi = \frac{AD}{BC} \quad \text{et} \quad R = \frac{A}{A+C} \Big| \frac{B}{B+D} = \frac{P_1}{P_0} \qquad (P \text{ probabilité de maladie})$$

alors que l'échantillon estime

$$\Psi^* = \frac{A'D'}{B'C'} \quad \text{et} \quad R^* = \frac{A'}{A'+C'} \Big| \frac{B'}{B'+D'}$$

On a les relations évidentes

$$\Psi^* = \frac{\alpha\delta}{\beta\gamma}\Psi \quad \text{et} \quad R^* = \frac{\alpha A}{\alpha A+\gamma C} \Big| \frac{\beta B}{\beta B+\delta D} \qquad (1)$$

.../...

Si on introduit les quantités

$\tau_M = \frac{\alpha}{\beta}$: rapport des probabilités de sélection (ou de surveillance) des E^+ et E^- chez les malades

$\tau_{\bar{M}} = \frac{\gamma}{\delta}$: rapport des probabilités de sélection (ou de surveillance) des E^+ et E^- chez les non malades

$\tau_E = \frac{\alpha}{\gamma}$: rapport des probabilités de sélection (ou de surveillance) des M^+ et M^- chez les exposés

$\tau_{\bar{E}} = \frac{\beta}{\delta}$: rapport des probabilités de sélection (ou de surveillance) des M^+ et M^- chez les non exposés.

les relations (1) s'écrivent

$$\psi^* = \frac{\alpha\delta}{\beta\gamma}\psi = \frac{\tau_M}{\tau_{\bar{M}}}\psi = \frac{\tau_E}{\tau_{\bar{E}}}\psi$$

$$R^* = R\,\frac{\alpha(A+C)(\beta B+\delta D)}{\beta(B+D)(\alpha A+\gamma C)} = R\,\frac{\alpha}{\beta}\,\frac{\beta P_0+\delta(1-P_0)}{\alpha P_1+\gamma(1-P_1)}$$

$$= R\,\frac{\alpha}{\beta}\cdot\frac{\delta}{\gamma}\,\frac{\tau_{\bar{E}}P_0+(1-P_0)}{\tau_E P_1+(1-P_1)} = R\,\frac{\tau_E}{\tau_{\bar{E}}}\,\frac{\tau_{\bar{E}}P_0+(1-P_0)}{\tau_E P_1+(1-P_1)}$$

(dans le cas d'une maladie rare P_0 et P_1 sont voisins de 0 et $R^* \sim R\,\frac{\alpha\delta}{\beta\gamma}$ qui est la même relation que pour ψ).

3-2. Discussion

3-2-1. Odds-ratio

$\psi^* = \psi$ si $\tau_M = \tau_{\bar{M}}$ ou $\tau_E = \tau_{\bar{E}}$

Un cas particulier est celui où $\tau_M = \tau_{\bar{M}} = 1$, qui signifie que la sélection (ou la surveillance) est indépendante de l'exposition tant chez les malades que chez les non malades. De même il n'y a

.../...

pas de biais si la sélection est indépendante de l'état du sujet chez les exposés et les non exposés.

Ψ^* est plus grand que Ψ si $\tau_M > \tau_{\bar{M}}$ (ou $\tau_0 > \tau_{\bar{0}}$), et plus petit dans le cas contraire.

Si Ψ est supérieur à 1 (relation positive entre le facteur et la maladie) et $\Psi^* > \Psi$ la relation est surestimée ; si Ψ est inférieur à 1 (relation négative) et $\Psi^* > \Psi$, la relation est sous-estimée si $\Psi^* < 1$, mais peut être trouvée inversée si $\Psi^* > 1$.

3-2-2. Risque relatif

$$R^* = R \text{ si : } \quad \frac{\tau_{\bar{E}}}{\tau_E} = \frac{\tau_{\bar{E}} P_0 + (1-P_0)}{\tau_E P_1 + (1-P_1)} \ldots$$

Contrairement à ce qui se passe pour Ψ, même si $\tau_E = \tau_{\bar{E}} = \tau$ il y a biais. Il vaut

$$\frac{rP_0 + (1-P_0)}{rP_1 + (1-P_1)}$$

Il est positif si $rP_0 + (1-P_0) > rP_1 = (1-P_1)$ soit $(r-1)(P_1-P_0)<0$.
On peut donc dresser le tableau ci-dessous qui donne le sens du biais selon la valeur de r et $R = \frac{P_1}{P_0}$

	$R < 1$	$R > 1$
$r < 1$	$R^* < R$	$R^* > R$
$r > 1$	$R^* > R$	$R^* < R$

Si $r < 1$ ($\alpha < \gamma$ et $\beta < \delta$) les malades sont moins sélectionnés (ou moins suivis que les non malades) ; c'est la situation inverse qui prévaut si $r > 1$.

.../...

3-3. Quelques situations types (parmi beaucoup d'autres)

a) Etude transversale des relations entre M et une exposition professionnelle : les malades exposés sont plus souvent soustraits au risque que les non malades, ce qui se traduit par $\alpha < \gamma$ et $\beta > \delta$

b) Etude cas-témoins (sur des cas prévalents) et mortalité différentielle selon E : $\alpha < \beta$

c) Problème des non répondants (à une convocation, ...) de façon différentielle

d) Enquête cohorte où les sujets exposés (par exemple à l'amiante) sont mieux suivis quant à la maladie (par exemple, le cancer du poumon)

e) Enquête cas-témoins où les malades sont mieux interrogés sur E (ou répondent plus facilement)

f) L'exposition entraine un symptome qui conduit à rechercher la maladie. Un exemple célèbre est constitué par les études destinées à rechercher si la prise d'oestrogènes est liée au cancer de l'endomètre. La prise d'oestrogènes occasionne des saignements qui poussent la femme à consulter. Ainsi, la proportion de cas diagnostiqués sera plus grand chez E^+ que chez E^- ($\alpha > \beta$).

3-4. Un exemple classique : le biais de Berkson

Il concerne les enquêtes cas-témoins où les sujets des 2 groupes sont choisis dans des populations hospitalières.

Soit à rechercher la relation entre M et E où E est lui-même une maladie (ou un symptome) qui peut conduire à l'hospitalisation. Les témoins sont des sujets atteints de la maladie M'.

Appelons p_1, p_2, p_3 les fréquences de M, E, M' dans la population générale et admettons qu'il y a indépendance entre ces 3 maladies.

Le tableau suivant donne les probabilités des diverses combinaisons des maladies dans la population générale.

.../...

	E^+		E^-	
M	MEM' $p_1p_2p_3$	ME $p_1p_2q_3$	MM' $p_1q_2p_3$	M $p_1q_2q_3$
M'	M'E $q_1p_2p_3$		M' $q_1q_2p_3$	
ni M ni M'	E $q_1q_2p_3$		- $q_1q_2q_3$	

Soient α_1, α_2, α_3 les taux d'hostpialisation des 3 maladies et admettons encore qu'il y ait indépendance. Ainsi un sujet (MEM') a la probabilité $1 - (1-\alpha_1)(1-\alpha_2)(1-\alpha_3)$ d'être hospitalisé, etc ...

Exemple numérique [2] : $p_1 = p_2 = p_3 = .10$

$\alpha_1 = .05 \quad \alpha_2 = .15 \quad \alpha_3 = .20$

Le tableau ci-dessous donne les effectifs de chacun des groupes (on a supposé que la population comprend 100.000 sujets) et leurs probabilités d'hospitalisation.

	E^+		E^-	
M	MEM' 100 .354	ME 900 .1925	MM' 900 .24	M 8100 .05
M'	M'E 900 .32		M' 8100 .20	

	E^+	E^-
M	1000	9000
M'	900	8100

Population générale

$\Psi = 1$

	E^+	E^-
M	208,65	621
M'	288	1620

Population hospitalière

$\Psi^* = 1,89$

Le biais crée une apparente liaison positive.

On décrit généralement le biais de Berkson en disant qu'un sujet qui a 2 maladies (ici M et E) a plus de chances d'être hospitalisé que s'il n'en a qu'une et donc plus de chances de figurer dans l'échantillon.

En réalité, le biais de Berkson peut créer une apparente liaison négative. Dans l'exemple précédent, supposons $\alpha = 1$. (Tous les cas sont hospitalisés). Les effectifs dans la population hospitalisée sont alors

	E^+	E^-
M	1000	9000
M'	288	1620

$\Psi = .625$

Pendant longtemps le biais de Berkson a été considéré comme une possibilité théorique. Ces dernières années, plusieurs exemples réels ont été publiés [1].

4 - LES ERREURS DE CLASSIFICATION

4-1. Le modèle

Les 2 tables IV et V donnent les effectifs théoriques en l'absence et en présence d'erreurs de classification.

	E^+	E^-
M^+	A	B
M^-	C	D

(IV)

	$\tilde{E}^+$	$\tilde{E}^-$
$\tilde{M}^+$	A'	B'
$\tilde{M}^-$	C'	D'

(V)

.../...

Introduisons les notions de sensibilité u et spécificité v

$$u(M/E^+) = Pr[\tilde{M}^+/M^+ \text{ et } E^+]$$
$$u(M/E^-) = Pr[\tilde{M}^+/M^+ \text{ et } E^-]$$

$$v[M/E^+] = Pr[\tilde{M}^-/M^- \text{ et } E^+]$$
$$v[M/E^-] = Pr[\tilde{M}^-/M^- \text{ et } E^-]$$

et des expressions analogues relatives à E.

On dit que les classifications de M et de E sont sans biais si

$$u(M/E^+) = u(M/E^-) = u_M \qquad u[E/M^+] = u[E/M^-] = u_E$$
$$v(M/E^+) = v(M/E^-) = v_M \quad \text{et} \quad v[E/M^+] = v[E/M^-] = v_E$$

c'est-à-dire si la classification de l'une des deux variables, M ou E, ne dépend pas du niveau de l'autre.

On peut, si les classifications de M et E sont indépendantes, calculer A', B', C', D' en fonction de A, B, C, D et des u et v, et donc mesurer le biais.

4-2. Discussion

En général, il n'existe pas de réponse claire quant à la valeur et au sens de ce biais.

Toutefois, si les classifications sont sans biais, on peut démontrer le résultat suivant, évident a priori : les erreurs de classifications sur l'une ou l'autre des 2 variables sous-estime la mesure de la relation facteur-maladie.

Vérifions le sur un exemple.

	E^+	E^-			
M^+	100	50	150	$u_E = .80$	$v_E = .90$
M^-	25	50	75	$u_M = .90$	$v_M = .80$

$\Psi = 4$

.../...

Le tableau ci-dessous donne les probabilités de classement

	$\tilde{M}^+ \tilde{E}^+$	$\tilde{M}^+ \tilde{E}^-$	$\tilde{M}^- \tilde{E}^+$	$\tilde{M}^- \tilde{E}^-$
$M^+ E^+$	$u_M u_E$	$u_M(1-u_E)$	$(1-u_M)u_E$	$(1-u_M)(1-u_E)$
$M^+ E^-$	$u_M(1-v_E)$	$u_M v_E$	$(1-u_M)(1-v_E)$	$(1-u_M)v_E$
$M^- E^+$	$(1-v_M)u_E$	$(1-v_M)(1-u_E)$	$v_M u_E$	$v_M(1-u_E)$
$M^- E^-$	$(1-v_M)(1-v_E)$	$(1-v_M)v_E$	$v_M(1-v_E)$	$v_m v_E$

Tous calculs faits on obtient les calssements

	$\tilde{E}^+$	$\tilde{E}^-$	
$\tilde{M}^+$	81,5	68,5	150
$\tilde{M}^-$	28,5	46,5	75
	110	115	

$\psi^* = 1,94 < 4$ Le biais est très important

4-3. Erreurs de classification en présence d'un facteur de confusion

Limitons-nous au cas d'une enquête cas-témoins et étudions diverses situations possibles à partir d'exemples.

4-3-1. Erreurs de classification sur E seul

Supposons $u_E = .80$ et $v_E = .90$

1er cas C est une variable de confusion.

Si les effectifs "réels" sont

100	50
25	50

C^+

$\psi = 4$

80	25
80	100

C^-

$\psi = 4$

180	75
105	150

Total

$\psi_p = 3,42$

.../...

les effectifs qui résultent du classement sont

85	65
25	50

C^+

$\psi^* = 2,62$

66,5	38,5
74	106

C^-

$\psi^* = 2,47$

151,5	103,5
99	150

Total

$\psi^*_p = 2,31$

Comme annoncé, la mesure de la relation maladie-facteur est sous-estimée ; de plus il y a création d'une "interaction".

On peut à l'inverse imaginer des situations ou une interaction réelle est annulée par des erreurs de classification.

Effectifs réels

85	100
80	200

C^+

$\psi^* = 2,125$

45	200
30	400

C^-

$\psi = 3$

Après classement

78	107
84	196

$\psi^* = 1,70$

56	189
64	366

$\psi^* = 1,69$

<u>2ème cas</u> C n'est pas facteur de confusion car $\psi' = 1$ (indépendance conditionnelle de M et C)

Effectifs réels

120	100
80	200

C^+

$\psi = 3$

45	200
30	400

C^-

$\psi = 3$

165	300
110	600

Total

$\psi_p = 3$

Après classement

106	114
84	196

C^+

$\psi^* = 2,17$

56	189
64	366

C^-

$\psi^* = 1,69$

162	303
148	562

Total

$\psi^* = 2,03$

Les conclusions sont les mêmes : il y a sous estimation de l'effet et création d'une interaction apparente.

3ème cas C n'est pas facteur de confusion car C et E sont conditionnellement indépendants

Effectifs réels

100	50
100	200

C^+

$\psi = 4$

200	100
50	100

C^-

$\psi = 4$

300	150
150	300

Total

$\psi_T = 4$

Après classement

85	65
100	200

$\psi^* = 2,615$

170	130
50	100

$\psi^* = 2,615$

255	195
150	300

$\psi^*_T = 2,615$

La conclusion est cette fois différente : la mesure de la relation est toujours sous-estimée, mais il n'y a pas création d'une interaction apparente.

4-3-2. Erreurs de classification sur C seul

Supposons $u_c = .80$ et $v_c = .90$

Effectifs réels

240	200
80	200

C^+

$\psi = 3$

30	100
40	400

C^-

$\psi = 3$

270	300
120	600

Total

$\psi_p = 4,5$

Dans ce cas, ne pas tenir compte de C conduit à une même sous-estimation de .

Après erreurs de classement sur C

195	170
68	200

$\widetilde{C^+}$

$\psi^* = 3,37$

75	130
52	400

$\widetilde{C^-}$

$\psi^* = 4,44$

Il y a surestimation de ψ : c'est normal, puisque ne pas tenir compte de C surestime ψ et que faire des erreurs sur C revient à ne pas le prendre totalement en compte. De plus il y a création d'une interaction (mais on peut construire des exemples ou une interaction réelle n'apparaît plus sur des observations mal classées).

4-3-3. Erreurs de classification sur C et E

Aucun résultat général ne peut être énoncé ; toutes les situations sont possibles.

5 - CONCLUSIONS

En ce qui concerne les biais de sélection, la notion de l'existence possible de très nombreux biais doit être présente à l'esprit de ceux qui établissent le protocole. Les enquêtes cas-témoins sont généralement les plus sujettes à des biais et posent le problème difficile du choix des témoins non-malades. Cet aspect est abordé dans la plupart des ouvrages d'épidémiologie générale.

Pour ce qui est des erreurs de classification l'exemple numérique qui figure en 4-2. montre que des erreurs relativement faibles peuvent entraîner une sous-estimation importante de la mesure de la relation exposition-maladie. Elles se traduisent donc d'une certaine façon par un "manque de puissance" dans le test de l'existence de cette relation.

Dans l'exemple, il y avait erreur sur, à la fois M et E ; toutefois dans les enquêtes cas-témoins il n'y a généralement pas d'erreur sur M ; de même dans les enquêtes cohorte ou un groupe exposé est comparé à un groupe non exposé, il n'y a vraisemblablement pas d'erreur sur E. Par contre dans les enquêtes cohorte où de nombreux facteurs E sont enregistrés sur un échantillon d'une population, il peut y avoir erreur sur E aussi bien que sur M.

CHAPITRE VIII - REFERENCES

[1] - D.L. SACKETT
Biais in Analytic Research
1979, J. Chron. Dis, 32, 51-63

[2] - D.G. KLEINBAUM, L.L. KUPPER, H. MORGENSTERN
Epidemiologic Research. Principles and quantitative methods.
1982, Lifetime Learning Pub.

[3] - S. GREENLAND
The effects of misclassification in the presence of covariates
1980, Am. J. Epid. 112, n° 4, 564-569.

CHAPITRE IX

ETUDE SIMULTANEE DE DEUX FACTEURS QUALITATIFS, INTERACTION, FACTEUR DE CONFUSION, TRAITEMENT STATISTIQUE SIMPLE

Lorsqu'on analyse l'association entre l'exposition à un facteur qualitatif et une maladie comme on l'a vu au Chapitre VI, on est amené dans certains cas à faire cette analyse en prenant en compte un facteur de confusion (Chapitre VII), et ce qu'il y ait eu ou non stratification a priori (voir Chapitre XII). Ce chapitre présente des méthodes simples de tests et de mesures d'association en ajustant sur un facteur de confusion. Il présente également des méthodes permettant de tester l'existence d'une interaction entre deux facteurs d'exposition, interaction qui a été définie au Chapitre V.

1 - NOTATIONS

1-1. Dans la population d'étude

Si C est un facteur de confusion à p classes : C_1, C_2, ..., C_i, .. C_p ; dans la classe (ou strate) i, on note :

$$P_{1i} = P\ (M^+ / E^+,\ C_i)$$

$$P_{oi} = P\ (M^+ / E^-,\ C_i)$$

$$P_{1i} = P\ (E^+ / M^+,\ C_i)$$

$$P_{oi} = P\ (E^+ / M^-,\ C_i)$$

et RR_i et OR_i respectivement les risque relatif et odds-ratio.

Rappelons (Chapitre VI) que dans les enquêtes de type échantillon représentatif ou exposés-témoins :

$$RR_i = P_{1i} \Big/ P_{oi} \qquad OR_i = \frac{P_{1i}}{1-P_{1i}} \Big/ \frac{P_{oi}}{1-P_{oi}}$$

.../...

Dans les enquêtes de type cas-témoins :

$$OR_i = \frac{p_{1i}}{1-p_{1i}} \Big/ \frac{p_{oi}}{1-p_{oi}}$$

1-2. Dans un échantillon de taille n, avec n_i sujets, dans la strate i ($\sum_1^p n_i = n$) :

	E^+	E^-	
M^+	a_i	b_i	n_{1i}
M^-	c_i	d_i	n_{oi}
	m_{1i}	m_{oi}	n_i

Dans chaque strate, on peut utiliser les tests et mesures d'association décrits au Chapitre VI, en particulier estimer RR_i ou OR_i selon le type d'enquête.

Si, avec la mesure d'association choisie, il n'y a pas d'interaction entre les deux facteurs C et E, c'est-à-dire si, selon le cas, les RR_i ou les OR_i sont les mêmes dans toutes les strates, on peut tester si, après ajustement sur le facteur de confusion, il y a association entre le facteur d'exposition et la maladie (paragraphe 2). On peut estimer cette association par des méthodes appropriées selon le type d'enquête (paragraphe 3). Pour faciliter l'exposé, le test d'existence d'une interaction entre E et C est présenté au paragraphe 4, mais les paragraphes 2 et 3 supposent vérifiée l'absence d'interaction.

2 - TEST DE L'ASSOCIATION ENTRE L'EXPOSITION AU FACTEUR ET LA MALADIE APRES AJUSTEMENT SUR UN FACTEUR DE CONFUSION

Ce test suppose l'absence d'interaction entre les deux facteurs, c'est-à-dire que dans toutes les strates les odds-ratio ont une valeur commune OR_c

$$OR_i = OR_c, \forall i$$

ou alternativement les risques relatifs ont une valeur commune RR_c

$$RR_i = RR_c, \forall i$$

.../...

L'hypothèse nulle d'absence d'association entre facteur et maladie s'exprime par $OR_c = 1$ (ou alternativement $RR_c = 1$).

Si on appelle A_i la variable aléatoire : nombre de malades chez les exposés dans la classe i du facteur C, et si les effectifs sont suffisants pour supposer que chaque A_i suit une loi normale conditionnellement aux marges* sous H_o, on peut calculer la moyenne et la variance de A_i par (voir formules (8) et (9) du Chapitre VI):

$$E(A_i, \psi = 1) = \frac{m_{1i}\, n_{1i}}{n_i}$$

$$V(A_i, \psi = 1) = \frac{n_{1i}\, m_{oi}\, m_{1i}\, m_{oi}}{n_i^2\,(n_i - 1)}$$

$$\text{soit} \quad V(A_i, \psi = 1) \sim \frac{n_{1i}\, n_{oi}\, m_{1i}\, m_{oi}}{n_i^3}$$

On compare le nombre total de malades observés chez les exposés: $\sum a_i$, au nombre attendu : $\sum E(A_i, \psi = 1)$ à l'aide de la quantité :

$$\frac{\left(\sum a_i - \sum E(A_i ; \psi = 1)\right)^2}{\sum V(A_i ; \psi = 1)}$$

qui suit sous H_o une loi du χ^2 à 1 ddl : c'est le test de Mantel Haenszel (MANTEL-HAENSZEL).

Remarques :

1. Ce test est parfois utilisé avec une correction de continuité :

$$\frac{\left(\left|\sum a_i - \sum E(A_i ; \psi = 1)\right| - \frac{1}{2}\right)^2}{\sum V(A_i ; \psi = 1)}$$

.../...

* : Le meilleur test, c'est-à-dire non biaisé, uniformément le plus puissant est conditionnel aux marges, c'est-à-dire que dans chaque strate on suppose connus les effectifs de sujets malades et non malades et de sujets exposés et non-exposés.

2. Le test de Mantel-Haenszel est identique au test d'ajustement de Cochran (COCHRAN, RUMEAU et al.) : en effet, on retrouve la formule de Cochran en remplaçant dans (4) les estimations des espérances et variances des A_i par les expressions (1) et (3); par exemple, dans le cas d'une enquête type exposés-témoins on arrive à la formule :

(6)
$$\frac{\sum \frac{m_{1i}\, m_{oi}}{n_i} (P_{1i} - P_{oi})^2}{\sum P_i (1 - P_i) \frac{m_{1i}\, m_{oi}}{n_i}}$$

$P_i = P(M^+)$ dans la strate i.

3 - ESTIMATION DE LA MESURE D'ASSOCIATION ENTRE L'EXPOSITION AU FACTEUR ET LA MALADIE, APRES AJUSTEMENT SUR UN FACTEUR DE CONFUSION dans l'hypothèse d'absence d'interaction entre les deux facteurs.

Rappelons que l'on pourra estimer le odds-ratio dans les trois types d'étude, alors qu'on ne pourra estimer le risque relatif que dans les études type échantillon représentatif ou exposés-témoins.

3-1. Estimation du risque relatif commun

(Etudes type échantillon représentatif ou exposés-témoins).

On suppose à nouveau que tous les RR_i sont égaux à une même quantité RR_c. Dans chaque strate on estime les risques relatifs $\widehat{RR}_i$ selon la formule (13) du Chapitre VI. L'estimateur du risque relatif commun $\widehat{RR}_c$ est construit à partir d'une moyenne pondérée des $\widehat{RR}_i$. En fait on estime Log $\widehat{RR}_c$ et non directement $\widehat{RR}_c$.

(7)
$$\text{Log } \widehat{RR}_c = \frac{\sum W_{iR} \text{ Log } \widehat{RR}_i}{\sum W_{iR}}$$

où :

(8)
$$W_{iR} = 1 \,/\, \text{Var Log } \widehat{RR}_i :$$

Var Log $\widehat{RR}_i$ étant donné par la formule (VI - 14)

On a alors :

(9)
$$\text{Var Log } \widehat{RR}_c = 1 \,/ \sum W_{iR}$$

.../...

Si on suppose que Log $\hat{RR}_c$ suit une loi normale, on peut donc obtenir un intervalle de confiance au risque α pour Log RR_c. puis RR_c.

Remarque : La transformation logarithmique choisie pour estimer le risque relatif commun permet :

- le calcul de la variance de Log $\hat{RR}_c$
- le maintien de la symétrie entre les 2 niveaux d'exposition, c'est-à-dire que :

$$\hat{RR}_c \; E^+ / E^- = 1 / \hat{RR}_c \; E^- / E^+$$

3-2. Estimation du odds-ratio commun (quel que soit le type d'enquête)

On suppose que tous les OR_i (notés ψ_i) sont égaux à une même quantité ψ_c.

3-2-1. Estimation ponctuelle de ψ_c

3-2-1-1. Estimation du maximum de vraisemblance

L'estimateur du maximum de vraisemblance de ψ_c, $\hat{\psi}_{cmv\ cond}$, conditionnellement aux marges comme dans le cas non ajusté (Chapitre VI, paragraphe 4) est la solution de l'équation :

$$\sum a_i = \sum E\,(A_i / m_{1i}, m_{oi}, n_{1i}, n_{oi} ; \psi_c) \tag{10}$$

qui est compliquée à résoudre. L'estimation de la variance de $\psi_{cmv\ cond}$ nécessite aussi des calculs compliqués.

Dans la pratique cette méthode est peu utilisée.

3-2-1-2. Estimation par la méthode des logits (WOOLF, ARMITAGE)

On estime Log $\psi_{c\ell}$ par une moyenne des Log $\hat{\psi}_i$ pondérés par l'inverse des variances. (Comme pour RR_c, on estime Log $\hat{\psi}_{c\ell}$ et non directement $\hat{\psi}_c$).

$$\text{Log}\,\hat{\psi}_{c\ell} = \frac{\sum W_{i\,l}\ \text{Log}\,\hat{\psi}_i}{\sum W_{i\,l}} \tag{11}$$

Rappelons que :

$$\hat{\psi}_i = \frac{a_i d_i}{b_i c_i}$$

.../...

et $$W_{il} = 1 / \text{Var Log } \hat{\psi}_i = \left(\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}\right)^{-1}$$

(Chapitre VI, paragraphe 4-3-2.)

$$\text{Var Log } \hat{\psi}_{cl} = 1 / \sum W_{il}$$

Cette méthode n'est utilisable que si les effectifs de chaque cellule sont suffisamment grands. Si, dans une strate, un des effectifs est nul, la contribution de cette strate n'est pas définie et la méthode ne peut pas être utilisée.

3-2-1-3. Méthode de Mantel-Haenszel (MANTEL-HAENSZEL, ARMITAGE)

(12) $$\hat{\psi}_{c\,mh} = \frac{\sum a_i d_i / n_i}{\sum b_i c_i / n_i}$$

(moyenne des $\hat{\psi}_i$ pondérés par $\frac{b_i c_i}{n_i}$, $\frac{n_i}{b_i c_i}$ étant de l'ordre de grandeur de la variance de $\hat{\psi}_i$, dans le cas où ψ_c est voisin de 1

C'est l'estimateur de ψ_c le plus utilisé en pratique utilisable même si certains effectifs observés sont faibles ou nuls, et très proche de $\hat{\psi}_{cmv\,cond}$ si les effectifs sont importants (BRESLOW-DAY).

3-2-2. Intervalle de confiance pour le odds-ratio commun

3-2-2-1. Généralisation de la méthode de Cornfield (BRESLOW-DAY)

Les limites ($\hat{\psi}_{cI}$, $\hat{\psi}_{cS}$) de l'intervalle de confiance au risque α pour ψ_c sont obtenues en supposant que $\sum a_i$ suit une loi normale de moyenne

$$\sum E(A_i ; \psi_c)$$

et de variance

$$\sum \text{Var}(A_i ; \psi_c)$$

(voir Chapitre VI, paragraphe 4-2-2)

$\hat{\Psi}_{cl}$ et $\hat{\Psi}_{cS}$ sont solutions de :

(13)
$$\begin{cases} \sum a_i - \sum E(A_i ; \Psi_{cl}) - \frac{1}{2} = \varepsilon_\alpha \sqrt{Var\,(A_i ; \Psi_{cl})} \\ \sum a_i - \sum E(A_i ; \Psi_{cS}) + \frac{1}{2} = -\varepsilon_\alpha \sqrt{Var(A_i ; \Psi_{cS})} \end{cases}$$

3-2-2-2. Méthode des logits

Intervalle de confiance construit à partir de Log $\hat{\Psi}_{c\ell}$:

(14)
$$(\text{Log}\,\hat{\Psi}_{cl} , \text{Log}\,\hat{\Psi}_{cS}) : \text{Log}\,\hat{\Psi}_{c\ell} \pm \varepsilon_\alpha \sqrt{\frac{1}{\Sigma W_{i1}}}$$

et on obtient $(\hat{\Psi}_{cl}, \hat{\Psi}_{cS})$ en prenant les exponentielles des bornes de cet intervalle.

3-2-2-3. Méthode de Miettinen

Comme dans le cas non ajusté (Chapitre VI, paragraphe 4-2-4.) en utilisant $\hat{\Psi}_{cmh}$ comme estimateur de Ψ_c et le test de Mantel Haenszel comme valeur pour le χ^2

(15)
$$(\hat{\Psi}_{cl}, \hat{\Psi}_{cS}) : \hat{\Psi}_{cmh}^{\,(1 \pm \varepsilon_\alpha / \sqrt{\chi^2_{mh}})}$$

3-2-2-4. Comparaison des trois méthodes

Ces trois méthodes ont les mêmes avantages et inconvénients respectifs que dans le cas non ajusté (Chapitre VI, paragraphe 4-2).

La méthode de Cornfield est la meilleure, mais beaucoup plus compliquée et rarement utilisée. La méthode des logits est utilisable à condition que les effectifs observés ne soient pas trop faibles, la méthode de Miettinen, à condition que Ψ_c soit voisin de 1 (BRESLOW-DAY).

4 - TEST D'INTERACTION ENTRE DEUX FACTEURS D'EXPOSITION

Alors que le test de l'association entre un premier facteur d'exposition et une maladie, après ajustement sur un deuxième facteur, est indépendant de la mesure d'association choisie, la

.../...

définition même de l'interaction dépend de la mesure d'association choisie, risque relatif ou odds-ratio (Chapitre IV).

4-1. Test d'interaction basé sur la décomposition du Chi-2

4-1-1. Principe (FLEISS)

Soit Y une mesure d'association, Y_i cette mesure dans la strate i et $w_i = \frac{1}{\text{var } Y_i}$. On suppose de plus que les Y_i suivent une loi normale. Donc sous l'hypothèse nulle d'absence d'association (c'est-à-dire espérance des Y_i nulle), $Y_i \sqrt{w_i}$ suit une loi normale centrée réduite, et

$$\sum_{i=1}^{p} W_i Y_i^2 \quad \text{suit une loi du } \chi^2 \text{ à p ddl.}$$

Cette quantité peut être décomposée en deux quantités permettant l'une de tester l'association moyenne sur les p strates

(16) $$\frac{(\sum W_i Y_i)^2}{\sum W_i} \qquad \chi^2 \text{ à 1 ddl}$$

l'autre, l'interaction entre le facteur d'exposition et le facteur de stratification :

(17) $$\sum W_i Y_i^2 - \frac{(\sum W_i Y_i)^2}{\sum W_i} : \chi^2 \text{ à p - 1 ddl}$$

Si Y_c est l'estimation commune de la mesure d'association Y sur l'ensemble des strates : $Y_c = \Sigma w_i Y_i / \Sigma w_i$, la formule (17) peut s'écrire :

(17bis) $$\Sigma w_i (Y_i - Y_c)^2$$

4-1-2. Test de l'interaction à partir du risque relatif

(enquêtes de type échantillon représentatif ou exposés-témoins). $Y = \text{Log RR}$.

Test de l'interaction par la quantité

(18) $$\sum W_{iR} (\text{Log } \hat{RR}_i)^2 - \frac{(\sum W_{iR} \text{ Log } \hat{RR}_i)^2}{\sum W_{iR}} = \sum W_{ie} (\text{Log } \hat{RR}_i - \text{Log } \hat{RR}_e)^2$$

4-1-3. Test de l'interaction à partir du odds-ratio en utilisant les logits.

$Y = \text{Log } \psi$

Test de l'interaction par la quantité

(19) $$\sum W_{i\ell} (\text{Log } \hat{\psi}_i)^2 - \frac{(\sum W_{i\ell} \text{ Log } \hat{\psi}_i)^2}{\sum W_{i\ell}} = \sum W_{ie} (\text{Log } \hat{\psi}_i - \text{Log } \hat{\psi}_e)^2$$

Remarque : Cette méthode ne peut être utilisée que si les effectifs observés sont suffisamment élevés. Par ailleurs, en pratique, on n'utilise pas la formule 16 pour tester l'existence d'une association, mais plutôt le test de Mantel-Haenszel vu plus haut.

4-2. Test construit sur l'estimateur du odds-ratio commun

Si $\hat{\Psi}_c$ est l'estimateur du odds-ratio commun estimé sous l'hypothèse d'absence d'interaction, mais s'il y a interaction, et si, dans la strate i le odds-ratio est plus élevé que la moyenne, le nombre a_i de cas observés chez les exposés sera plus élevé que le nombre attendu calculé sur $\hat{\Psi}_c$, $E(A_i ; \hat{\Psi}_c)$; on peut donc tester l'hypothèse d'un odds-ratio commun par la quantité

$$\sum \frac{(a_i - E(A_i ; \hat{\Psi}_c))^2}{V(A_i ; \hat{\Psi}_c)} \qquad (20)$$

qui suit un χ^2 à $p - 1$ ddl.

En pratique, on peut prendre pour $\hat{\Psi}_c$ l'estimateur de Mantel-Haenszel $\hat{\Psi}_{cmh}$ (formule 12), et $E(A_i ; \hat{\Psi}_{cmh})$ et $V(A_i ; \hat{\Psi}_{cmh})$ sont calculés dans chaque strate à partir des formules (22) et (23) du Chapitre VI.

Remarques : 1) Si le nombre de strates est élevé, et les effectifs observés faibles, la distribution de (20) peut ne pas suivre une loi du χ^2.

2) Ce test est peu puissant contre certaines hypothèses alternatives comme celle d'une tendance linéaire du odds-ratio par rapport au critère de stratification.

Dans ce cas on peut utiliser :

$$\frac{\left[\sum_1^p x_i (a_i - E(A_i ; \hat{\Psi}_c))\right]^2}{\sum x_i^2 V(A_i ; \hat{\Psi}_c) - \left[\sum x_i V(A_i ; \hat{\Psi}_c)\right]^2 / \sum V(A_i ; \hat{\Psi}_c)} \qquad (21)$$

qui suit une loi du χ^2 à 1 ddl.

3) Il n'existe pas de bon test d'interaction, dans la mesure où tous les tests sont peu puissants (BRESLOW-DAY). Les conséquences en sont très gênantes lorsqu'on cherche à détecter une interaction plutôt qu'à vérifier une homogénéité pour faire un test d'association ou estimer cette association.

.../...

5 - FACTEUR D'EXPOSITION A PLUSIEURS NIVEAUX

La statistique qui généralise au cas où l'on ajuste sur un facteur de confusion le test d'association entre une maladie et un facteur d'exposition à plusieurs niveaux (Chapitre VI, paragraphe 5-2) fait appel à du calcul matriciel (BRESLOW-DAY).

Les méthodes exposées au Chapitre XIII permettent de prendre en compte simultanément plusieurs facteurs à plusieurs niveaux et de tester les interactions.

6 - EXEMPLES

Exemple 1 : Association entre consommation d'alcool pendant la grossesse et hypotrophie en ajustant sur l'usage du tabac. (exemple 1 du Chapitre VI, paragraphe 2-2). Données tirées de KAMINSKI et al. (Detail des calculs : voir page IX-12).

On ajuste sur l'usage du tabac, parce que cette variable est un facteur de confusion : liée à l'hypotrophie et à la consommation d'alcool.

1 - Test de Mantel-Haenszel ajusté sur l'usage du tabac (en 2 classes : non fumeuses et fumeuses).

Sous l'hypothèse d'absence d'interaction

$$\frac{\left(\sum a_i - \sum \frac{m_{1i}\, n_{1i}}{n_i} \right)^2}{\sum \frac{n_{1i}\, n_{oi}\, m_{1i}\, m_{oi}}{n_i^3}} = \frac{(20 - 11{,}34)^2}{10{,}36} = 7{,}24$$

Chi 2 à 1 ddl significatif à $p < 0{,}01$.

2 - Estimation du risque relatif commun

Sous l'hypothèse d'absence d'interaction

$$\text{Log } \widehat{RR}_c = \frac{\sum W_{iR} \text{ Log } \widehat{RR}_i}{\sum W_{iR}} = \frac{12{,}03}{18{,}96} = 0{,}63$$

$$\left(W_{iR} = \left[\frac{c_i}{m_{1i}\, a_i} + \frac{d_i}{m_{oi}\, b_i} \right]^{-1} \right)$$

d'où $\widehat{RR}_c = 1{,}88$ (alors que le risque relatif non ajusté était de 1,94).

.../...

$$\text{Var Log } \hat{RR}_c = \frac{1}{\sum W_{iR}} = {}^{1}/_{18,96} = 0,0527$$

Intervalle de confiance à 95 % pour Log $\hat{RR}_c$

$$0,63 \pm 2\sqrt{0,0527}$$

$$[0,17 \ ; \ 1,09]$$

Intervalle de confiance à 95 % pour $\hat{RR}_c$

$[1,19 \ ; \ 2,97]$ très voisin de l'intervalle obtenu pour le risque relatif non ajusté.

3 - Test d'homogénéité du risque relatif chez les non fumeuses et les fumeuses.

$$\sum W_{iR}(\text{Log } \hat{RR}_i)^2 - \frac{(\sum W_{iR} \text{ Log } \hat{RR}_i)^2}{\sum W_{iR}} = 8,49 - \frac{12,03^2}{18,96} = 0,86$$

χ^2 à 1 ddl : test non significatif ; on ne rejette pas l'hypothèse d'homogénéité.

4 - Si dans cette étude type échantillon représentatif on s'intéresse au odds-ratio, si on estime le odds-ratio commun par la méthode Mantel-Haenszel on obtient

$$\hat{\Psi}_{mh} = 18,40 \, / _{9,74} = 1,89$$

très voisin de $\hat{RR}_c$, puisque la fréquence de l'hypotrophie est faible.

- Exemple 1 -

CONSOMMATION D'ALCOOL ET HYPOTROPHIE : AJUSTEMENT SUR L'USAGE DU TABAC

	Consommation de vin >40cl	Consommation de vin ≤40cl	
oui hypotr.	a_i	b_i	n_{1i}
non	c_i	d_i	n_{oi}
	m_{1i}	m_{oi}	n_i

Non fumeuses

16	154	170
316	6544	6860
332	6698	7030

Fumeuses

4	35	39
76	828	904
80	863	943

	a_i	$\hat{E}(A_i)$	$\hat{V}(A_i)$	P_{1i} %	P_{oi} %	$\widehat{RR}_i$	$\text{Log}\widehat{RR}_i$	W_{iR}	$W_{iR}\text{Log}\widehat{RR}_i$	$W_{iR}\text{Log}\widehat{RR}_i^2$	ψ_i	$a_i d_i/n_i$	$b_i c_i/n_i$
TOTAL	20	11,34	10,36					18,96	12,03	8,49		18,40	9,71

Exemple 2 : Mononucléose infectieuse et antécédent d'amygdalectomie en ajustant sur l'âge (exemple 2 du chapitre VI paragraphe 2-2 : données tirées de MILLER). (Detail des calculs, voir page IX.16).

1 - Test d'association ajusté sur l'âge (en 7 classes)

Test de Mantel-Haenszel - sous l'hypothèse d'absence d'interaction

$$\chi^2_{mh} = \frac{\left(\sum a_i - \sum \frac{m_{1i}\, n_{1i}}{n_i}\right)^2}{\sum \frac{m_{1i} m_{oi} n_{1i} n_{oi}}{n_i^3}} = \frac{(40 - 56{,}08)^2}{28{,}90} = 8{,}95$$

χ^2 à 1 ddl significatif à $p < 0{,}01$: la fréquence d'antécédents d'amygdalectomie est plus faible chez les sujets atteints de mononucléose infectieuse que chez les témoins.

2 - Estimation du odds-ratio commun par la méthode de Mantel-Haenszel

$$\hat{\psi}_{cmh} = \frac{\sum a_i\, d_i / n_i}{\sum b_i\, c_i / n_i} = 19{,}84 \,/\, 35{,}89 = 0{,}55$$

3 - Estimation du odds ratio commun par la méthode des logits

$$\text{Log } \hat{\Psi}_{c\ell} = \frac{\sum W_{i\ell} \text{ Log } \hat{\Psi}_i}{\sum W_{i\ell}} = - \frac{12,81}{23,70} = - 0,54$$

où $W_{i\ell} = \left(\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}\right)^{-1}$

d'où $\hat{\Psi}_{c\ell} = 0,58$

$$\text{Var Log } \hat{\Psi}_{c\ell} = \frac{1}{\Sigma W_{i\ell}} = 1:23,70 = 0,0422$$

Intervalle de confiance à 95 % pour Log $\Psi_{c\ell}$:

$$- 0,54 \pm 2 \sqrt{0,0422}$$

$$[-0,95 \ ; \ - 0,13]$$

Intervalle de confiance à 95 % pour $\Psi_{c\ell}$:

$$[0,39 \ ; \ 0,88]$$

4 - Intervalle de confiance pour Ψ_c par la méthode de Miettinen

$$\hat{\Psi}_{cmh}^{\left(1 \pm 2/\sqrt{\chi^2_{mh}}\right)} :$$

$$0,55^{\,1 \pm 2/\sqrt{8,95}}$$

$$(0,55^{\,1,67} \ ; \ 0,55^{\,0,33})$$

$$[0,37 \ ; \ 0,82]$$

5 - Test d'homogénéité du odds ratio dans les strates en utilisant les logits

$$\sum W_{i\ell} (\text{Log } \hat{\Psi}_i)^2 - \frac{(\sum W_{i\ell} \text{ Log } \hat{\Psi}_i)^2}{\sum W_{i\ell}}$$

$$= 12,76 - \frac{(12,81)^2}{23,70} = 5,84$$

Test du χ^2 à 6 ddl = non significatif

6 - Test d'homogénéité construit sur $\hat{\Psi}_{cmh}$

$$\sum \frac{(a_i - E(A_i, \hat{\Psi}_{cmh}))^2}{V(A_i, \hat{\Psi}_{cmh})}$$

On obtient $E(A_i, \hat{\Psi}_{cmh})$ et $V(A_i, \hat{\Psi}_{cmh})$ en résolvant les équations suivantes :

(on note $E(A_i, \hat{\Psi}_{cmh})$ par a_i)

$$\frac{a_i (n_{oi} - m_{1i} + a_i)}{(n_{1i} - a_i)(m_{1i} - a_i)} = \hat{\Psi}_{cmh}$$

et

$$V(A_i, \hat{\Psi}_{cmh}) = \left[\frac{1}{a_i} + \frac{1}{n_{1i} - a_i} + \frac{1}{m_{1i} - a_i} + \frac{1}{(n_{oi} - m_{1i} + a_i)}\right]^{-1}$$

Pour la 1ere strate (15 ans) : cela donne

$$a_1 (49 - 23 + a_1) = 0{,}55 (23 - a_1)(23 - a_1)$$

équation du 2ème degré qui a pour racine positive : 5,42,

et $V(A_1, \hat{\Psi}_{cmh}) = \left[\frac{1}{5{,}42} + \frac{1}{23-5{,}42} + \frac{1}{23-5{,}42} + \frac{1}{49-23+5{,}42}\right]^{-1}$

On trouve finalement ici

$$\sum \frac{(a_i - E(A_i, \hat{\Psi}_{cmh}))^2}{V(A_i, \hat{\Psi}_{cmh})} = 6{,}14$$

χ^2 à 6 ddl non significatif.

Si on veut utiliser le test de tendance linéaire :

$$\frac{\sum x_i (a_i - E(A_i, \hat{\Psi}_{cmh}))^2}{\sum x_i^2 V(A_i, \hat{\Psi}_{cmh}) - \left[\sum x_i V(A_i, \hat{\Psi}_{cmh})\right]^2 / \sum V(A_i, \hat{\Psi}_{cmh})}$$

$$= \frac{3{,}80^2}{383{,}95 - \frac{88{,}17^2}{24{,}82}} = 0{,}21$$

χ^2 à un ddl non significatif.

- Exemple 2 -

MONONUCLEOSE INFECTIEUSE ET ANTECEDENTS D'AMYGDALECTOMIE EN AJUSTANT SUR L'AGE. ETUDE CAS-TEMOINS

	Amygdalectomie oui	Amygdalectomie non	
mononucléose	a_i	b_i	n_{1i}
témoin	c_i	d_i	n_{0i}
	m_{1i}	m_{0i}	n_i

				a_i	$\hat{E}(A_i)$	$\hat{V}(A_i)$	$\hat{\psi}_i$	$a_i d_i/n_i$	$b_i c_i/n_i$	$\log \hat{\psi}_i$	$w_i e$	$w_i \log \hat{\psi}_i$	$w_i \log \hat{\psi}_i^2$	$E(A_i, \hat{\psi}_{i-1})$	$a_i - E(A_i, \hat{\psi}_{i-1})$	$V(A_i, \hat{\psi}_{i-1})$	$\frac{(a_i - E(A_i, \hat{\psi}_{i-1}))^2}{V(A_i, \hat{\psi}_{i-1})}$	x_i	$x_i \times [a_i - E(A_i, \hat{\psi}_{i-1})]$	$x_i \times V(A_i, \hat{\psi}_{i-1})$	$x_i^2 \times V(A_i, \hat{\psi}_{i-1})$
18 ans	6	17	23	6	7,35	3,40	0,66	2,67	4,01	-0,42	3,17	-1,33	0,56	5,42	0,58	3,03	0,11	1	0,58	3,03	3,03
	17	32	49																		
	23	49	72																		
19 ans	3	39	42	3	8,83	4,85	0,21	1,52	7,35	-1,56	2,43	-3,79	5,91	6,17	-3,17	4,04	2,49	2	-6,34	8,08	16,16
	26	70	96																		
	29	109	138																		
20 ans	12	29	41	12	12,33	6,31	0,95	6,12	6,44	-0,05	6,25	-0,31	0,02	8,83	3,17	5,42	1,85	3	9,51	16,26	48,78
	34	78	112																		
	46	107	153																		
21 ans	8	38	46	8	13,92	7,30	0,40	3,94	9,86	-0,92	5,46	-5,02	4,62	9,90	-1,90	6,20	0,58	4	-7.6	24,80	99,20
	48	91	139																		
	56	129	185																		
22 ans	5	10	15	5	5,64	3,12	0,81	2,74	3,38	-0,21	2,98	-0,63	0,13	3,92	1,08	2,62	0,45	5	5,40	13,10	65,50
	45	73	118																		
	50	83	133																		
23 ans	2	7	9	2	3,72	1,92	0,36	0,99	2,71	-1,02	1,42	-1,45	1,48	2,64	-0,64	1,67	0,25	6	-3,84	10,02	60,12
	29	37	66																		
	31	44	75																		
24 ans	4	5	9	4	4,29	2,00	0,87	1,86	2,14	-0,14	1,99	-0,28	0,04	3,13	0,87	1,84	0,41	7	6,09	12,88	90,16
	36	39	75																		
	40	44	84																		
	Total			40	56,08	28,90		19,84	35,80		23,70	-12,81	12,76			24,82	6,14		3,80	88,17	382,95

CHAPITRE IX - REFERENCES

METHODOLOGIE :

- ARMITAGE P.
  Statistical methods in medical research
  Blackwell, Oxford, 1977, 4e edition, chapitre 16

- BRESLOW N.E., DAY N.E.
  Statistical methods in Cancer Research
  Volume 1 - The analysis of case-control studies
  IARC Scienfific Publications n° 32
  IARC, Lyon, 1980, Chapitre 4.

- COCHRAN W.G.
  Some methods for strengthening the common χ^2 tests
  Biometrics, 1954, 10, 417-451.

- FLEISS J.L.
  Statistical methods for rates and proportions
  Wiley, New York, 1981, 2e edition, Chapitre 10.

- MANTEL N., HAENSZEL W.
  Statistical aspects of the analysis of data from retrospective studies of disease
  J. Natl. Cancer Inst. 1959, 22, 719-748.

- RUMEAU-ROUQUETTE C., BREART G., PADIEU R.
  Méthodes en épidémiologie
  Flammarion, Paris, 1981, 2è édition, chapitre 23.

- WOOLF B.
  On estimating the relationship between blood group and disease.
  Ann. Human Gen., 1955, 19, 251-253.

EXEMPLES ADAPTES DE :

- KAMINSKI M., RUMEAU-ROUQUETTE C., SCHWARTZ D.
  Consommation d'alcool chez les femmes enceintes et issue de la grossesse.
  Rev. Epidem. et Santé Publ. 1976, 24, 27-40.

- MILLER R.B.
  Combining 2 x 2 contingency tables.
  in : Biostatishes casebook, MILLER RG, EFRON B, BROWN BW, MOSES LE.
  Wiley, New York, 1980, pp 73-83.
  (exemple sur la mononucléose infectieuse).

CHAPITRE X

APPARIEMENT : TRAITEMENT STATISTIQUE SIMPLE

Lorsque la méthode de stratification utilisée dans une enquête de type exposés-témoins, ou de type cas-témoins est un appariement individuel, l'analyse relève de technique statistiques particulières exposées dans ce chapitre (pour plus de détail voir FLEISS).

1 - NOTATIONS

Les notations utilisées dans la population sont celles du Chapitre VI (voir paragraphe 6-1-1.)

Sur un échantillon de N paires par exemple de type cas-témoin, on répartit les paires en 4 groupes selon le schéma suivant

		Témoins E+	Témoins E-
Cas	E+	n_{11}	n_{10}
	E-	n_{01}	n_{00}

cas et témoins exposés : n_{11} paires
cas exposé-témoin non exposés : n_{10}
cas non exposé-témoin exposé : n_{01}
cas et témoin non exposés : n_{00}

2 - APPARIEMENT COMME CAS PARTICULIER DE STRATIFICATION

On considère chaque paire cas-témoin comme une strate d'effectif 2, cas particulier du Chapitre IX . Les résultats possibles sont l'un des quatre tableaux 2 x 2 ci-dessous.

.../...

Exposition

	+	−	+	−	+	−	+	−
Cas	1	0	1	0	0	1	0	1
Témoin	1	0	0	1	1	0	0	1
	2	0	1	1	1	1	0	2
-type de paire	(1)		(2)		(3)		(4)	
-nombre de paires de ce type	n_{11}		n_{1o}		n_{o1}		n_{oo}	

Pour chaque strate, compte tenu de son effectif, on utilise le modèle exact, décrit au Chapitre VI, paragraphe 2-1, et qui est conditionnel aux marges : on utilise la distribution conditionnelle du nombre de cas exposés (l'équivalent de a dans le Chapitre VI), les marges étant fixées.

Conditionnellement aux marges, les tableaux de type (1) ou (4), dans lesquels un total marginal est nul, ont une probabilité 1 d'être observés et ne fournissent pas d'information sur OR.

Les tableaux de type (2) ou (3) sont les deux configurations possibles avec les mêmes marges : un sujet exposé et un sujet non exposé (paires discordantes). Soit π la probabilité qu'une paire soit de type (2) : cas exposé x témoin non exposé, sachant que c'est une paire discordante.

Si p_1 et p_o sont les probabilités d'exposition respectivement chez les cas et chez les témoins, la probabilité pour qu'une paire soit de type (2) est $p_1(1-p_o)$ et de type (3) : $p_o(1-p_1)$

$$\text{donc } \pi = \frac{p_1(1-p_o)}{p_1(1-p_o) + p_o(1-p_1)}$$

$$= \psi / \psi + 1 \qquad (1)$$

Remarque : (1) est un cas particulier de la formule (3) du Chapitre VI avec $m_1 = m_o = n_1 = n_o = a = 1$.

Soit N_{10} la variable aléatoire : nombre de paires de type (2) ; la probabilité pour qu'il y ait n_{10} paires de type (2) sachant qu'il y a $n_{10} + n_{01}$ paires discordantes, est donnée par la loi binomiale :

(2) $$P(N_{10} = n_{10} / n_{10} + n_{01} \; ; \; p_1, \; p_o) = C^{n_{10}}_{n_{10}+n_{01}} \pi^{n_{10}} (1-\pi)^{n_{01}}$$

3 - TEST DE L'ASSOCIATION ENTRE FACTEUR D'EXPOSITION ET MALADIE

L'hypothèse nulle d'absence d'association peut s'exprimer par (voir Chapitre VI, paragraphe 2-1).

$$\Psi = 1$$

c'est-à-dire ici par

$$\pi = \frac{1}{2}$$

On peut tester H_o en utilisant directement les probabilités exactes de la loi binomiale :

(3) $$P(N_{10} \geq n_{10} \quad / \pi = \frac{1}{2} \;)$$

Si l'échantillon est assez grand, N_{10} suit une loi normale de moyenne :

(4) $$E(N_{10}) = \frac{1}{2} \times (n_{10} + n_{01})$$

et variance :

(5) $$V(N_{10}) = \frac{1}{2} \times \frac{1}{2} \times (n_{10} + n_{01}) = \frac{1}{4} (n_{10} + n_{01})$$

Alors :

(6) $$\frac{\left(n_{10} - \frac{n_{10}+n_{01}}{2}\right)^2}{\frac{1}{4}(n_{10} + n_{01})}$$ suit une loi du χ^2 à 1 ddl,

Il s'écrit sous la forme

(7) $$(n_{10} - n_{01})^2 / (n_{10} + n_{01})$$

C'est le test de Mc Nemar

.../...

Remarque : C'est un cas particulier du test ajusté de Mantel-Haenszel (Chapitre IX, formule (4)).

En effet, ici,

$$\sum a_i = n_{11} + n_{10} \text{ (nombre de cas exposés)}$$

$$\sum \hat{E}(A_i, \Psi = 1) = \sum \frac{m_{1i}\, n_{1i}}{n_i}$$

- ici n_{1i} vaut 1 et n_i, 2, dans toutes les strates
- m_{1i} vaut 2 pour les paires de type (1)
  1 pour les paires de type (2) ou (3)
  0 pour les paires de type (4).

donc

$$\sum \hat{E}(A_i ; \Psi = 1) = \frac{1}{2} (2\, n_{11} + n_{10} + n_{01})$$

$$\sum \hat{V}(A_i ; \Psi = 1) = \sum \frac{m_{1i}\, m_{oi}\, n_{1i}\, n_{oi}}{n_i^{\,2} (n_i - 1)}$$

(dans chaque strate l'effectif est 2, donc on prend l'estimation exacte de la variance).

$$V(A_i ; \Psi = 1) = \frac{1}{4} (n_{10} + n_{01})$$

car $m_{1i}\, m_{oi}\, n_{1i}\, n_{oi}$ est nul pour toutes les paires de type (1) ou (4).

Finalement le test de Mantel Haenszel donne

$$\frac{(n_{11} + n_{10} - n_{11} - \frac{1}{2}(n_{10} + n_{01}))^2}{\frac{1}{4}(n_{10} + n_{01})} = \frac{(n_{10} - n_{01})^2}{n_{10} + n_{01}}$$

4 - ESTIMATION DU ODDS RATIO

4-1. Estimation ponctuelle

L'estimateur du maximum de vraisemblance de Π conditionnellement aux marges, est le pourcentage observé de paires de type (2) parmi les paires discordantes :

(8)
$$\hat{\Pi}_{mv} = \frac{n_{10}}{n_{10} + n_{01}}$$

on en déduit :

$$\hat{\Psi}_{mv} = n_{10}/n_{01}$$

Remarque : $\hat{\Psi}_{mv}$, estimateur du maximum de vraisemblance conditionnellement aux marges pour des données stratifiées, dans le cas particulier des séries appariées est également l'estimateur de Mantel Haenszel (Chapitre IX, formule (12)).

$$\hat{\Psi}_{mh} = \frac{\sum a_i d_i / N_i}{\sum b_i c_i / N_i}$$

$N_i = 2, \forall i$

$a_i d_i$ non nul seulement si paire de type (2) : il y en a n_{10}

$b_i c_i$ non nul seulement si paire de type (3) : il y en a n_{01}

$$\hat{\Psi}_{mh} = \frac{n_{10}/2}{n_{01}/2} = n_{10}/n_{01} = \hat{\Psi}_{mv}$$

4-2. Intervalle de confiance pour Ψ

On établit un intervalle de confiance pour $\overline{\Pi}$, paramètre de la loi binomiale (1), et à partir de (2) on calcule l'intervalle correspondant pour Ψ.

L'intervalle de confiance pour $\overline{\Pi}$ peut être établi à partir de tables, dans le cas de petits effectifs, ou par approximation par la loi normale si les effectifs sont suffisamment grands (SCHWARTZ).

Remarque : On a vu ci-dessus comment, dans le cas de séries appariées, tester l'existence d'une association entre exposition à un facteur et maladie : le test a été construit à partir du odds ratio et peut donc être utilisé pour les études exposés-témoins comme pour les études cas-témoins. On a vu aussi comment estimer le odds ratio, estimation possible dans les deux types d'étude également. Par contre, dans le cas d'etude exposés-témoins, on pourrait souhaiter estimer le risque relatif, mais on ne dispose pas de méthode directe pour cela. On peut toutefois, si on connait P_o, estimer RR à partir de ψ graâce à la relation existant entre ces trois paramètres

(10) $$RR = \frac{\Psi}{1-P_o+\Psi P_o}$$

Par ailleurs on peut estimer d'autres mesures d'associations décrites par FLEISS.

5 - CAS PLUS COMPLIQUES

Les méthodes décrites ci-dessus s'appliquent au cas où

1) à chaque cas, on n'a apparié qu'un témoin et
2) le facteur d'exposition n'a que deux niveaux.

Les cas où :

- on a plus d'un témoin, que ce nombre de témoins soit fixe, ou variable selon le cas,
- le facteur d'exposition a plusieurs niveaux,
- on veut tenir compte de un ou plusieurs facteurs autres que celui sur lequel on a apparié, soit pour faire un ajustement, soit pour rechercher des interactions,

se traitent mieux avec des modèles appropriés (voir Chapitre XV)

Des techniques classiques sont toutefois décrites par BRESLOW et DAY.

6 - EXEMPLE

Cas : femme ayant un cancer du sein vérifié histologiquement et traitée dans une des 73 cliniques privées participantes

Témoin apparié : femme ayant une intervention chirurgicale pour une maladie bénigne dans la même clinique et de la même classe d'âge

Relation entre cancer du sein et nombre d'enfants

Les 1032 "paires" se répartissent de la manière suivante :

		Témoins	
		0 enfant	1 ou +
Cas	0 enfant	25	178
	1 ou +	102	727

1 - Test de Mc Nemar

$$\frac{(178 - 102)^2}{178 + 102} = 20,63$$

χ^2 à 1 ddl significatif à $p < 0,001$.
Les femmes ayant un cancer du sein ont significativement moins d'enfants.

2 - Estimation du odds ratio

$$\hat{\Psi} = 178/102 = 1,75$$

3 - Intervalle de confiance pour Ψ

$$\hat{\pi} = \frac{178}{178 + 102} = 0,64$$

Intervalle de confiance à 95 % pour π :

$$0,64 \pm 2 \sqrt{\frac{0,64 \times 0,36}{280}}$$

$$(0,58 \ ; \ 0,70)$$

$$\Psi = \frac{\pi}{1 - \pi}$$

D'où intervalle de confiance pour Ψ :

$$(1,38 \ ; \ 2,33)$$

CHAPITRE X - REFERENCES

- FLEISS J.L.
Statistical methods for rates andproportions
Wiley, New York, 1981, 2e edition, chapitre 8

- SCHWARTZ D.
Méthodes statistiques à l'usage des médecins et biologistes
Flammarion, Paris, 1969, 3ème édition Chapitre 4

- BRESLOW N.E., DAY N.E.
Statistical methods in cancer research
Volume 1 - The analysis of case control studies
IARC Scientific Publications n° 32
IARC, Lyon, 1980, Chapitre 5.

EXEMPLE ADAPTE DE :

- LE M.
Cancer du sein et nutrition
Gaz. Méd. de France, 1980, 87, 2857.

CHAPITRE XI

NOMBRE DE SUJETS NECESSAIRE

1 - RAPPELS

On sait (1) que pour que le test de comparaison (au risque de première espèce α) des 2 distributions binomiales de paramètres P_A et P_B ait une puissance $\pi = 1-\beta$ (β risque de seconde espèce), il est nécessaire que les effectifs, supposés égaux, de chacun des 2 groupes soient au moins égaux à

$$n = \frac{(\varepsilon_{2\alpha} + \varepsilon_{2\beta})^2}{2(\arcsin\sqrt{p_A} - \arcsin\sqrt{p_B})^2} \qquad (1)$$

où ε est donné par la table de l'écart-réduit.

Cette formule est valable dans le cas d'un test unilatéral ; dans le cas bilatéral, $\varepsilon_{2\alpha}$ doit être remplacé par ε_{α}.

Dans le cas plus général d'effectifs inégaux, les effectifs n_A et n_B doivent vérifier

$$\frac{1}{n_A} + \frac{1}{n_B} = \frac{2}{n} \qquad (2)$$

où n est donné par la formule (1). Aussi, si le rapport n_B/n_A a une valeur fixée λ

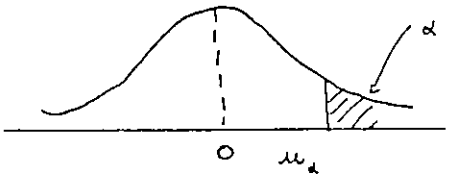
$$n_A = \frac{n}{2}\left(\frac{1+\lambda}{\lambda}\right) \quad \text{et} \quad n_B = \frac{n}{2}(1+\lambda)$$

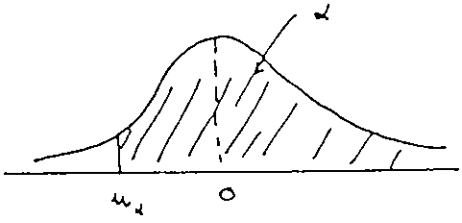
Les formules (1) et (2) peuvent être utilisées de 2 façons : soit calculer les effectifs qui assurent une puissance donnée π, soit inversement déterminer la puissance correspondant à des effectifs donnés, ce qui revient à calculer $\varepsilon_{2\beta}$ à partir des éléments connus.

.../...

L'utilisation de l'écart-réduit peut poser de petits problèmes pratiques : si $\beta > .5, 2\beta$ est supérieur à 1 ; on peut trouver des valeurs négatives pour $\varepsilon_{2\beta}, \ldots$.

Aussi vaut-il mieux introduire les quantités u, définies par $\Phi(u_\alpha) = 1 - \alpha$ où Φ est l'intégrale de la densité de la loi normale réduite (voir figures ci-dessous)





Aussi si $\alpha < \frac{1}{2}$ $\quad u_\alpha = \varepsilon_{2\alpha}$, tandis que si $\alpha > \frac{1}{2}$ $u_\alpha = -\varepsilon_{2(1-\alpha)}$

Il est facile de voir qu'en fonction de ces quantités, (1) s'écrit

$$n = \frac{(u_\alpha - u_\pi)^2}{2(\arcsin\sqrt{p_A} - \arcsin\sqrt{p_B})^2} \qquad (3)$$

si le test est bilatéral, u_α est remplacé par $u_{\alpha/2}$.

2 - AUTRES FORMULES POSSIBLES

Les formules précédentes ne sont, en toute rigueur, qu'approchées. D'autres approximations sont possibles; une, souvent utilisée, est la suivante [2]

$$n' = \frac{\left(u_\alpha\sqrt{2PQ} - u_\pi\sqrt{P_AQ_A + P_BQ_B}\right)^2}{(P_A - P_B)^2} \qquad (4)$$

où $P = \frac{P_A + P_B}{2}$, $Q_A = 1 - P_A$, ...

.../...

On pourra vérifier sur des exemples numériques que (3) et (4) conduisent à des valeurs très voisines. Dans le cas d'effectifs inégaux avec $\frac{n_A}{n_B} = \lambda$, l'équivalent de (4) est

$$n'_A = \frac{\left(u_\alpha \sqrt{(1+1/\lambda)\bar{P}\bar{Q}} - u_\pi \sqrt{P_A Q_A/\lambda + P_B Q_B}\right)^2}{(P_A - P_B)^2}$$

où $\bar{P} = \frac{P_B + \lambda P_A}{1 + \lambda}$

On peut montrer que les formules (3) et (4) conduisent à des effectifs légèrement trop faibles pour assurer la puissance requise. Une formule plus exacte, mais plus compliquée a été proposée par [3].
Cependant, les formules pour calculer des nombres de sujets nécessaires ne sont destinées qu'à donner des ordres de grandeur. On peut donc utiliser sans problèmes les formules 3 ou 4.

3 - APPLICATION AUX ENQUETES EPIDEMIOLOGIQUES

On distinguera le cas de l'enquête "Cas-témoins" de celui de l'enquête "Cohorte".

3-1. Enquête cas-témoins

Les P_A et P_B des formules sont alors p_o et p_1, pourcentages d'exposés chez les témoins et les malades. Raisonnons sur un exemple :

- supposons que l'on connaisse, au moins approximativement, le pourcentage d'exposés chez les témoins, soit $p_o = .30$. Combien faut-il de sujets par groupe si l'on veut avoir une probabilité $\pi = .9$ de détecter un risque relatif $R = 2$, le test (unilatéral) étant effectué avec un risque de première espèce $\alpha = .05$?

.../...

On sait que dans une telle enquête on ne peut calculer que le odds-ratio

$$\Psi = \frac{P_1}{1-P_1} \Big/ \frac{P_o}{1-P_o} = \frac{P_1}{1-P_1} \Big/ \frac{P_o}{1-P_1}$$

De cette relation on tire

$$P_1 = \frac{\Psi P_o}{1+P_o(\Psi-1)}$$

si la maladie est rare, on peut approcher Ψ par R, donc

$$P_1 = \frac{2 \times .30}{1 + .30} = .462$$

La formule (3) donne donc $n = \dfrac{(1,645 + 1.282)^2}{2(\text{arc sin } \sqrt{.462} - \text{arc sin}\sqrt{.30})^2}$

$= 152$

3-2. Enquête cohorte

On peut distinguer 2 situations.

3-2-1. Les sujets sont suivis pendant une même durée T.

Les P_A et P_B sont alors les probabilités de faire la maladie pendant la période du temps T.

Supposons connue la probabilité de la maladie P_o chez les non-exposés (ou dans la population générale, si l'exposition est rare), et qu'on veuille détecter un risque relatif R.

Alors $P_1 = R.P_o$ et $n = \dfrac{(u_\alpha . u_\pi)^2}{2(\text{arc sin}\sqrt{R p_o} - \text{arc sin}\sqrt{p_o})^2}$

Dans le cas, très fréquent en pratique, où P_o est faible, cette formule se réduit à

$$n = \frac{(u_\alpha . u_\pi)^2}{2 p_o [\sqrt{R}-1]^2} \qquad (5)$$

Exemple numérique : $\alpha = .05$; $\pi = .90$; $P_o = 2\ \%$; $R = 3$.
on trouve $n = 400$.

.../...

3-2-2. Les sujets sont suivis pendant une durée variable

Si i_o est l'incidence de la maladie chez les non-exposés pendant l'unité de temps, disons l'année, l'incidence en k ans sera, a très peu près, ki_o. Si on veut déceler un risque relatif r et si les sujets sont suivis un an, il faut des groupes au moins égaux à

$$n = \frac{(u_\alpha . u_\beta)^2}{2 i_o [\sqrt{r}-1]^2} \qquad (6)$$

Si les sujets sont suivis 2 ans, le dénominateur est double, et il faut 2 fois moins de sujets. Ainsi la formule (6) doit-elle être interprétée comme le nombre de personnes x années nécessaire : n sujets suivis 1 an, n/2 suivis 2 ans, ou n' sujets suivis de façon différente, mais donnant au total n personnes x années.

4 - INTERET DES EFFECTIFS INEGAUX

Pour une discussion générale, on pourra se reporter à [1].

4-1. Enquête cas-témoins

Dans d'assez nombreuses situations, on est amené à prendre plusieurs témoins pour un même malade. La motivation n'en est presque jamais une diminution du nombre des malades pour arriver à la même puissance, mais plutôt une variété des témoins qui peut rendre plus crédibles les différences constatées.

4-2. Enquête cohorte

Soit une population dont le pourcentage d'exposés est $p = 1 - q$. Plutôt que suivre un groupe d'exposés et un groupe de non-exposés, définis a priori, on peut décider de suivre un échantillon de cette population.

Le rapport des effectifs des 2 groupes (non exposés et exposés) est $\lambda = \frac{q}{p}$.

.../...

Le nombre de non-exposés nécessaire est :

$n_o = \frac{n}{2}(1 + \lambda) = \frac{n}{2} \cdot \frac{1}{p}$, celui des exposé est :

$n_1 = \frac{n}{2}\frac{1+\lambda}{\lambda} = \frac{n}{2q}$ soit un total de $N = n_o + n_1 = \frac{n}{2pq}$

n étant donné par les formules (5) ou (6).

5 - CALCUL D'UNE PUISSANCE

Ce calcul peut être utile pour interpréter des résultats négatifs d'études basées sur des effectifs réduits, comme c'est souvent le cas d'enquêtes "Cas-témoins" publiées dans la littérature médicale.

Exemple 1 : Quelle est la puissance d'une enquête Cas-Témoins basée sur 2 x 50 sujets lorsque la fréquence de l'exposition chez les témoins (ou la population générale) est $p_o = .30$ et que le risque relatif R est 2.

Alors $p_1 = .4615$ (en admettant que $R = \Psi$)

$$50 = \frac{(u_\alpha - u_\pi)^2}{2 \times (.1672)^2} \longrightarrow u_\alpha - u_\pi = 1.672$$

comme $u_\alpha = 1{,}645$ (test unilatéral), $u_\pi = -.027$ et $\pi = .49$

On a plus d'une chance sur deux de ne pas mettre en évidence la différence ; mais l'enquête n'est pas très lourde.

Exemple 2 : Sans une population de 1000 sujets dont 100 sont exposés et 900 non-exposés, quelle est la probabilité de détecter un risque relatif R de 2, lorsque $P_o = .05$ (en 5 ans)

$$\frac{1}{100} + \frac{1}{900} = \frac{4 \times .05 \times (\sqrt{2}-1)^2}{(u_\alpha - u_\pi)^2} \longrightarrow u_\alpha - u_\pi = 1{,}757$$

$$u_\pi = -.112 \quad \text{et} \quad \pi = .45$$

Peut-être vaut-il mieux ne pas se lancer dans une enquête aussi lourde pour des résultats aussi incertains.

.../...

6 - PLUS PETIT RISQUE RELATIF (OU ODDS-RATIO) DETECTABLE

Une autre façon d'évaluer la puissance potentielle d'une enquête est de calculer le plus petit risque relatif (ou odds-ratio) qu'elle est capable de détecter avec une puissance donnée

Exemple 1 : 50 malades sont comparés à 50 témoins. On estime à $p_o = .25$ H le pourcentage d'exposés chez les témoins (ou dans la population générale). On décide de faire un test bilatéral au risque de première espèce $\alpha = .05$. Combien doit valoir p_1, pourcentage d'exposés chez les malades, pour que la puissance du test soit $\pi = .80$?

On utilise toujours la formule (1), mais cette fois, n, α, β et p_o sont connus et p_1 est inconnu.

l'équation $50 = \dfrac{(1,960 + 0.842)^2}{2(\text{arc sin } \sqrt{p_1} - .52360)^2}$ donne

$$\left|\text{arc sin } \sqrt{p_1} - .52360\right| = .2802$$

soit arc sin $\sqrt{p_1} = .8038$ $\quad p_1 = .5184$

arc sin $\sqrt{p_1} = .2434$ $\quad p_1 = .0581$

La première valeur conduit à un odds-ratio de

$\dfrac{.5184}{.4816} \Big/ \dfrac{.25}{.75} = 3.23$ et la seconde à un odds ratio de

$\dfrac{.0581}{.9419} \Big/ \dfrac{.25}{.75} = .19$

Si on peut faire l'approximation $\Psi = R$, on peut dire que seuls des risques relatifs supérieurs à 3.2. ou inférieurs à .2 ont une probabilité au moins égale à .8 d'être détectés.

Exemple 2 : $n = 50$; $\alpha = .05$ (bilatéral) ; $\pi = 90$; $p_o = .90$

l'équation $50 = \dfrac{(1,960 + 1,645)^2}{2(\text{arc sin } \sqrt{p_1} - 12490)^2}$ donne

$$\left|\text{arc sin } \sqrt{p_1} - 1.2490\right| = .3605$$

soit arc sin $\sqrt{p_1} = 1.6095$ ou arc sin $\sqrt{p_1} = .8885$

Or il n'existe aucun p tel que arc sin $\sqrt{p_1}$ = 1.6095, la valeur maximum étant arc sin 1 = $\frac{\pi}{2}$ = 1.57. Ainsi, quelle que soit la valeur p_1 ($> p_o$), l'enquête ne peut avoir la puissance désirée.

On peut inversement trouver une limite inférieure à arc sin $\sqrt{p_1}$ qui soit négative. Ceci indique que quel que soit p_1 ($<p_o$), l'enquête ne peut avoir la puissance désirée.

7 - ANALYSE SEQUENTIELLE

On a mis au point [4] des procédés séquentiels qui peuvent être sommairement décrits de la façon suivante :

Comparer k malades (ou k exposés) à k témoins (ou k non-exposés) ; si la différence est significative, conclure ; sinon prendre 2 nouveaux sous-groupes de k malades et k témoins ; comparer les 2k malades aux 2k témoins ; si la différence est significative, conclure ; sinon prendre 2 nouveaux sous-groupes etc ...

On se fixe généralement le nombre total K d'étapes. En fonction des exigences (α et π), l'effectif k est calculé. Il faut par ailleurs remarquer que les limites de signification à chaque étape ne sont pas les limites habituelles, pour tenir compte de la multiplication des tests effectués.

Ces procédés ne semblent pas avoir été utilisés dans des enquêtes réelles.

8 - NOMBRE DE SUJETS NECESSAIRE DANS LE CAS DE SERIES APPARIEES

Raisonnons sur l'exemple d'une enquête cas-témoins.

On rappelle que le test est basé sur les seules paires discordantes (-+)=,(malades non exposés, témoins exposés) en nombre n_{01} et (+-) (malades exposés, témoins non exposés) en nombre n_{10} ($n_{10} + n_{01} = N$) et qu'il revient à comparer à 1/2 le pourcentage $\frac{n_{10}}{N}$ (ou $\frac{n_{01}}{N}$) si P est le pourcentage théorique de paires (+-), le test aura une puissance égale à π, si le nombre nécessaire de paires discordantes est donné par l'une des 2 formules :

.../...

$$N = \frac{(u_{\alpha/2} - u_{\pi})^2}{4(\arcsin\sqrt{P} - \arcsin\sqrt{.5})^2} \qquad (7)$$

$$N = \frac{\left(\frac{u_{\alpha/2}}{2} - u\sqrt{PQ}\right)^2}{(P - 1/2)^2} \qquad (8)$$

La formule (7) est l'analogue de la formule (1) et la formule (8) est l'analogue de la formule (4). La formule (7) est plus volontiers utilisée.

Par ailleurs on sait $P = \frac{\Psi}{1+\Psi}$ ($\sim \frac{R}{1+R}$ si la maladie est rare).

Exemple : Combien de paires discordantes sont nécessaires pour détecter un risque relatif de 2 avec une puissance π = .9.

$$P = \frac{2}{1+2} = \frac{2}{3}$$

La formule (7) donne $N = \frac{(1.960 + 1.282)^2}{4(.95532 - .78540)^2} = 91$

tandis que (8) donne $N = \frac{\left[\frac{1.960}{2} + 1.282\sqrt{\frac{2}{3} \times \frac{1}{3}}\right]^2}{\left(\frac{2}{3} - \frac{1}{2}\right)^2} = 90$

Les valeurs trouvées sont à peu près égales.

Il est cependant important de se souvenir que les formules donnent le nombre nécessaire de paires discordantes.

Si p_d désigne le pourcentage de ces paires parmi toutes les paires possibles, alors le nombre total des paires nécessaires doit être $M = \frac{N}{p_d}$.

.../...

p_d dépend de la corrélation entre les expositions des 2 éléments de la paire, malade et témoin, si on fait l'hypothese qu'il y a indépendance entre ces expositions $p_d = p_0 q_1 + p_1 q_0$, où comme d'habitude, p_0 et p_1 désignent les pourcentages d'exposés chez les témoins et les malades.

En fait cette hypothèse d'indépendance est peu plausible et on doit s'attendre à une corrélation positive puisque les 2 éléments de la paire sont appariés sur des critères vraisemblablement liés à l'exposition. Ceci entraîne que le p_d réel est plus grand que celui calculé sous l'hypothèse d'indépendance.

La formule $M = \dfrac{N}{p_0 q_1 + p_1 q_0}$ apparaît donc comme "conservative" dans la mesure où elle conduit à un nombre trop élev de paires.

Dans l'exemple précédent, supposons qu'on ait l'estimation $p_0 = .30$. Alors $p_1 = .462$ (valeur calculée plus haut) et $p_d = .30 \times .54 + .46 \times .70 = .484$.

Ainsi M = 186 paires

On peut évidemment utiliser (7) ou (8) pour calculer en fonction de N la puissance du test.

CHAPITRE XI - REFERENCES

1 - D. SCHWARTZ, R. FLAMANT, J. LELLOUCH
L'essai thérapeutique chez l'homme
1970, Flammarion, Paris

2 - J.L. FLEISS
Statistical methods for rates and proportions
1973, J. Wiley, New York

3 - J.T. CASAGRANDE, M.C. PIKE
An improved approximate formula for calculating sample sizes for comparing two binomial populations
1978, Biometrics, 34, 483-486.

4 - B.S. PASTERNACK, R.E. SHORE
Group sequential methods for cohort and case control studies
1980, J. Chron Dis, 33, 365-373.

CHAPITRE XII

STRATIFICATION OU AJUSTEMENT ?

1 - INTRODUCTION

Le Chapitre IX a montré comment au moment de l'analyse des résultats, il est possible de tenir compte d'un facteur de confusion C et présenté les techniques statistiques appropriées (ajustement). Cependant comme indiqué au Chapitre VIII on peut essayer d'en tenir compte au moment de la planification même de l'étude (stratification). Se pose alors le problème de la comparaison de ces 2 attitudes, du double point de vue du biais dans l'estimation de la mesure de la relation facteur de risque x maladie et de la précision de cette estimation.

Nous distinguerons comme d'habitude les situations où la mesure de la relation est exprimée par le risque relatif (habituellement études de cohorte) et celles où elle est exprimée par le odds-ratio (enquêtes cas-témoins).

2 - ENQUETES COHORTE

2-1. Le modèle

Nous rappelons la formule

$$R_p = R \frac{R'p_1 + (1-p_1)}{R'p_2 + (1-p_2)} \qquad (1)$$

.../...

en renvoyant au Chapitre VIII pour la définition des symbôles qui y figurent.

Pour tenir compte de C on peut comme on l'a dit, soit ajuster (ou comme on dit parfois "post-stratifier"), soit stratifier, c'est-à-dire faire en sorte que la distribution de C soit la même chez les exposés et les non exposés ($p_1 = p_2 = p$). Alors $R_p = R$, quel que soit R'. Le biais sur le risque relatif est supprimé par la stratification.

Ainsi du point de vue du biais sur R_p, l'ajustement et la stratification sont équivalents. Toutefois, après stratification on a, au moment de l'analyse, encore 2 options : ne pas tenir compt du fait qu'il existe 2 strates, ou au contraire en tenir compte.

Au total 3 stratégies sont possibles : Ajustement (A) ; Stratification sans ajustement (S) ; Stratification suivie d'un ajustement (SA) ; nous désignerons de plus par le symbôle (o), la stratégie consistant à ne pas tenir compte de C.

Ces attitudes vont être comparées du point de vue de la précision de l'estimation de R.

Les N sujets étudiés se distribuent en fonction des classes de C et E selon les valeurs théoriques données par le tableau 1. Le tableau 2 donne les incidences théoriques pour chacune des combinaisons de C et E (sous le modèle de non-interaction).

	E^+	E^-
C^+	P_{p1}	$(1-P)p_2$
C^-	$P(1-p_1)$	$(1-P)(1-p_2)$
	P	(1-P)

Tableau 1 : Distribution de C et E

	E^+	E^-
C^+	RR'i	R'i
C^-	Ri	i

Tableau 2 : Incidences

2-2. Ajustement

Ajuster revient à estimer R séparément chez C^+ ($\hat{R}^+$) et C^- ($\hat{R}^-$) puis à calculer une estimation commune $\hat{R}$.

.../...

Un procédé habituel pour obtenir $\hat{R}$ est de l'estimer par la moyenne de $\hat{R}^+$ et $\hat{R}^-$ pondérés par l'inverse de leurs variances. Toutefois, ainsi qu'il a été dit au Chapitre VI, il est préférable de travailler avec les logarithmes des risques relatifs qu'avec les risques relatifs eux-mêmes.

On rappelle que si $R = \frac{P_1}{P_o}$ (P_1 et P_o étant 2 incidences calculées respectivement sur n_1 et n_o sujets), $\log R = \log P_1 - \log P_o$;

$$\text{var}\left[\log (R)\right] = \text{var} \log P_1 + \text{var} \log P_0 = \frac{\text{var } P_1}{P_1^{\,2}} + \frac{\text{var } P_o}{P_o^{\,2}} \quad \frac{Q_1}{n_1 P_1} + \frac{Q_o}{n_o P_o} \qquad (2)$$

Ainsi la variance de $\log \hat{R}^+$ est (au facteur $\frac{1}{N}$ près)

$$V_A^+ = \frac{1 - R'i}{R'i(1-P)p_2} + \frac{1 - RR'i}{RR'i\ Pp_1}$$

de même la variance de $\log \hat{R}^-$ est

$$V_A^- = \frac{1 - i}{i(1-P)(1-p_2)} + \frac{1 - Ri}{Ri\ P(1-p_1)}$$

On sait que l'estimateur combiné a pour variance V_A telle que $\frac{1}{V_A} = \frac{1}{V_A^+} + \frac{1}{V_A^-}$

soit $$V_A = \frac{V_A^+ V_A^-}{V_A^+ + V_A^-} \qquad (3)$$

2-3. Stratification

On stratifie généralement sur la distribution de C chez les exposés ; c'est-à-dire que l'on choisit $p_2 = p_1$ et le tableau 1 devient :

.../...

	E^+	E^-	
C^+	Pp_1	$(1-P)p_1$	p_1
C^-	$P(1-p_1)$	$(1-P)(1-p_1)$	$1-p_1$
	P	$1-P$	

L'incidence de la maladie chez l'ensemble des exposés (en proportion P) est

$$I^+ = \frac{RR'iPp_1 + Ri\,P(1-p_1)}{P} = Ri\left[R'\,p_1 + (1-p_1)\right] = Ri\rho$$

on a posé $\rho = R'p_1 + (1-p_1)$.

L'incidence de la maladie chez l'ensemble des non exposés (en proportion 1-P) est de même

$$I^- = \frac{R'i(1-P)p_1 + i(1-P)(1-p_1)}{1-P} = i\rho$$

(on vérifie bien que $R = \frac{I^+}{I^-}$: la stratification élimine le bia

Ainsi V_S (variance de logarithme de l'estimation de $\hat{R} = \frac{I^+}{I^-}$) est d'après (2)

$$V_S = \frac{1-Ri\rho}{Ri\rho\;P} + \frac{1-i\rho}{i\rho(1-P)} = \frac{1 + P(R-1) - Ri\rho}{Ri\rho\;P(1-P)} = \frac{A - Ri\rho}{Ri\rho\;P(1-P)}$$

(on a posé $A = 1 + P(R-1)$

2-4. Stratification suivie d'un ajustement

La variance de l'estimateur chez C^+ est

$$V_{SA}^+ = \frac{1 - RR'i}{RR'iPp_1} + \frac{1 - R'i}{R'i(1-P)p_1} = \frac{A - RR'i}{RR'\;p_1\;P(1-P)}$$

de même chez C^-

$$= \frac{A - Ri}{Ri(1-p_1)P(1-P)}$$

.../...

La variance de l'estimateur combiné est

$$V_{SA} = \frac{V_{SA}^{+} \; V_{SA}^{-}}{V_{SA}^{+} + V_{SA}^{-}} = \frac{(A - R_i)(A - RR'_i)}{R_i \; P(1-P)(A_\rho - RR'_i)} \qquad (5)$$

2-5. Comparaison de V_A , V_S, V_{SA}

a) On peut montrer que $V_S > V_{SA}$

En effet $\frac{V_S}{V_{SA}} = \frac{A-R_i\rho}{\rho} \Big/ \frac{(A-R_i)(A-RR'_i)}{A_\rho - RR'_i}$

donc $V_S - V_{SA}$ a le signe de $(A-R_i\rho)(A_\rho-RR'_i)- \rho(A-R_i)(A-RR'_i)$

On trouve après des calculs algébriques élémentaires que V_S-V_{SA} a le signe de $AR_i\, p_1(1-p_1)(R'-1)^2$. Il est donc toujours positif.

Ainsi il vaut mieux, du point de vue de la précision des estimateurs, ajuster même après une stratification.

Ceci est l'équivalent de ce qui se passe avec une "méthode des blocs" en analyse de la variance, où bien qu'il y ait équilibre, il vaut mieux, pour augmenter la puissance de l'expérience, tenir compte de l'effet "bloc".

La différence entre V_S et V_{SA} dépend des 5 paramètres R, R', i, p_1 et P. En particulier elle est d'autant plus grande que R' est différent de 1. Mais elle n'est réellement importante que si la relation facteur de confusion x maladie est forte, voire très forte (R' grand).

b) Comparaison de V_{SA} et V_A

Il n'y a aucun résultat général, ainsi que le montrent les exemples numériques ci-après.

	E^+	E^-
C^+	.125	.375
C^-	.375	.125
	.5	.5

Distribution de E et C

.30	.15
.02	.01

R = 2

R' = 15

Incidences

Ici $p_1 = Pr\left[C^+/E^+\right] = .25$ et $p_2 = Pr\left[C^+/E^-\right] = .75$

La formule (1) donne $R_p = 2 \times \dfrac{15 \times 1/4 + 3/4}{15 \times 3/4 + 1/4} = 2 . \dfrac{18}{46} = .78$

Le biais introduit par C est énorme puisqu'il renverse même le sens de la liaison ; nous avons une illustration du paradoxe de SIMPSON : existence d'une relation positive chez C^+ et C^- séparément, mais négative au total.

$$\left.\begin{aligned} V_A^+ &= \frac{.85}{.15 \times .375} + \frac{.70}{.30 \times .125} = 33,78 \\ V_A^- &= \frac{.99}{.01 \times .125} + \frac{.98}{.02 \times .375} = 922,67 \end{aligned}\right\} \quad V_A = 32.6$$

$$\left.\begin{aligned} V_{SA}^+ &= \frac{.70}{.30 \times .125} + \frac{.85}{.15 \times 125} = 64 \\ V_{SA}^- &= \frac{.98}{.02 \times .375} + \frac{.99}{.01 \times .375} = 394,67 \end{aligned}\right\} \quad V_{SA} = 55,07$$

Dans cet exemple $V_A < V_{SA}$

Exemple 2 : même tableau des incidences, mais distribution de C et E donnée par la table ci-dessous

	E^+	E^-
C^+	.375	.125
C^-	.125	.375

On vérifiera aisément que $V_A = 47.79 > V_{SA} = 20.96$

Dans chacun des 2 exemples, les différences entre V_A et V_{SA} apparaissent importantes ; mais à la fois $|p_1 - p_2|$ et R' sont très grands. Dans des situations plus réalistes, l'écart entre V_A et V_{SA} est moins important.

.../...

2-6. Le facteur C n'est pas facteur de confusion

Alors soit $p_1 = p_2$ soit $R' = 1$.

2-6-1. $p_1 = p_2$ (indépendance entre C et E)

Les observations sont de fait stratifiées. On a vu en 2-5 a) que $V_{SA} < V_S$ (sauf si $R' = 1$, auquel cas il y a égalité), c'est-à-dire qu'il vaut théoriquement mieux ajuster. Ce résultat conduirait à tenir compte au moment de l'analyse de tous les facteurs liés à la maladie mais pas à l'exposition. Cependant le gain est minime sauf si R' est grand. Ainsi, dans l'étude d'un facteur étiologique E du cancer du poumon on a intérêt à ajuster sur la consommation de tabac, même si E n'est pas lié à cette consommation.

2-6-2. $R' = 1$

La stratégie la plus simple consiste évidemment à ne pas tenir compte de C.

	E^+	E^-
C^+	Pp_1	$(1-P)p_2$
C^-	$P(1-p_1)$	$(1-P)(1-p_2)$

Distribution de C et E

	E^+	E^-
C^+	Ri	i
C^-	Ri	i

Incidences

La variance de log R, estimé sur l'ensemble de la population est

$$V_0 = \frac{1-i}{i(1-P)} + \frac{1-Ri}{RiP} = \frac{1}{i}\left[\frac{1-i}{1-P} + \frac{1-Ri}{RP}\right] = C + D$$

où l'on a posé $C = \frac{1-i}{i(1-P)}$ et $D = \frac{1-Ri}{RPi}$

Par ajustement on obtient une estimation dont la variance est donnée par (3) où l'on à fait $R' = 1$.

on a ainsi $V_A^- = \frac{C}{p_2} + \frac{D}{p_1}$

de même $V_A^+ = \frac{C}{1-p_2} + \frac{D}{1-p_1}$

.../...

Quelques manipulations algébriques élementaires montrent que l'on a $V_A \geq V_o$ (avec égalité si $p_1 = p_2$).

On peut vérifier ce résultat général sur l'exemple particulier ci-dessous.

Exemple :

	E^+	E^-
C^+	.30	.10
C^-	.30	.30

Distribution de C et E

	E^+	E^-
C^+	.20	.10
C^-	.20	.10

Incidences

$$V_O = \frac{.80}{.20 \times .60} + \frac{.90}{.10 \times 40} = 29,17$$

$$V_A^- = \frac{.80}{.20 \times .30} + \frac{.90}{.10 \times .30} = 43,33$$

$$V_A^+ = \frac{.80}{.20 \times .30} + \frac{.90}{.10 \times .10} = 103,33$$

$$\Big\} \quad V_A = 30,53$$

Ainsi, si R' = 1 (C non lié à la maladie, mais lié à E), l'ajustement est une procédure qui fait perdre de la puissance.

2-7. Conclusions

La décision de stratifier au lieu de simplement ajuster doit résulter de la constatation que V_S est assez nettement inférieure à V_A. Si on stratifie, on doit en plus ajuster si R' est assez grand.

3 - ENQUETES CAS-TEMOINS

3-1. Introduction

On rappelle la formule

$$\Psi_P = \frac{\Psi p_1 + (1-p_1)}{\Psi' p_2 + (1-p_2)} \qquad (6)$$

où les Ψ sont des odds-ratio.

ψ' mesure la relation (conditionnelle) facteur de confusion x maladie chez les exposés (ou les non exposés) ; p_1 et p_2 sont les pourcentages de C^+ chez les témoins exposés et non exposés.

Comme indiqué au Chapitre VIII, la seule stratification possible dans une enquête cas-témoins consiste à rendre M et C marginalement indépendants (même distribution de C chez les malades et les témoins), mais ceci n'assure pas que $\psi' = 1$ (qui est une indépendance conditionnelle).

En conséquence la <u>stratification n'élimine pas le biais</u> sur ψ . On sait seulement qu'elle biaise vers 1 le odds-ratio, c'est-à-dire que

si $\psi < 1$ $\quad \psi < \psi_p < 1$ $\quad$ et $\quad$ si $\psi > 1$ $\quad 1 < \psi_p < \psi$

Elle tend donc à masquer la relation facteur de risque x maladie.

Aussi, toute stratification doit-elle être nécessairement <u>suivie d'un ajustement</u>. On est alors en droit de se demander quel est l'intérêt d'une stratification (qui sera suivie d'un ajustement) par rapport à un ajustement seul. La réponse à cette question résulte de la comparaison des variances V_{SA} et V_A des estimateurs de ψ sous les 2 stratégies.

3-2. <u>Le modèle</u>

Soit P le pourcentage d'exposés chez les témoins ; p_1 et p_2 étant toujours les pourcentages de C^+ chez ces mêmes témoins, exposés et non exposés.

On a ainsi la distribution suivante ches les témoins

$C^+ E^+$	$C^+ E^-$	$C^- E^+$	$C^- E^-$
Pp_1	$(1-P)p_2$	$P(1-p_1)$	$(1-P)(1-p_2)$

Il est facile de calculer (le faire !) cette même distribution chez les malades en fonction de P, p_1, p_2 et ψ , ψ' . On trouve au total (tableau 3) :

.../...

	E^+	E^-	Total
M^+	$\Psi\Psi' p p_1 / \Sigma$	$\Psi'(1-P)p_2/\Sigma$	K^+
M^-	$P p_1$	$(1-P)p_2$	K^-
	C^+		

	E^+	E^-	Total
M^+	$\Psi P(1-p_1)/\Sigma$	$(1-P(1-p_2)/\Sigma$	$1-K^+$
M^-	$P(1-p_1)$	$(1-P)(1-p_2)$	$1-K^-$
	C^-		

Tables 3

où l'on a posé $\Sigma = \Psi\Psi' P p_1 + \Psi'(1-P)p_2 + \Psi P(1-p_1) + (1-P)(1-p_2)$

3-3. Ajustement

Il revient à estimer Ψ (ou mieux $\log \Psi$) séparément chez C^+ et C^- puis à calculer une moyenne pondérée par l'inverse de variances (voir 2-2).

On sait par ailleurs (Chapitre VI) que var $\log \Psi$ est estimée par la somme des inverses des effectifs des 4 cases.

Ainsi $V_A = \dfrac{V_A^+ \, V_A^-}{V_A^+ + V_A^-}$ avec $V_A^+ = \sum \dfrac{1}{\text{cases de } C^+}$

et $V_A^- = \sum \dfrac{1}{\text{cases de } C^-}$

3-4. Stratification

Elle consiste à faire en sorte que la distribution de C chez les témoins soit la même que chez les malades (c'est-à-dire un pourcentage K^+ de C^+ et un pourcentage $1 - K^+$ de C^-). La stratification conduit ainsi à la distribution théorique ci-dessous (Tables 4)

$\Psi\Psi' P p_1 / \Sigma$	$\Psi'(1-P)p_2/\Sigma$	K^+
$P p_1 \dfrac{K^+}{K^-}$	$(1-P)p_2 \dfrac{K^+}{K^-}$	K^+
C^+		

$\Psi P(1-p_1)/\Sigma$	$(1-P)(1-p_2)/\Sigma$	$1-K^+$
$P(1-p_1)\dfrac{1-K^+}{1-K^-}$	$(1-P)(1-p_2)\dfrac{1-K^+}{1-K^-}$	$1-K^+$
C^-		

Tables 4

Conclusion : le schéma stratifié (plus un ajustement) est plus puissant que le schéma non stratifié (avec ajustement).

Exemple 2 : $\psi = 5$; $\psi' = 1$ (C n'est pas facteur de confusion ; $P = .5$; $p_1 = .9$; $p_2 = .1$

On vérifiera que cette fois la conclusion est inversée

$V_A = 27,47 < V_{SA} = 35,78$

La table en annexe empruntée à [1] donne le rapport (x 100) $\frac{V_{SA}}{V_A}$ pour différentes combinaisons des paramètres. Le schéma stratifié est meilleur si ce rapport est inférieur à 100. La différence ne semble importante que pour ψ' grand ou p_1 très différent de p_2.

3-6. C n'est pas un facteur de confusion

La plus simple des stratégies possibles est évidemment de ne pas tenir compte de C ; ceci conduit à une estimation de log ψ dont la variance est notée V_o.

Trois stratégies sont alors à comparer : ne pas tenir compte de C, stratifier puis ajuster, ajuster.

Nous distinguerons les 2 situations où C n'est pas facteur de confusion.

3-6-1. $\psi' = 1$

On remarque que même dans ce cas, stratifier (sans ajuster) conduit à un biais, comme on le voit sur l'exemple suivant (les nombres sont des effectifs)

	E^+	E^-	
M^+	80	10	90
M^-	5	5	10

C^+

$\psi = 8$

	E^+	E^-	
M^+	16	6	22
M^-	1	3	4

C^-

$\psi = 8$

	E^+	E^-	
M^+	96	16	112
M^-	6	8	14

Total

$\psi_T = 8$

.../...

Il faut ajuster. La variance de l'estimateur combiné est

$$V_{SA} = \frac{V_{SA}^{+} \cdot V_{SA}^{-}}{V_{SA}^{+} + V_{SA}^{-}}$$ avec pour V_{SA}^{+} et V_{SA}^{-} la valeur $\sum \frac{1}{\text{cases}}$

3-5. Comparaison de V_A et V_{SA}

Il n'existe pas de résultat général ainsi que le montrent les exemples ci-dessous.

Exemple 1 : $\Psi = 2$; $\Psi' = 5$; $P = .1$; $p_1 = .4$; $p_2 = .2$

Les tables 3 sont

	E^+	E^-	
M^+	$\frac{.40}{2.14}$	$\frac{.90}{2.14}$	$\frac{1.30}{2.14}$
M^-	.04	.18	.22

C^+

$V_A^+ = 38,28$

	E^+	E^-	
	$\frac{.12}{2.14}$	$\frac{.72}{2.14}$	$\frac{.84}{2.14}$
	.06	.72	.78

C^-

$V_A^- = 38;86$

$$V_A = 19,28$$

Les tables 4 sont

	E^+	E^-	
M^+	$\frac{.40}{2.14}$	$\frac{.90}{2.14}$	$\frac{1.30}{2.14}$
M^-	$.04 \times \frac{1.30}{2.14} \times \frac{1}{.22}$	$.18 \times \frac{1.30}{2.14} \times \frac{1}{.22}$	$\frac{1.30}{2.14}$

C^+

$V_{SA}^+ = 18,79$

	E^+	E^-	
	$\frac{.12}{2.14}$	$\frac{.72}{2.14}$	$\frac{.84}{2.14}$
	$.06 \times \frac{.84}{2.14} \times \frac{1}{.78}$	$.72 \times \frac{.84}{2.14} \times \frac{1}{.78}$	$\frac{.84}{2.14}$

C^-

$V_{SA}^- = 56,68$

$$V_{SA} = 14,11$$

La stratification conduirait aux effectifs

80	10	90	16	6	22	96	16	
45	45	90	5,5	16,5	22	50,5	61,5	
C^+			C^-			Total		

$\psi_{MH} = 8$ $\qquad$ $\psi = 8$ $\qquad$ $\psi = 7,3 < 8$

En ce qui concerne les variances des divers estimateurs :

a) V_A et V_{SA} ne sont pas en relation simple : on lit en effet dans la table en annexe que pour $\psi' = 1$, il existe des valeurs des paramètres pour lesquelles $V_A > V_{SA}$ et d'autres pour lesquelles $V_A < V_{SA}$. Ce résultat n'est toutefois pas très intéressant eu égard aux résultats b) et c).

b) on peut démontrer que $V_A > V_o$. On perd de la puissance en ajustant.

c) de très nombreux exemples numériques suggèrent fortement que l'on a aussi $V_{SA} > V_o$.

La conclusion importante est qu'on perd en puissance en tenant compte d'une façon ou d'une autre d'un facteur qui n'est pas lié à la maladie (c'est un cas particulier de ce qu'on appelle le "sur-ajustement").

3-6-2. $p_1 = p_2 = p$

Dans cette situation et contrairement à la précédente, la stratification ne crée pas de biais comme le montre la démonstration ci-après :

Le Σ qui figure dans les tables 3) vaut

$$\psi\psi' Pp + \psi'(1-P)p + \psi P(1-p) + (1-P)(1-p) = \rho\rho'$$

avec $\rho = \psi P + (1-P)$ et $\rho' = \psi' p + (1-p)$

.../...

Les tables (3) sont

$\Psi\Psi' \frac{Pp}{pp'}$	$\Psi' \frac{(1-P)p}{pp'}$	$\Psi' \frac{p}{p'}$
Pp	$(1-P)p$	p

C^+

$\Psi \frac{P(1-p)}{pp'}$	$\frac{(1-P)(1-p)}{pp'}$	$\frac{1-p}{p'}$
$P(1-p)$	$(1-P)(1-p)$	$1-p$

C^-

qui se totalisent en la table (5)

$\Psi \frac{P}{p}$	$\frac{1-P}{p}$
P	$1-P$

On construira aisément les tables (4), qui résultent de la stratification et on vérifiera qu'elles se somment aussi selon la table 5.

La case (M^- E^+) par exemple est $Pp \frac{\Psi'}{p'} + \frac{P(1-p)}{p'} = P$ etc ...

Donc stratifier (sans ajuster) revient très exactement à ne pas tenir compte de C.

Des trois stratégies possibles : ne pas tenir compte de C, ajuster, stratifier puis ajuster, on peut montrer que $V_o = V_{SA} < V_A$

On a intérêt à ne pas tenir compte de C.

3-6-3. Conclusion (si C n'est pas facteur de confusion)

Il y a perte de puissance à en tenir compte.

3-7. Conclusion générale

1) ne stratifier que dans les situations où V_{SA} est nettement inférieur à V_A ;

2) ajuster obligatoirement après une stratification.

CHAPITRE XII - REFERENCES

[1] - P.G. SMITH, N.G. DAY
Matching and confounding in the design and analysis of epidemiological case control studies.

Table D1. Relative efficiency of an unmatched to a matched design, in both cases with a stratified analysis, when the extra variable is a positive confounder. The body of the table shows the values of $100 \times V_{MS}/V_S$.

				$R_E = 2$				$R_E = 5$				$R_E = 10$			
				R_C				R_C				R_C			
P	p_1	p_2	R_{CE}	1	2	5	10	1	2	5	10	1	2	5	10
0.1	0.5	0.5	1.0	100	97	87	79	100	98	88	79	100	98	89	80
	0.6	0.4	2.3	99	91	78	69	98	90	78	70	98	91	79	71
	0.4	0.2	2.7	99	90	73	60	98	90	74	62	98	90	76	64
	0.8	0.6	2.7	99	92	84	80	98	91	84	79	97	91	84	79
	0.7	0.3	5.4	98	84	69	60	93	82	68	61	92	82	71	63
	0.9	0.1	81.0	89	69	51	43	79	66	53	47	80	71	62	57
0.3	0.5	0.5	1.0	100	97	87	79	100	97	87	80	100	97	88	81
	0.6	0.4	2.3	100	95	84	77	101	97	88	80	102	100	91	84
	0.4	0.2	2.7	100	96	83	71	101	99	88	76	102	102	92	82
	0.8	0.6	2.7	100	96	89	85	100	97	91	87	102	100	94	90
	0.7	0.3	5.4	99	93	82	74	102	99	89	81	107	106	96	88
	0.9	0.1	81.0	96	88	76	69	108	105	94	85	130	129	115	102
0.5	0.5	0.5	1.0	100	97	88	80	100	97	89	83	100	98	91	87
	0.6	0.4	2.3	100	99	90	82	101	100	93	87	102	101	96	91
	0.4	0.2	2.7	101	100	90	79	101	102	94	84	102	102	97	89
	0.8	0.6	2.7	100	99	93	89	101	100	96	93	102	101	98	95
	0.7	0.3	5.4	101	100	92	84	105	106	99	92	107	108	102	96
	0.9	0.1	81.0	106	107	98	90	129	133	121	108	152	157	139	122

Table D1. (*Continued*) . . .

				$R_E = 2$				$R_E = 5$				$R_E = 10$			
				R_C				R_C				R_C			
P	p_1	p_2	R_{CE}	1	2	5	10	1	2	5	10	1	2	5	10
0.7	0.5	0.5	1.0	100	97	88	82	100	98	91	87	100	99	94	91
	0.6	0.4	2.3	101	101	93	87	101	101	96	91	101	101	97	94
	0.4	0.2	2.7	101	102	95	84	101	102	97	89	101	102	98	93
	0.8	0.6	2.7	101	101	96	93	101	101	98	96	101	101	99	97
	0.7	0.3	5.4	102	105	99	92	104	106	102	96	103	105	102	98
	0.9	0.1	81.0	112	122	117	107	131	141	133	120	138	146	137	124
0.9	0.5	0.5	1.0	100	97	89	83	100	98	93	89	100	99	96	94
	0.6	0.4	2.3	100	102	96	90	100	101	97	94	100	100	98	96
	0.4	0.2	2.7	100	103	98	88	100	101	97	92	100	101	98	95
	0.8	0.6	2.7	100	102	98	95	100	101	99	97	100	101	99	98
	0.7	0.3	5.4	101	106	103	97	101	104	101	98	101	102	101	99
	0.9	0.1	81.0	109	125	128	120	112	122	122	116	109	115	115	111

See footnote to Table A.

Table D2. Relative efficiency of an unmatched to a matched design, in both cases with a stratified analysis, when the extra variable is a negative confounder. The body of the table shows values of V_{SRS}/V_S.

				$R_E = 0.1$				$R_E = 0.2$				$R_E = 0.5$			
				R_C				R_C				R_C			
P	p_1	p_2	R_{CE}	1	2	5	10	1	2	5	10	1	2	5	10
0.1	0.5	0.5	1.0	100	99	96	94	100	98	93	89	100	97	89	83
	0.6	0.4	2.3	100	98	93	90	100	96	89	84	100	94	83	76
	0.4	0.2	2.7	100	97	89	82	100	95	83	74	100	93	77	64
	0.8	0.6	2.7	100	98	96	95	100	97	93	91	100	95	89	86
	0.7	0.3	5.4	101	97	90	86	101	94	84	78	101	90	76	68
	0.9	0.1	81.0	109	98	79	67	112	95	71	57	109	86	61	48
0.3	0.5	0.5	1.0	100	99	94	91	100	98	91	87	100	97	88	82
	0.6	0.4	2.3	101	97	91	87	101	96	87	82	101	95	84	77
	0.4	0.2	2.7	101	97	86	77	101	96	82	71	101	95	80	68
	0.8	0.6	2.7	101	98	95	93	101	97	92	90	101	96	90	86
	0.7	0.3	5.4	103	98	88	82	104	96	84	77	102	93	80	72
	0.9	0.1	81.0	138	118	88	71	131	110	81	66	112	94	73	62
0.5	0.5	0.5	1.0	100	98	91	87	100	97	89	83	100	97	88	80
	0.6	0.4	2.3	102	97	89	83	101	97	87	80	100	96	86	79
	0.4	0.2	2.7	102	98	81	73	101	97	83	71	100	97	85	72
	0.8	0.6	2.7	102	98	91	91	101	98	92	89	101	97	91	87
	0.7	0.3	5.4	107	100	87	80	105	98	85	78	101	96	85	77
	0.9	0.1	81.0	152	131	99	81	129	113	90	76	106	98	84	75

Table D2. (*Continued*) . . .

				$\psi = R_E = 0.1$				$R_E = 0.2$				$R_E = 0.5$			
				$\psi' = R_C$				R_C				R_C			
P	P_1	P_2	R_{CE}	1	2	5	10	1	2	5	10	1	2	5	10
0.7	0.5	0.5	1.0	100	97	88	81	100	97	87	80	100	97	87	79
	0.6	0.4	2.3	102	[illegible]	88	81	101	98	88	80	100	99	90	82
	0.4	0.2	2.7	102	99	86	74	100	99	88	76	100	100	91	80
	0.8	0.6	2.7	102	99	93	89	101	98	92	89	100	99	93	89
	0.7	0.3	5.4	107	102	90	81	102	99	90	82	99	100	92	85
	0.9	0.1	81.0	130	122	103	90	108	106	96	87	96	100	96	90
0.9	0.5	0.5	1.0	100	98	89	80	100	98	88	79	100	97	87	79
	0.6	0.4	2.3	98	101	96	88	98	102	97	89	99	103	97	88
	0.4	0.2	2.7	97	102	102	94	98	104	104	95	99	104	102	91
	0.8	0.6	2.7	98	100	97	92	98	101	98	93	99	102	98	93
	0.7	0.3	5.4	92	100	101	95	93	103	104	97	98	107	105	98
	0.9	0.1	81.0	80	91	100	100	79	94	105	105	89	107	118	113

See footnote to Table A.

CHAPITRE XIII

LE MODELE LOGISTIQUE : I - ENQUETES "COHORTE"

1 - INTRODUCTION

Les chapitres précédents ont uniquement considéré le cas d'un facteur de risque dichotomique (présent ou absent), étudié soit isolément (Chapitre VI), soit en présence d'un facteur de confusion qualitatif à plusieurs classes (Chapitre IX). Cependant dans de très nombreuses situations, on a à étudier simultanément de nombreux facteurs (de risque aussi bien que de confusion) de nature quelconque (qualitative ou quantitative). Ceci ne peut se faire de façon efficace qu'en ayant recours à un modèle dont on pense qu'il décrit de façon convenable les faits observés. Le modèle logistique qui va être maintenant présenté est, pour un certain nombre de raisons qui apparaitront clairement dans la suite, extrèmement utilisé.

2 - LE MODELE LOGISTIQUE DANS L'ENQUETE COHORTE

Soit X un facteur de risque quantitatif (par exemple la pression artérielle, le taux de cholestérol ...) ou une fonction (par exemple le logarithme) de ce facteur. Appelons P(x) la probabilité qu'un sujet chez qui X = x développe la maladie au cours d'un délai T fixé, le même pour tous les sujets (c'est la situation décrite au Chapitre IV, paragraphe 1.1). Dans le cas fréquent où T est variable d'un sujet à l'autre on doit utiliser les techniques introduites plus loin au Chapitre XVI.

Le modèle logistique stipule que

$$P(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} \qquad (1)$$

Si X croit de $-\infty$ à $+\infty$, P(x) croit de 0 à 1 selon une courbe en S.

Par définition on appelle <u>logit</u> d'un pourcentage P, la quantité logit $P = \log \frac{P}{1-P}$. Si P = 0, le logit est $-\infty$, si P = 5, il vaut 0 ; si P = 1, le logit est $+\infty$.

Le modèle logistique est donc : logit P(x) = $\alpha x + \beta$. Le logit de l'incidence est une fonction linéaire de la valeur du facteur de risque.

Le choix du modèle logistique conduit logiquement à mesurer la relation facteur-maladie par le coefficient β.

Si X est une variable dichotomique (absence, présence) on peut lui attribuer des valeurs 0 ou 1 et le mçdèle se réduit à :

$$P_0 = \frac{e^{\alpha}}{1+e^{\alpha}} \quad , \quad P_1 = \frac{e^{\alpha+\beta}}{1+e^{\alpha+\beta}} \quad , \text{ d'où} \quad e^{\beta} = \frac{P_1}{1-P_1} \Big/ \frac{P_0}{1-P_0} = \psi$$

Ainsi l'utilisation du modèle logistique implique que la relation facteur-maladie soit mesurée par le "odds-ratio" (et non par le risque relatif) même dans le cas d'une enquête cohorte. Une conséquence particulièrement importante de ce point, est que ainsi que nous verrons au chapitre suivant, le modèle logistique peut être également utilisé dans les enquêtes cas-témoins, lesquelles comme nous le savons permettent d'estimer ψ . Par ailleurs, si l'incidence de la maladie est faible $\psi \simeq RR$.

3 - <u>CAS D'UN FACTEUR QUALITATIF X A PLUSIEURS CLASSES</u>

Appelons $x_0, x_1, \ldots x_k$ les classes en nombre k + 1. S'il y a un ordre sur les classes il est quelquefois possible de leur attribuer une valeur numérique transformant le facteur qualitatif en un facteur quantitatif. Si tel n'est pas le cas, ou si on ne

peut raisonnablement assigner des valeurs numériques aux classes, on peut remplacer la variable X par k variables $(X_1, .., X_i, .., X_k)$ qui prennent les valeurs 0 ou 1 selon que $X_1 \neq x_i$ ou $X_i = x_i$ (si $X = x_0$, les k variables X_i prennent la valeur 0).

On est ainsi ramené au modèle logistique multivariable qui sera décrit plus loin.

4 - ANALYSE STATISTIQUE

Différents problèmes peuvent se poser :

a) y-a-t-il une relation entre le facteur et la maladie ? ceci revient à tester l'hypothèse nulle $H_o : \beta = 0$.

b) mesurer cette relation, c'est-à-dire estimer β.

c) éventuellement, tester l'adéquation du modèle aux observations.

Ces observations sont constituées de l'ensemble des N (N = nombre de sujets étudiés) paires (x_i, y_i) où x_i est la valeur de X du sujet i et y_i est la variable (malade ou non malade) du même sujet.

La vraisemblance de la paire $(x_i\ y_i)$ est :

$$p_i = \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}} \qquad \text{si le sujet est malade}$$

ou $$q_i = \frac{1}{1+e^{\alpha+\beta x_i}} \qquad \text{s'il est non malade.}$$

Il est alors facile de voir que la vraisemblance de l'ensemble des observations est

$$V = \frac{\tilde{\pi}\left(e^{\alpha+\beta x_i}\right)}{\tilde{\pi}\left(1+e^{\alpha+\beta x_i}\right)} \qquad \text{où n désigne le nombre des malades} \qquad (2)$$

son logarithme est $$L = n\alpha + \beta \tilde{\Sigma} x_i - \sum^{N} \log\left(1+e^{\alpha+\beta x_i}\right) \qquad (3)$$

4-1. Estimation des paramètres inconnus α et β

Sous certaines hypothèses assez restrictives on peut estimer α et β par les techniques de discrimination linéaire.

(Cette méthode est présentée et discutée à la fin du chapitre suivant).

Autrement on peut estimer α et β par la méthode habituelle du maximum de vraisemblance : les estimations $\hat{\alpha}$ et $\hat{\beta}$ sont les valeurs qui rendent maximum (2) ou de façon équivalente (3), ou encore qui annulent les dérivées de L par rapport à α et β, c'est-à-dire sont solution du système

$$\frac{\partial L}{\partial \alpha} = n - \sum^{N} \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}} = 0$$

$$\frac{\partial L}{\partial \beta} = \sum^{n} x_i - \sum^{N} \frac{x_i e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}} = 0 \qquad (4)$$

On remarque qua la première équation signifie que le nombre total attendu de malades sous le modèle est égal au nombre observé.

De façon générale le système (4) n'a pas de solution explicite. On ne peut le résoudre que sur ordinateur.

On sait que les estimateurs du maximum de vraisemblance ont de "bonnes propriétés puisque lorsque les effectifs sont grands (tendent vers l'infini)

a) ils convergent vers les valeurs des paramètres inconnus

b) leur distribution tend vers la distribution normale.

Par ailleurs on peut montrer que leur matrice de variances covariances (asymptotique) est I^{-1} où I est l'opposée de la matrice des dérivées secondes de L par rapport aux paramètres inconnus.

Ces notions théoriques sont explicitées par l'exemple très simple ci-dessous.

Un exemple : X est dichotomique de sorte que les données se présentent sous la forme de la table 2 × 2 classique

	X=1	X=0	
M^+	a	b	n
M^-	c	d	m
	N_1	N_0	N

(3) est simplement

$$L = n\alpha + \beta a - N_1 \log_e(1+e^{\alpha+\beta}) - N_0 \log_e(1+e^{\alpha})$$

de sorte que le système (4) est

$$\frac{\partial L}{\partial \alpha} = n - \frac{N_1 e^{\alpha+\beta}}{1+e^{\alpha+\beta}} - \frac{N_0 e^{\alpha}}{1+e^{\alpha}} = 0$$

$$\frac{\partial L}{\partial \beta} = a - \frac{N_1 e^{\alpha+\beta}}{1+e^{\alpha+\beta}} = 0 \qquad (4')$$

La résolution de ce système est immédiate. On trouve les résultats, évidents a priori

$$\hat{\alpha} = \log \frac{b}{d}$$

$$\hat{\beta} = \log \frac{ad}{bc} = \log \hat{\psi}$$

On vérifiera après quelques calculs algébriques élementaires que :

$$I = -\begin{vmatrix} \frac{\partial^2 L}{\partial \alpha^2} & \frac{\partial^2 L}{\partial \alpha \partial \beta} \\ \frac{\partial^2 L}{\partial \alpha \partial \beta} & \frac{\partial^2 L}{\partial \beta^2} \end{vmatrix} = \begin{vmatrix} N_0 \frac{e^{\alpha}}{(1+e^{\alpha})^2} + N_1 \frac{e^{\alpha+\beta}}{(1+e^{\alpha+\beta})^2} & \frac{N_1 e^{\alpha+\beta}}{(1+e^{\alpha+\beta})^2} \\ N_1 \frac{e^{\alpha+\beta}}{(1+e^{\alpha+\beta})^2} & N_1 \frac{e^{\alpha+\beta}}{(1+e^{\alpha+\beta})^2} \end{vmatrix}$$

dont l'inverse est :

$$I^{-1} = \begin{vmatrix} \frac{1}{N_0}\frac{(1+e^{\alpha})^2}{e^{\alpha}} & -\frac{1}{N_0}\frac{(1+e^{\alpha})^2}{e^{\alpha}} \\ -\frac{1}{N_0}\frac{(1+e^{\alpha})^2}{e^{\alpha}} & \frac{1}{N_0}\frac{(1+e^{\alpha})^2}{e^{\alpha}} + \frac{1}{N_1}\frac{(1+e^{\alpha+\beta})^2}{e^{\alpha+\beta}} \end{vmatrix} \qquad (5)$$

Cette matrice est la matrice de variances-covariances de $(\hat{\alpha}, \hat{\beta})$. Pour en avoir une estimation, il suffit d'y remplacer e^{α} et e^{β} par leurs estimations $\frac{b}{d}$ et $\frac{ad}{bc}$. Tous calculs faits on trouve

$$\hat{I}^{-1} = \begin{vmatrix} \frac{1}{b}+\frac{1}{d} & -\left(\frac{1}{b}+\frac{1}{d}\right) \\ -\left(\frac{1}{b}+\frac{1}{d}\right) & \frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d} \end{vmatrix}$$

Ce résultat est bien celui qu'on attendait : en particulier l'estimation de la variance de $\hat{\beta} = \log\psi$ est $1/a+1/b+1/c+1/d$ comme prouvé directement au Chapitre VI.

4-2. Test du coefficient β

Plusieurs tests sont disponibles

4-2-1. Test exact

Considérons le logarithme de la vraisemblance (3). Elle contient les paramètres inconnus α et β, et des quantités qui sont aléatoires (n : nombre de malades et $\sum x_i$: somme des x des sujets malades). La théorie [1] indique, que le "meilleur" test (uniformément le plus puissant non biaisé) de H_o : $\beta = 0$ rejette H_o si $\sum x_i$ est soit trop grand, soit trop petit, les limites de rejet étant obtenues à partir de la distribution de x_i, conditionnelle au fait que nombre de malades est n.

Ce résultat est assez intuitif : s'il y a une liaison positive entre facteur et maladie, les malades correspondront plutôt à des valeurs élevées de x_i, et à des valeurs plutôt basses si la liaison est négative.

La distribution exacte de $\sum x_i$ conditionnelle à n est assez compliquée : on peut par contre calculer assez facilement sa moyenne μ et sa variance σ^2 sous H_o. On trouve :

$$\mu = \frac{n}{N}\sum x = n\mu_1 \quad (\mu_1 \text{ moyenne des N valeurs } x_i) \qquad (6)$$

$$\sigma^2 = \frac{n(N-n)}{N-1}\left[\sum x_o^2 - \frac{\left(\sum x\right)^2}{N}\right] = \frac{n.(N-n)}{N-1}\sigma_x^2 \quad (\sigma_x^2 \text{ variance des N valeurs } x_i).$$

Comme $\sum x_i$ a une distribution qui tend vers la normale si N tend vers l'infini, une forme approchée du test exact revient à considérer que la qualité :

$$\frac{\sum x_o - n\mu_1}{\sqrt{\sigma^2}}$$

a sous H_o une distribution normale réduite.

Exemple : X est dichotomique

$\sum x_i$ est la somme des x des n malades. C'est donc a.

Le test exact est donc basé sur la distribution de a conditionnellement à n. C'est le test de Fisher comme indiqué au Chapitre VI.

Comme X a une distribution de Bernouilli de paramètres $\frac{N_1}{N}$

$\frac{N_1}{N'}$ $\frac{N_1}{n}$ $p_1 = \frac{N_1}{N'}$ $\sigma^2_x = \frac{N_1}{N} \cdot \frac{N_0}{N}$

la moyenne de a sous H_o est $\frac{n}{N}.N_1$ et sa variance est $\sigma^2 = \frac{n.m}{N-1}.\frac{N_1}{N}\frac{N_0}{N}$

ce qui conduit à la statistique

$$\chi^2 = \frac{\left(a - \frac{n}{N}N_1\right)^2 \left[N^2(N-1)\right]}{nm\,N_1\,N_0} = \frac{(O-E)^2\,N^2(N-1)}{nm\,N_1\,N_0} \qquad (7)$$

On peut facilement vérifier que le χ^2 habituel $\sum \frac{(O-E)^2}{E}$ s'écrit $\frac{(O-E)^2\,N^3}{nm.N_1\,N_0}$ de sorte que (7) est le χ^2 habituel au facteur $\frac{N-1}{N}$ près.

4-2-2. Test basé sur la distribution de β

D'après les propriétés énoncées plus haut des estimateurs du maximum de vraisemblance (normalité asymptotique) la quantité $\frac{\hat{\beta}}{s_{\hat{\beta}}}$ suit, sous H_o, une loi normale réduite.

Dans l'exemple de la table 2 x 2

$\hat{\beta} = \log \frac{ad}{bc}$ et $\hat{s}^2_{\beta} = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$

Donc sous H_o

$$\frac{\log \frac{ad}{bc}}{\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}}$$ suit une loi normale réduite.

Une forme alternative/a estimer $\hat{s}^2_{\beta}$ sous l'hypothèse nulle $H_o : \beta = 0$ consiste

Le terme correspondant de la matrice (5) est sous H_o

$$\frac{1}{N_0} + \frac{1}{N_1} \frac{(1 + e^{\alpha})^2}{e^{\alpha}}$$

Comme sous H_o la vraisemblance $L = N\alpha - n \log (1+e^{\alpha})$ est maximum pour $e^{\hat{\alpha}} = \frac{n}{m}$ $(1+e^{\alpha} = \frac{N}{m})$

l'estimation de $s^2_{\hat{\beta}}$ est $(\frac{1}{N_0} + \frac{1}{N_1}) \frac{N^2}{m^2} \frac{m}{n} = \frac{N^3}{nm\, N_1\, N_0}$ ce qui conduit à la statistique

$$\chi^2 = \frac{(\log \frac{ad}{bc})^2 . N3}{N_1\, N_0\, n\, m}$$

Le lecteur pourra essayer de voir pourquoi $(O-E)^2$ n'est pas très différent de $(\log \frac{ad}{bc})^2$.

4-2-3. Test du rapport des vraisemblances et score-test

Nous en donnons une présentation très intuitive dans le cas simplifié où il n'y a qu'un seul paramètre inconnu β. Le graphe du logarithme de la vraisemblance comme une fonction de β a l'allure représentée sur la figure 1.

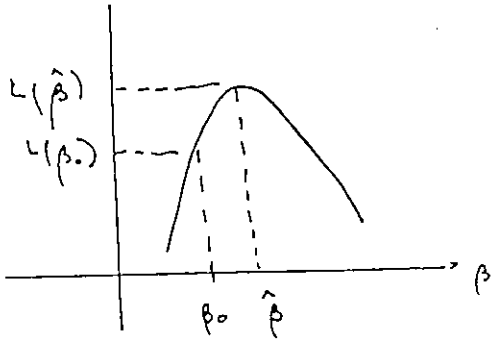


Figure 1.

Par définition l'estimateur $\hat{\beta}$ du maximum de vraisemblance vérifie $L(\hat{\beta})$ maximum et $L'(\hat{\beta})=0$

Supposons qu'on veuille tester l'hypothèse nulle $H_o : \beta = \beta_0$.

Si H_o est vraie, $\hat{\beta}$ n'est pas "trop" loin de β_o ce qui se traduit de différentes façons (i) l'écart $\hat{\beta}-\beta_o$ n'est pas "trop" grand (ii) $L(\hat{\beta})$ n'est pas "très" différent de $L(\beta_o)$ (iii) $L'(\beta_o)$ pente de $L(\beta)$ en β_o n'est pas "trop" différente de 0.

L'approche (i) est celle qui a été décrite en 4-2-2.

(test basé sur $\frac{\hat{\beta}-\beta_0}{s_{\hat{\beta}}}$)

L'approche (ii) conduit au test dit G^2 du rapport des maximum de vraisemblance. Il s'enonce rigoureusement de la façon suivante : soit un modèle (I) considéré comme exact et un sous modèle (II), cas particulier du modèle I.
L et L_0 sont les maximum des logarithmes de la vraisemblance sous I et II . Alors sous l'hypothèse H_o que le sous modèle II est vrai la quantité $2 (L - L_0)$ suit un χ^2 dont le nombre de degrés de liberté est la différence entre les nombres de paramètres des modèles (I) ou (II), ou de façon équivalente le nombre des paramètres spécifiés par le modèle (II).

Dans le cas particulier du modèle logistique le modèle (I) est logit $P(x) = \alpha + \beta x$; le modèle (II) est logit $p(x) = \alpha$. Il y a 2 paramètres dans le premier modèle, 1 dans le second ; le χ^2 aura 1 d.d.l qui est aussi le nombre de paramètres dont la valeur est spécifiée par (II) ($\beta = 0$).

L'approche (iii) conduit à ce qui est appelé le "score test". On peut le décrire de la façon ci-après :
soit L_D le vecteur des dérivées premières du logarithme de la vraisemblance par rapport aux paramètres inconnus du modèle (I) et I la matrice des dérivées secondes de L.

Calculons la quantité $L'_D I^{-1} L_D$, dans laquelle les paramètres inconnus sont remplacés par leurs valeurs, fixées ou estimées, sous le modèle (II).

Alors sous H_o, cette quantité suit un χ^2 qui a le même nombre de d.d.l que le test du rapport des maximum de vraisemblance.

Ces résultats théoriques sont explicités toujours sous l'exemple de la table 2 x 2.

. Sous le modèle (I) $L = m\alpha + \beta a - N_1 \log(1+e^{\alpha+\beta}) - N_0 \log(1+e^{\alpha})$

qui atteint son maximum pour $\hat{\alpha} = \log \frac{b}{d}$ et $\hat{\beta} = \log \frac{ad}{bc}$ Ce maximum vaut

$$L = (a+b)\log\frac{b}{d} + a\log\frac{ad}{bc} - (a+c)\log\frac{N_1}{c} - (b+d)\log\frac{N_0}{d}$$

. Sous le modèle (II) $L = m\alpha + N\log(1+e^{\alpha})$

qui est maximum pour $\hat{\alpha} = \log\frac{n}{m}$

Ce maximum vaut

$$L_0 = (a+b)\log\frac{n}{m} - (a+b+c+d)\log\frac{N}{m}$$

Pour calculer $L - L_0$, cherchons par exemple le coefficient de a. Il vaut

$$\log \frac{b}{d} + \log \frac{ad}{bc} - \log \frac{N_1}{C} - \log \frac{n}{m} + \log \frac{N}{m} =$$

$$\log \frac{b}{d} \times \frac{ad}{bc} \times \frac{c}{N_1} \times \frac{m}{n} \times \frac{N}{m} = \log \frac{aN}{nN_1}$$

$$= \log \frac{a}{\frac{nN_1}{N}} = \log \frac{O}{E} \quad \text{où O désigne l'effectif observé (a) et E l'effectif calculé sous } H_o \left(\frac{n\,N_1}{N}\right)$$

On aboutit ainsi à la statistique

$$\boxed{G^2 = 2 \sum O \log \frac{O}{E}}$$

Venons en maintenant au score test

L_D est le vecteur $\begin{cases} h - \dfrac{N_1 e^{\alpha+\beta}}{1+e^{\alpha+\beta}} - \dfrac{N_0 e^{\alpha}}{1+e^{\alpha+\beta}} \\ \omega - N_1 \dfrac{e^{\alpha+\beta}}{1+e^{\alpha+\beta}} \end{cases}$

sous le modèle (II), $\beta=0$ et l'estimation de β est $\hat{\alpha} = \log \frac{n}{m}$.

L'estimation de L_D sous H_o est ainsi :

$$\begin{cases} h - N\frac{n}{N} = 0 \\ \omega - N_1 \frac{n}{N} = O - E \end{cases}$$

Pour calculer I^{-1} il suffit de remplacer α et β dans la matrice (5) par $\hat{\alpha}$ et O. Seul le dernier terme de la diagnole est utile. Il vaut $\left(\frac{1}{N_0} + \frac{1}{N_1}\right) \frac{N^2}{m^2} \frac{m}{n} = \frac{N^3}{N_0 N_1 mn}$

$$\chi^2 = \begin{pmatrix} o, & O-E, \end{pmatrix} \times \begin{vmatrix} \times & \times \\ \times & \frac{N^3}{N_0 N_1 mn} \end{vmatrix} \times \begin{vmatrix} o \\ O-E \end{vmatrix} = \frac{(O-E)^2\, N^3}{N_0 N_1\, mn}$$

qui est très exactement le χ^2 habituel $\sum \frac{(O-E)^2}{E}$.

4-2-4. Discussion

Nous avons donné des méthodes très générales de constitution de tests d'hypothèses. Par ailleurs pour une même hypothèse H_o, il peut exister plusieurs tests possibles. A priori ces tests sont, du point de vue de leurs qualités, équivalents. Mais il se peut que pour des problèmes particuliers, l'un soit plus facile à obtenir que les autres.

4-3. Test du modèle

Il s'agit de savoir si les observations sont compatibles avec le modèle logistique(1). Un test possible est le suivant :
- on découpe la variable X en un certain nombre c de classes. Dans la classe j(j=1, .. c) sont observés les nombres n_j et m_j de malades et de non malades ($n_j + m_j = N_j$) qu'on compare aux nombres attendus. Ces nombres attendus s'obtiennent très simplement : si un sujet i de la classe j à x_i pour valeur de X, sa probabilité d'être malade est si le modèle est vrai

$$\pi_i = \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}$$

estimée par $\hat{p}_i$, calculée en remplaçant α et β par leurs estimations $\hat{\alpha}$ et $\hat{\beta}$. Le nombre attendu des malades est donc $\sum \hat{p}_i$, où la somme porte sur tous les sujets de la classe j, et celui des non-malades est de même $\sum \hat{q}_i$.

La statistique de test est comme d'habitude $\chi^2 = \sum \frac{(O-E)^2}{E}$ (ou $G^2 = 2 \sum O \log \frac{O}{E}$) qui sous l'hypothèse H_o que le modèle est valide a une distribution du χ^2 avec c - 2 d.d.l (nombre de classes - nombre de paramètres estimés). Un exemple numérique est donné plus loin dans le cas multivariate.

Ce test est assez peu puissant : il est cependant utile de l'effectuer pour détecter de gros écarts au modèle. D'autres tests ont été proposés : ils reviennent à tester le modèle logistique comme sous-modèle d'un modèle plus compliqué ; mais il peut y avoir des difficultés dans la spécification de ce modèle plus large.

5 - PLUSIEURS FACTEURS DE RISQUE - ANALYSE MULTIVARIATE

5-1. Le modèle

Ce qui précéde s'étend immédiatement au cas de plusieurs facteurs de risque X_1, X_2, ... X_k. Le modèle correspondant est

$$\text{logit } P(x_1, \ldots x_k) = \beta_\bullet + \beta_1 x_1 \ldots + \beta_k x_k \qquad (S)$$

qui conduit à une vraisemblance dont le logarithme est

$$L = n\beta_\bullet + \beta_i \tilde{\Sigma} x_1 + \ldots + \beta_k \tilde{\Sigma} x_k - \tilde{\Sigma} \log (1 + e^{\beta_\bullet + \beta_1 x_1 \ldots + \beta_k x_k})$$

Les β sont estimés par la méthode du maximum de vraisemblance (ou éventuellement par analyse discriminante). Les estimations sont solution du système

$$\frac{\partial L}{\partial \beta_\bullet} = \bullet \qquad \frac{\partial L}{\partial \beta_1} = \bullet \ , \ \ldots \ , \ \frac{\partial L}{\partial \beta_k} = 0$$

La matrice des variances covariances des k + 1 estimations $\beta_\bullet$, β_i est encore l'opposée de l'inverse de la matrice des dérivées secondes de L par rapport aux $\beta_\bullet$ et β_i ; elles permet en particulier d'obtenir des intervalles de confiance pour les paramètres inconnus.

Quant aux tests sur les β, ils peuvent être construits soit à partir des expressions $\frac{\tilde{\beta}_i}{s_{\beta_i}}$ soit comme tests du rapport des maximum de vraisemblance, soit comme score-tests.

5-2. Un exemple

Il est emprunté à l'enquête de Framingham [2] . 742 hommes âgés de 40-45 ans ont été suivis 12 ans. 88 ont développé une cardiopathie ischémique. Les variables considérées ont été : l'âge (années) le cholestérol (mg/dl), la pression artérielle (nm Hg), le poids relatif ($100 \times \frac{\text{poids}}{\text{poids idéal}}$), l'hémoglobine (g/dl), la consommation quotidienne de cigarettes (0 = non fumeurs, 1 = < 20 cig ; 2 = 20 cig ; 3 = > 20 cig), l'E.C.G. (0 = normal, 1 = anormal).

Les résultats de l'analyse ont été

	β	s_β	$t = \frac{\beta}{s_\beta}$	Coefficients standardisés
	-13.2573	-	-	-
Age	.1216	.0437	2,78	.3376
Cholestérol	.0070	.0025	2,80	.3034
Pression artérielle	.0068	.0060	1,13	.1320
Poids relatif	.0257	.0091	2,82	.3452
Hémoglobine	- .0010	.0098	0,10	-.0012
Cigarettes	.4223	.1031	4,10	.4952
E.G.G.	.7206	.4009	1,80	.1750

5-3. Importance relative des variables. Coefficients standardisés

On peut se poser le problème de l'importance relative des variables.

Il est évident que la comparaison des β n'a aucun sens puisque leur valeur dépend des unités avec lesquelles sont exprimés les x_i. Une façon de tourner cette difficulté est de les exprimer avec comme unités les écarts types des distributions des variables correspondantes.

Si à x_i correspond β_i, à $\frac{x_i}{\sigma_{x_i}}$ correspond $\beta_i \sigma_{x_i}$ (puisque le produit $\beta_i x_i$ doit être invariant). Ces coefficients standardisés sont très souvent utilisés, mais il faut bien comprendre qu'aucune conclusion biologique, ou même opérationnelle, ne peut être tirée du seul examen de ces coefficients.

5-4. Test du modèle

On opère comme indiqué en 4-3, mais avec l'adaptation suivante : les classes sont constituées à partir de la variable $\beta_1 x_1 + \ldots + \beta_k x_k$

Le nombre de classes est habituellement fixé autour de la dizaine, les limites de classe étant telles que les nombres de sujets par classe soient à peu près les mêmes.

Les données numériques relatives à l'exemple qui précède sont :

Deciles de $\hat{p}_i$	1	2	3	4	5	6	7	8	9	10	Total
O(malades)	1	4	3	8	5	9	11	6	14	27	88
E(malades)	2,1	3,3	4,2	5,1	6,2	7,4	8,9	11,5	15,2	24,1	88
O(non-malades	73	-									654
E(non-malades)	71,9										654

On trouve $\chi^2 = 7,84$ avec 8 d.d.l.
l'ajustement paraît très bon.

6 - DISCUSSION ET INTERPRETATION DU MODELE LOGISTIQUE

6-1. Facteurs de risque ou facteur de confusion

Le modèle (8) étudie l'effet de X_i, les autres facteurs $X_1 \ldots, X_{i-1}, X_{i+1} \ldots X_k$ étant contrôlés.

Ces autres X peuvent donc être aussi bien des facteurs de risque que des facteurs de confusion (auxquels on ne s'intéresse pas directement, mais dont on veut tenir compte). Le modèle (8) permet de traiter de façon symétrique - sur le plan des calculs - les différentes catégories de variables qui interviennent dans le problème étudié.

6-2. Interaction

Tel qu'il est écrit le modèle (8) est un modèle sans interaction. Pour nous en convaincre, supposons qu'il ne contient que 2 variables X_1 et X_2 toutes 2 dichotomiques (absence = 0 ; présence = 1).

$$\frac{p_{11}}{q_{11}} = e^{\beta_0+\beta_1+\beta_2} \qquad \frac{p_{10}}{q_{10}} = e^{\beta_0+\beta_1} \qquad \frac{p_{01}}{q_{01}} = e^{\beta_0+\beta_2} \qquad \frac{p_{00}}{q_{00}} = e^{\beta_0}$$

On a $\dfrac{p_{11}}{q_{11}} \Big/ \dfrac{p_{00}}{q_{00}} = \dfrac{p_{10}}{q_{10}} \Big/ \dfrac{p_{00}}{q_{00}} \times \dfrac{p_{01}}{q_{01}} \Big/ \dfrac{p_{00}}{q_{00}}$ d'où

$$\Psi(1,1) = \Psi(1,0) \times \Psi(0,1)$$

Il y a multiplication des odds-ratio, qui est la définition de l'absence d'interaction (voir Chapitre V).

Une façon de prendre en compte l'interaction est de considérer le modèle

$$\text{logit } P = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \gamma x_1 x_2 \qquad (9)$$

où $x_1 x_2$ désigne le produit du x_1 par x_2.
(Avec ce modèle, l'effet de X_1 mesuré par le odds-ratio, dépend du niveau de x_2 = c'est la définition même de l'interaction).

Tester s'il y a une interaction revient donc à tester l'hypothèse nulle $H_0 : \gamma = 0$, soit au moyen du rapport γ/σ_γ, soit par le test des rapports de maximum de vraisemblance $G^2 = 2\left[L - L_0\right]$, L maximum de la vraisemblance sous (9) et L_0 maximum sous (S).

Si le modèle (9) est retenu, les estimations de β_1 et β_2 et γ sont imprécises du fait de la forte correlation entre les variables X_1, X_2 d'une part et la variable X_1X_2. Une façon de tourner la difficulté est de remplacer X_1 et X_2 par $X_1 - E(X_1)$ et $X_2 - E(X_2)$.

En effet cherchons la covariance de $X_1 - E(X_1)$ avec $(X_1 - E(X_1)).(X_2 - E(X_2))$. C'est par définition

$$E\left[(X_1 - E(X_1)).(X_1 - E(X_1))(X_2 - E(X_2)\right] = E\left\{\left[X_1 - E(X_1)\right]^2.(X_2 - E(X_2))\right\}$$

Si X_1 et X_2 sont indépendants, cette covariance vaut $E((X_1 - E(X_1))^2).E(X_2 - E(X_2))$. Le deuxième terme du produit étant nul, $(X_1 - E(X_1))$ est indépendant du produit $(X_1 - E(X_2))(X_2 - E(X_2))$.

L'estimation du (9) n'est pas modifiée par ce changement de variable : cependant les variances des estimations de β_1 et β_2 (pas celle de γ) sont plus petites quand on travaille sur les

variables transformées. Ce résultat est important : si dans u
équation de régression (linéaire ou logistique) on introduit
des termes produit, il faut travailler avec des variables
transformées, X-E(X) (ou X - mode (X)).

7 - APPLICATIONS DU MODELE LOGISTIQUE

Elles sont très nombreuses. Les techniques d'ajustement,
comme cas particulier de l'utilisation du modèle logistique,
seront présentées oralement à partir d'un exemple numérique.

Pour une discussion approfondie voir [3].

CHAPITRE XIII - REFERENCES

[1] - E.L. LEHMANN
Testing statistical hypothesis.
1959. J. Wiley, New York

[2] - M. HALPERIN, W.C. BLACKWELDER, J.I. VERTER
Estimation of the multivariate logistic risk function : a comparison of the discriminant function and maximum likelihood approches
1971, J. Chron. Dis. 24, 125-158.

[3] - D.G. KLEINBAUM, L.L. KUPPER, L.E. CHAMBLESS
Logistic Regression Analysis of Epidemiologic Data : Theory and practice
1982. Comm. Statist. Theor. Meth. 1982, 11, 485-547.

CHAPITRE XIV

LE MODELE LOGISTIQUE : II - LES ENQUETES CAS-TEMOINS

Le chapitre précédent a présenté une modélisation particulièrement intéressante de la probabilité de survenue d'une maladie en fonction des valeurs prises par un certain nombre de facteurs de risque (modèle logistique). Tel qu'il a été décrit, il ne semble pouvoir s'appliquer qu'aux enquêtes cohorte (dans lesquelles les valeurs des facteurs de risque sont considérées comme fixées et l'état malade/non malade aléatoire).

Cependant un très important résultat est qu'il est possible d'analyser au moyen de ce modèle une enquête cas-témoins (où les nombres de malades et de témoins sont fixés par l'investigateur, et où les valeurs des facteurs de risque sont aléatoires) exactement comme s'il s'agissait d'une enquête cohorte.

Ce résultat peut paraître très surprenant ; en fait on en a déjà rencontré un cas particulier. C'est celui où il n'y a qu'une seule variable de risque dichotomique : le odds-ratio

$$\psi = \frac{P_1}{1-P_1} \Big/ \frac{P_o}{1-P_o}$$ (où les P sont des probabilités de maladie)

peut être estimé dans une enquête cas-témoins aussi bien que dans une enquête cohorte.

Le but de ce chapitre est de présenter plusieurs approches pour établir le résultat. Son intérêt est donc essentiellement théorique.

1 - APPROCHE DE MANTEL (1)

Supposons qu'on ait réalisé une enquête cohorte sur N sujets, et qu'à la fin de la durée de surveillance n_1 de ces sujets soient devenus malades, les $N.- n = n_o$ autres restant sains. Décidons

.../...

de faire l'analyse sur un pourcentage π_1 des malades et un pourcentage π_0 sur non malades, ou de façon plus précise qu'u malade rentrera dans l'analyse avec une probabilité π_1 et un non malade avec la probabilité π_0. Typiquement π_1 sera pris égal à 1, tandis que π_0 sera beaucoup plus faible.

Soit comme d'habitude :

$P(x) = Pr[M/x] = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$ la probabilité de faire la maladie, lorsque le vecteur X des facteurs de risque a la valeur x (β est lui-même un vecteur). Calculons la même probabilité, mais sur l'échantillon d'étude : il est clair qu'elle est d'autant plus grande que le rapport du nombre de malades au nombre de non malades étudiés estplus grand. De façon plus précise on a à calculer $P[M/x : Z = 1]$, en appelant Z la variable qui vaut 0 si le sujet ne fait pas partie de l'échantillon, et 1 s'il en fait partie.

Le théorème de Bayes (1) donne

$$Pr[M/x ; Z=1] = \frac{Pr[M/x] \cdot Pr[Z=1/M \text{ et } x]}{Pr[Z=1]} = \frac{P(x)\,\pi_1}{\pi_0[1-P(x)] + \pi_1 P(x)}$$

$$= \frac{\pi_1 e^{\alpha+\beta x}}{\pi_0 + \pi_1 e^{\alpha+\beta x}} = \frac{\frac{\pi_1}{\pi_0} e^{\alpha+\beta x}}{1+\frac{\pi_1}{\pi_0} e^{\alpha+\beta x}} = \frac{e^{\tilde{\alpha}+\beta x}}{1+e^{\tilde{\alpha}+\beta x}}$$

en posant $\tilde{\alpha} = \alpha + \log \frac{\pi_1}{\pi_0}$

Ainsi la probabilité d'être malade, dans l'échantillon s'exprime sous la forme logistique avec des paramètres β qui sont ceux cherchés ($\tilde{\alpha}$ est bien entendu différent de α, puisqu'il dépend de π_1 et π_0). (Le calcul précédent suppose que π_0 et π_1 ne dépendent pas de la valeur de x).

---

(1) le théorème de Bayes s'écrit de façon générale

$$Pr[A/B] = \frac{Pr[A] \cdot Pr[B/A]}{Pr[B]}.$$

Ici, l'événement A est M/x et l'événement B est Z=1

Si donc on admet que les malades d'une enquête cas-témoins constituent un échantillon de malades provenant d'une cohorte "sous-jacente" (peut être pas très bien définie) et qu'il en est de même des témoins, on peut estimer les β comme dans une enquête cohorte.

Cette approche a le mérite de la simplicité : elle souffre d'une petite imperfection théorique. Puisque chaque sujet de l'enquête cohorte a une probabilité (π_0 ou π_1) d'être inclus dans l'étude cas-témoins, les nombres de malades et de témoins de cette étude sont en fait aléatoires, alors qu'en réalité ils sont certains, puisque fixés par l'investigateur. Elle s'applique par contre parfaitement à l'enquête cohorte où une fraction des malades (généralement tous) est comparée à une fraction de non malades, par exemple s'il s'agit d'étudier un dosage couteux que l'on peut effectuer longtemps après le prélèvement.

2 - APPROCHE CONDITIONNELLE [2]

L'enquête cas-témoins compare n_1 malades et n_0 témoins ($n = n_0 + n_1$). Les n vecteurs des facteurs de risque sont $\{x_1, x_2, \ldots x_n\}$ dont certains, disons les n_1 premiers, correspondent à des malades et les autres à des témoins.

Etant donnés n_1, n_0 et l'ensemble $\{x_1, \ldots x_n\}$ on cherche la probabilité de la configuration observée c'est-à-dire la probabilité que les n_1 malades aient les vecteurs des facteurs de risque $x_1, x_2, \ldots x_{n_1}$.

A titre d'exemple, et pour bien comprendre le calcul, on va chercher cette probabilité lorsque $n_1 = 2$ et $n_0 = 3$.

Les vecteurs des facteurs de risque sont $\{x_1, x_2, x_3, x_4, x_5\}$, x_1 et x_2 correspondant aux malades.

Il y a $10 = \binom{5}{2}$ configurations, en ce qui concerne les x possibles pour les malades ; ce sont $(x_1\ x_2)$, (x_1, x_3) ..... (x_4, x_5).

.../...

Cherchons la probabilité d'une de ces configurations, par exemple (x_2, x_5). C'est

$$P_r[M/x_2].\ P_r[M/x_5].\ P_r[\bar{M}/x_1].\ P_r[\bar{M}/x_3].\ P_r[\bar{M}/x_4] = \frac{e^{\alpha+\beta x_2}.\ e^{\alpha+\beta x_5}}{\prod_{i=1}^{5}(1+e^{\alpha+\beta x_i})} \qquad (1)$$

La probabilité de la configuration réellement observée (x_1, x_2) est

$$\frac{e^{\alpha+\beta x_1}.\ e^{\alpha+\beta x_2}}{\prod_{i=1}^{5}(1+e^{\alpha+\beta x_i})} \qquad (2)$$

Ainsi la probabilité de cette configuration, conditionnelle à n_1, n_0 et x est le rapport de (2) à la somme des 10 probabilités de type (1). C'est après simplification

$$\frac{e^{\beta(x_1+x_2)}}{\sum e^{\beta(x_i+x_j)}}$$

De façon générale, la probabilité (ou vraisemblance) conditionnelle vaut

$$V = \frac{e^{\beta(x_1+\cdots+x_{n_1})}}{\sum e^{\beta(x_{i_1}+\cdots+x_{i_{n_1}})}} \qquad (3)$$

la somme figurant au dénominateur porte sur les $\binom{n}{n_1}$ façons de choisir n_1 vecteurs (x) parmi les n.

L'estimation de β s'obtient par maximisation de (3). Avant de discuter cette approche conditionnelle considérons un exemple simple de calcul de l'expression (3).

Exemple : Il n'y a qu'un seul facteur de risque X dichotomique, prenant la valeur 1 ou 0 selon qu'il est présent ou absent. Les données peuvent alors se résumer en le tableau 2 x 2 introduit au Chapitre VI.

.../...

	E^+	E^-	
M^+	a	b	n_1
M^-	c	d	n_0
	m_1	m_0	n

Le numérateur de V est simplement $e^{\beta a}$
Le dénominateur est à peine moins simp
Les n valeurs x sont les suivantes :

$\left.\begin{matrix}1\\ \cdot\\ \cdot\\ \cdot\\ 1\end{matrix}\right\}$ m_1 fois $\qquad$ $\left.\begin{matrix}0\\ \cdot\\ \cdot\\ \cdot\\ 0\end{matrix}\right\}$ m_0 fois

et il faut en choisir n_1; si on en choisit x parmi les m_1 et $n_1 - x$ parmi les m_0, la quantité correspondante vaut : $e^{\beta x}$. Ceci peut être fait de $\binom{m_1}{x}\binom{m_0}{n_1-x}$ façons différentes.

En posant $e^{\beta} = \psi$, on retrouve bien la formule (3) du Chapitre VI, au facteur multiplicatif $\binom{m_1}{a}\binom{m_0}{n_1-a}$ près qui ne dépend pas de ψ . (Dans la démarche du Chapitre VI, on cherche la probabilité que a parmi les n_1 soient exposés; ici il s'agit de la probabilité que a sujets nommément désignés soient exposés.

La vraisemblance conditionnelle (3) est la même pour les enquêtes cohortes et les enquêtes cas-témoins, puisqu'elle est calculée en supposant connus à la fois les nombres de malades et de non malades et les valeurs de facteurs de risque : seuls sont aléatoires les couples "malade-valeur des facteurs de risque".

De façon générale, maximiser (3) conduit à des calculs qui sont rapidement inextricables (même avec un ordinateur puissant) quand n_1 et n_0 deviennent grands. Cependant, on peut montrer [3] que les estimations de β qui maximisent la vraisemblance non conditionnelle (dans une enquête cohorte) qui figure au chapitre précédent, et la vraisemblance conditionnelle (3) sont proches numériquement et d'autant plus que n_0 et n_1 sont grands.

Donc, plutôt que maximiser (3), on maximise la vraisemblance non conditionnelle des observations d'une enquête cohorte. Ceci est une deuxième justification du résultat annoncé en début de ce chapitre.

.../...

Les considérations qui précédent, dont certaines généralisent ce qui a été dit au Chapitre VI, pourraient laisser croire que vraisemblance conditionnelle et vraisamblance inconditionnelle peuvent être utilisées indifféremment (aux complexités de calcul près). Le chapitre suivant montrera qu'il n'en est rien : dans certaines situations bien particulières, seule l'approche conditionnelle est valable.

3 - APPROCHE DE PRENTICE ET PYKE [4]

Cette approche est plus subtile que les précédentes. Elle utilise le fait que dans une enquête cas-témoins, n_0 et n_1 ($n_0 + n_1 = n$) sont donnés, mais les X aléatoires.

Appelons $f_0(x) = Pr[x/\bar{M}]$, la probabilité (ou la densité) de chez les témoins et $f_1(x) = Pr[x/M]$ la même quantité chez les malades.

Il existe une relation entre $f_0(x)$ et $f_1(x)$ qui tient à l'existence de la relation, $P(M/x)$, qui lie M à x (fonction logistique).

Considérons une valeur particulière x_0 de X. Les quantités

$$\frac{Pr(M/x)}{Pr(\bar{M}/x)} \Big/ \frac{Pr(M/x_0)}{Pr(\bar{M}/x_0)} \quad \text{et} \quad \frac{f_1(x)}{f_1(x_0)} \Big/ \frac{f_0(x)}{f_0(x_0)} \quad \text{sont égales}$$

Il suffit en effet d'appliquer le théorème de Bayes sous la forme

$$Pr(M/x) = \frac{Pr(M).Pr(x/M)}{Pr(x)} = \frac{Pr(M)f_1(x)}{Pr(x)} \quad \text{et}$$

$$Pr(\bar{M}/x) = \frac{Pr(\bar{M}).f_0(x)}{Pr(x)}$$

D'ailleurs les 2 quantités précédentes ne sont rien d'autre que $\frac{P_1}{\bar{1}-P_1} \Big/ \frac{P_0}{1-P_0}$ et $\frac{P_1}{1-p_1} \Big/ \frac{P_0}{1-p_0}$ où les P sont les probabilités de maladie pour x_0 et x et les p sont les probabilités de x chez les malades et les témoins (X n'a que 2 valeurs possibles x et

.../...

Par hypothèse $Pr(M/x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$. On trouve immédiatement que

$$\frac{f_1(x)}{f_1(x_0)} \Big/ \frac{f_0(x)}{f_0(x_0)} = e^{\beta(x-x_0)}$$

ou encore $f_1(x) = c\, f_0(x) e^{\beta x}$ où c est une constante normalisatrice (l'intégrale de f_1 doit être égale à 1).

Considérons la population constituée par le mélange de n_0 témoins et n_1 malades. La loi de X dans cette population est

$$f(x) = \frac{n_0}{n} f_0(x) + \frac{n_1}{n} f_1(x) = \frac{f_0(x)}{n} [n_0 + c n_1 e^{\beta x}]$$

On peut inversement exprimer $f_0(x)$ et $f_1(x)$ en fonction de $f(x)$. On trouve que

$$f_0(x) = \frac{1}{1+e^{\delta+\beta x}} \frac{n}{n_0} f(x)$$

$$f_1(x) = \frac{e^{\delta+\beta x}}{1+e^{\delta+\beta}} \frac{n}{n_1} f(x) \quad \text{(où l'on a posé } e^{\delta} = c \frac{n_1}{n_0}\text{)}$$

La vraisemblance des observations $x_1, x_2, \ldots x_n$ est

$$\prod^{n_0} f_0(x_j) \cdot \prod^{n_1} f_1(x_i)$$

Elle vaut (à une constante multiplicative près)

$$\frac{\prod^{n_1} (e^{\delta+\beta x_i})}{\prod^{n_0+n_1} (1+e^{\delta+\beta x_i})} \cdot f(x_1)\, f(x_2) \ldots f(x_n). \qquad (4)$$

Les inconnues sont δ, β et $f(x)$. $f(x)$ contient une certaine information sur β. Mais si l'on renonce à cette information, $f(x)$ doit être considérée comme totalement non explicitée. On a donc a estimer séparément δ, β et les $f(x_i)$ soit au total $n + 2$ paramètres

Il existe une contrainte entre ces paramètres : elle exprime que f_0 et f_1 sont des lois de probabilité, c'est-à-dire que

$$\frac{n_0}{n} = \int \frac{1}{1+e^{\delta+\beta x}} f(x)\, dx \quad \left(\text{ou } \frac{n_1}{n} = \int \frac{e^{\delta+\beta x}}{1+e^{\delta+\beta x}} f(x)\, dx\right).$$

.../...

Il faut donc trouver les paramètres sujets à cette contrai
qui maximisent (4).

Le calcul relativement simple montre que :

a) $f(x_i) = \frac{n(x_i)}{n}$, où $n(x_i)$ est le nombre de sujets dont le X vaut x_i

b) les estimations $\hat{\delta}$ et $\hat{\beta}$ qui satisfont la contrainte sont cell
qui maximisent la quantité

$$\frac{\prod^{n_1} (e^{\delta+\beta x_i})}{\prod^{n} (1+e^{\delta+\beta x_i})}$$ qui est précisément la vraisemblance

des observations considérées comme issues d'une enquête cohorte (x_i donnés, n_0 et n_1 aléatoires).

Ce dernier point démontre le résultat annoncé.

4 - APPROCHE PAR L'ANALYSE DISCRIMINANTE

La démarche précédente ne supposait rien sur la forme de $f_0(x)$ (et $f_1(x)$) distribution de facteurs de risque chez les témoins (et les malades). Faisons maintenant l'hypothèse que $f_0(x)$ est multinormale de vecteur moyen μ_0 et de matrice de variances-covariances Σ .

$$f_0(x) = K \exp -\frac{1}{2} [x - \mu_0]' \Sigma^{-1} [x - \mu_0].$$

La relation $f_1(x) = c\, f_0(x) e^{\beta x}$ prouve que f_1 est aussi multinormale et a la même matrice de variances-covariances Σ .

Si donc on fait l'hypothèse que f_0 et f_1 sont multinormales de moyennes μ_0 et μ_1 et ont la même matrice de variances-covariances Σ , c'est qu'on adopte le modèle logistique, dont les coefficients β sont simplement donnés par

$$\beta' x = \log \frac{f_1(x)}{f_0(x)} + K$$

Il est facile de montrer que

$$\log \frac{f_1(x)}{f_0(x)} = (\quad) \Sigma^{-1} (\mu_1 - \mu_0) \quad \text{donc} \quad \beta = \Sigma^{-1} (\mu_1 - \mu_0) \qquad (5)$$

$$\left(x - \frac{(\mu_0+\mu_1)}{2}\right)'$$

Ainsi sous l'hypothèse faite (f_0 et f_1 multinormales, de même matrice Σ) on peut estimer β par le vecteur $S^{-1}(m_1 - m_0)$, où S^{-1} est l'estimation de la matrice commune, m_1 et m_0, sont les estimations des vecteurs moyennes de 2 populations.

Or ce vecteur n'est autre que l'ensemble des coefficients du plan de discrimination linéaire entre M et $\bar{M}$.

On dispose ainsi d'une autre estimation de la fonction logistique, valable d'ailleurs aussi bien dans l'étude de cohortes que d'enquêtes cas-témoins.

Il est à noter que les coefficients de la fonction discriminante ne sont autres que les coefficients de la régression linéaire de la variable Y (1 pour les malades, 0 pour les témoins) sur les facteurs de risque.

Le tableau ci-dessous compare les avantages et les inconvénients des estimations par la méthode du maximum de vraisemblance et l'analyse discriminante.

	maximum de vraisemblance	analyse discriminante
Distribution des x	- quelconque - peuvent être discrets	multinormales
Calculs	- Itératifs	non-itératifs

On peut montrer [5] que si les hypothèses de normalité ne sont pas vérifiées, l'analyse discriminante conduit à des estimations de β qui peuvent être assez fortement biaisées, mais que leur test a 0 peut être considéré comme satisfaisant. On peut donc proposer (si on est limité au point de vue des calculs) la démarch suivante : 1) trouver au moyen d'analyses discriminantes les facteurs de risque à faire entrer dans une équation logistique (il y a généralement plusieurs essais) ; 2) trouver l'équation par la méthode du maximum de vraisemblance, les premières valeurs du calcul itératif étant fournies par l'analyse discriminante.

CHAPITRE XIV - REFERENCES

[1] - N. MANTEL
Synthetic retrospective studies and related topics.
1973. Biometrics, 29, 479-486.

[2] - N.E. BRESLOW, N.E. DAY
Statistical methods in cancer Reserach
Vol. 1. The analysis of case control studies
1980, IARC Lyon

[3] - V.T. FAREWELL
Some results on the estimation of logistic models based o
retrospective date
1979, Biometrika, 66, 27-32

[4] - R.L. PRENTICE, R. PYKE
Logistic disease incidence models and case control studie
1979, Biometrika, 66, 403-411.

[5] - S.J. PRESS, S. WILSON
Choosing between logistic regression and discrimination
analysis.
1978, J. Am. Stat. Ass. 70, 699-705.

CHAPITRE XV

LE MODELE LOGISTIQUE : III - ECHANTILLONS APPARIES

Ce chapitre généralise les notions qui ont été présentées au Chapitre X.

1 - ENQUETE COHORTE

Imaginons que pour l'étude d'un facteur de risque E, on décide d'apparier chaque sujet exposé à un sujet non exposé. A la fin du suivi de la cohorte, les N paires de sujets se répartissent en les 4 groupes ci-dessous

	E^+	E^-	nbre de paires	
(a)	$\bar{M}$	$\bar{M}$	n_{00}	
(b)	M	$\bar{M}$	n_{10}	$n_{10} + n_{01} = n$
(c)	$\bar{M}$	M	n_{01}	
(d)	M	M	n_{11}	
	TOTAL		N	

Le protocole des séries appariées peut être considéré comme un cas particulier d'un protocole avec stratification, les strates étant formées de 2 sujets, un exposé et un non exposé.

Il est donc raisonnable de vouloir appliquer les techniques vues au Chapitre XIII paragraphes 5 et 7.

Si X désigne l'exposition (X = 1 pour E^+ et X = 0 pour E^-) le modèle logistique qui stipule que $P(x) = \frac{e^{\alpha_i + \beta x}}{1 + e^{\alpha_i + \beta x}}$ se réduit ici à

$$P_1 = \frac{e^{\alpha_i + \beta}}{1 + e^{\alpha_i + \beta}} \quad \text{et} \quad P_0 = \frac{e^{\alpha_i}}{1 + e^{\alpha_i}}$$

où α_i est une caractéristique de la strate, ici de la paire.

La vraisemblance d'une paire de type (a) est $\frac{1}{1+e^{\alpha_i}} \cdot \frac{1}{1+e^{\alpha_i+\beta}}$

" " (d) est $\frac{e^{\alpha_i}}{1+e^{\alpha_i}} \cdot \frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}$

" " (c) est $\frac{1}{1+e^{\alpha_i}} \cdot \frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}$

" " (b) est $\frac{e^{\alpha_i}}{1+e^{\alpha_i}} \quad \frac{1}{1+e^{\alpha_i+\beta}}$

La vraisemblance totale est le produit de n_{00} vraisemblances (a), n_{11} vraisemblances (d) etc. ...

Il y a N + 1 inconnues, les N α_i et β.

Soit à estimer par la méthode du maximum de vraisemblance le α_i d'une paire de type (a). La dérivée de la vraisemblance par rapport à α_i est

$$-\left\{\frac{e^{\alpha_i}}{(1+e^{\alpha_i})^2} \quad \frac{e^{\alpha_i+\beta}}{(1+e^{\alpha_i+\beta})^2}\right\}$$

qui ne peut être nulle que si $\alpha_i = -\infty$. La vraisemblance correspondante est 1, quel que soit β; on verrait de même que l'estimation de α_i dans les paires des types (d) est $\alpha_i = +\infty$, et que la vraisemblance est également 1 quel que soit β.

La vraisemblance a maximum se réduit ainsi à

$$\prod^{n_{10}} \left(\frac{1}{1+e^{\alpha_i}} \cdot \frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}\right) \cdot \prod^{n_{01}} \left(\frac{e^{\alpha_i}}{1+e^{\alpha_i}} \cdot \frac{1}{1+e^{\alpha_i+\beta}}\right)$$

dont le logarithme vaut $\sum \alpha_i + n_{10}\beta - \sum \log_e(1+e^{\alpha_i}) - \sum \log_e(1+e^{\alpha_i+\beta})$

On peut facilement montrer que les estimateurs du maximum de vraisemblance sont

$$\begin{cases} \hat{\alpha}_i = -\frac{\hat{\beta}}{2} & \forall i \\ e^{\hat{\beta}} = \left(\frac{n_{10}}{n_{01}}\right)^2 \end{cases}$$

Ainsi la méthode du maximum de vraisemblance conduit pour le odds-ratio $\psi = e^{\beta}$ à la valeur incorrecte (dans le sens qu'elle est biaisée) $(\frac{n_{10}}{n_{01}})^2$ au lieu de la valeur $\frac{n_{10}}{n_{01}}$. L'explication de ce phénomène, a priori surprenant est la suivante : on sait que les estimateurs du maximum de vraisemblance ont de "bonnes" propriétés: en particulier, les estimations tendent vers les vraies valeurs des paramètres à estimer, lorsque les effectifs sont grands (tendent vers l'infini). Cependant, ceci n'est vrai, qui si les paramètres à estimer sont en nombre fini . Or dans le problème qui nous occupe, le nombre de paires, donc le nombre des α_i à estimer, augmente avec le nombre total des observations, et dans ce cas les estimateurs n'ont plus les propriétés "optimales" enoncées.

Comment peut-on échapper à cette difficulté ? En écrivant une vraisemblance qui ne dépende pas des paramètres α_i dont le nombre augmente avec l'effectif : ceci est possible si pour chaque strate (ici chaque paire) on se place conditionnellement aux marges. On se rappellera la formule (3) du Chapitre XIV qui montre bien que cette vraisemblance conditionnelle ne dépend que de β.

Pour chacun des cas (a) à (d) la vraisemblance conditionnell figure au Chapitre X. On peut aussi les obtenir comme cas particulier de la formule (3) du Chapitre XIV.

2 - ENQUETES CAS-TEMOINS

Dans le chapitre précédent on a insisté sur le fait que la vraisemblance conditionnelle est la même dans le cas de cohortes que dans celui d'études cas-témoins.

Ce qui suit concerne donc aussi bien le premier type d'enquêtes que le second. Dans la présentation, nous parlerons essentiellement en termes d'enquêtes cas-témoins.

3 - PLUSIEURS TEMOINS POUR LE MEME MALADE [1]

3-1. : 2 témoins

Supposons pour débuter qu'à chaque malade sont associés k = 2 témoins. Si l'exposition est dichotomique les "triplets" possibles sont les suivants

	E^+	E^-	
M^+	1	0	1
M^-	2	0	2
	3	0	

1	0	1
1	1	2
2	1	

0	1	1
2	0	2
2	1	

1	0	1
0	2	2
1	2	

0	1	1
1	1	2
1	2	

0	1	1
0	2	2
0	3	

en nombre

n_{12} n_{11} n_{02} n_{10} n_{01} n_{00}

(n_{ij} est le nombre de triplets où i(0,1) malades et j(0,1,2) témoins témoins sont exposés).

Les vraisemblances correspondantes sont

$$1 \qquad \frac{2\psi}{2\psi+1} \qquad \frac{1}{2\psi+1} \qquad \frac{\psi}{2+\psi} \qquad \frac{2}{2+\psi} \qquad 1$$

Comme application de la formule 3 du Chapitre VI. (Si on utilise la formule (3)du Chapitre XIV, les expressions sont les mêmes à des facteurs multiplicatifs près).

Par exemple, la probabilité de la 5è table est

$$\frac{\binom{1}{0}\binom{2}{1}\psi^0}{\binom{1}{0}\binom{2}{1}\psi^0+\binom{1}{1}\binom{2}{0}\psi} = \frac{2}{2+\psi}$$

La vraisemblance totale est ainsi proportionnelle à

$$\frac{\psi^{n_{11}+n_{10}}}{(2\psi+1)^{n_{11}+n_{02}}\,(2+\psi)^{n_{10}+n_{01}}}$$

dont le logaithme $(n_{11}+n_{10})\log\psi-(n_{11}+n_{02})\log(2\psi+1)-(n_{10}+n_{01})\log(2+\psi)$ est maximum

pour $\hat{\psi}$ vérifiant $\frac{n_{11}+n_{10}}{\hat{\psi}} = \frac{2(n_{11}+n_{02})}{2\hat{\psi}+1} + \frac{(n_{10}+n_{01})}{2+\hat{\psi}}$

$\hat{\psi}$ est solution d'une équation d'un second degré dont une seule est acceptable (positive).

3-2. : k témoins

Considérons maintenant le cas général de k-témoins. On peut opérer exactement comme il vient d'être fait pour k = 2.

Parmi les (k + 1) sujets exposés, on peut observer 0, 1, ...k+

S'il y a au total m exposés, ce peut être le malade et (k-1) témoins ou m témoins

Les tables correspondantes sont

	E^+	E^-	
M^+	1	0	1
M^-	m-1	k-m+1	k
	m	k-m+1	(k+1)

	E^+	E^-	
M^+	0	1	1
M^-	m	k-m	k
	m	k-m+1	k+1

en nombre

$$n_{1,m-1} \qquad n_{0,m}$$

Les vraisemblances correspondantes sont

$$\frac{\binom{m}{1}\binom{k-m+1}{0}\psi}{\binom{m}{0}\binom{k-m+1}{1}+\binom{m}{1}\binom{k-m+1}{0}\psi} \quad \text{et} \quad \frac{\binom{m}{0}\binom{k-m+1}{1}}{\text{id.}}$$

qui se simplifient en

$$p_m = \frac{\psi}{\psi + \frac{k-m+1}{m}} \qquad q_m = \frac{\frac{k-m+1}{m}}{\psi + \frac{k-m+1}{m}}$$

La vraisemblance totale est

$$\prod_{m=1}^{k} p_m^{n_{1,m-1}} q_m^{n_{0,m}}$$

(en principe le produit va de m = 0 à m = k+1, mais pour ces valeurs extrêmes la vraisemblance vaut 1).

Son logarithme (à une constante additive près)

$$\sum_m n_{1,m-1} \log \psi - (n_{1,m-1} + n_{0,m}) \log \left(\psi + \frac{k-m+1}{m}\right)$$

a une dérivée qui s'annule pour $\hat{\psi}$ satisfaisant à

$$\sum_{m=1}^{k} n_{1,m-1} = \psi \sum_{m=1}^{k} \frac{m T_m}{m\psi + k - m + 1} \qquad (1)$$

où T_m désigne le nombre total de (k+1) uplets où il y a exactement m exposés.

C'est une équation algébrique que l'on peut résoudre de façon itérative par les méthodes classiques.

3-3. Estimateur de Mantel-Haenszel

Les calculs numériques qui précédent peuvent être très lourds. Un estimateur immédiat a obtenir a été proposé par Mantel Haenszel. Son expression est la suivante

$$\psi_{M.H} = \frac{\sum_{m=1}^{k} (k-m+1)\, n_{1,m-1}}{\sum_{1}^{k} m\, n_{0,m}} \qquad (2)$$

3-4. Un exemple numérique (avec k=2)

Nbre de malades exposés \ Nbre de témoins exposés	0	1	≥ 2
0	$n_{00}=12$	$n_{01}=23$	$n_{02}=11$
1	$n_{10}=17$	$n_{11}=42$	$n_{12}=31$

$T_1 = 23 + 17 = 4$

$T_2 = 53$

L'équation du maximum de vraisemblance est

$$17 + 42 = \psi\left[\frac{40}{\psi+2} + \frac{2.53}{2\psi+1}\right]$$ qui se simplifie en

$$68\,\psi^2 + 57\,\psi - 118 = 0$$

dont la seule racine positive est $\hat{\psi} = 1,80$

quant à $\hat{\psi}_{M-H}$ il vaut $\dfrac{2 \times 17 + 42}{23 + 2 \times 11} = \dfrac{76}{45} = 1,69$

3-5. Précision des estimations

3-5-1. Estimateur du maximum de vraisemblance

On rappelle que si $\hat{\psi}$ est l'estimateur du maximum de vraisemblance, sa variance asymptotique est

$$\sigma^2_{\hat{\psi}} = I^{-1}(\psi)$$

où I est l'opposé de la dérivée seconde du logarithme de la vraisemblance par rapport à ψ. En fait, comme $\log \hat{\psi}$ a une distribution qui tend vers la normale plus rapidement que celle de $\hat{\psi}$, il vaut mieux chercher un intervalle de confiance pour $\log \hat{\psi}$.

La variance de $\log \hat{\psi}$ est liée à celle de $\hat{\psi}$ par la relation $\sigma^2_{\log \hat{\psi}} = \dfrac{\sigma^2_{\hat{\psi}}}{\psi^2}$ soit $\sigma^2_{\log \hat{\psi}} = \dfrac{1}{\psi^2} I(\psi)$

Calculons $I(\psi)$ dans le cas où k=2 (2 témoins pour un malade). La dérivée première du logarithme de la vraisemblance étant $\dfrac{(n_{11} + n_{10})}{\psi} - \dfrac{2T_2}{2\psi+1} - \dfrac{T_1}{2+\psi}$ la dérivée seconde vaut

$$-\frac{n_{11} + n_{10}}{\psi^2} + \frac{4T_2}{(2\psi+1)^2} + \frac{T_1}{(2+\psi)^2}$$

Pour en avoir une estimation, il suffit de remplacer ψ par son estimation $\hat{\psi}$.

or $\dfrac{n_{11} + n_{10}}{\hat{\psi}} = \dfrac{2T_2}{2\hat{\psi}+1} + \dfrac{T_1}{2+\hat{\psi}}$. En remplaçant $\dfrac{n_{11} + n_{10}}{\psi}$ par cette valeur, on trouve après des calculs simples que

$$\sigma^2_{\hat{\psi}} = \frac{\hat{\psi}}{2\left[T_2/(2+\hat{\psi})^2 + T_1/(1+2\hat{\psi})^2\right]}$$

d'où la variance de $\log \hat{\psi}$

$$\hat{\sigma}^2_{\log\hat{\psi}} = \frac{1}{2\hat{\psi}\left[\frac{T_1}{(1+\hat{\psi})^2} + \frac{T_2}{(1+2\hat{\psi})^2}\right]}$$

- dans le cas de k témoins on peut montrer de la même façon que

$$I(\psi) = \sum_{m=1}^{k} \frac{T_m\left(\frac{k+1-m}{m}\right)}{\psi\left[\psi + \frac{k+1-m}{m}\right]^2}$$

et

$$\text{Var}(\log\hat{\psi}) = \left[\sum_{m=1}^{k} \frac{\psi m T_m(k-m+1)}{(m\psi + k - m + 1)^2}\right]^{-1} \qquad (3)$$

3-5-2. Estimateur de Mantel-Haenszel

Un travail très récent [2] donne l'expression suivante pour la variance du log de $\hat{\psi}_{M-H}$

$$\text{Var}(\log\hat{\psi}_{M.H}) = \frac{\sum_{m=1}^{k}(k-m+1)^2 n_{1,m-1} n_{0,m} \hat{\psi}^2_{M-H} + \sum_{m=1}^{k} m^2 n_{0,m} n_{1,m-1}}{\left[\sum_{m=1}^{k}(k-m+1) n_{1,m-1} n_{0,m}\right]^2} \qquad (4)$$

Considérons l'exemple suivant qui figure dans ce travail

Malades \ Témoins	0	1	2	3	4
0	$n_{00} = 0$	4	1	1	1
1	$n_{10} = 3$	17	16	15	5

L'équation (1) du maximum de vraisemblance est

$$3 + 17 + \ldots + 15 = \psi\left[\frac{7}{\psi+4} + \frac{2\times 18}{2\psi+3} + \frac{3\times 17}{3\psi+2} + \frac{4\times 16}{4\psi+1}\right]$$

Sa solution est $\hat{\psi} = 7{,}95$

et var $(\log\hat{\psi})$ calculé selon la formule (3) vaut 0.177, d'où l'intervalle de confiance pour $\psi = 7.95(\exp.\pm 1.96\sqrt{0.177}) = (3.5, 18.$

L'estimateur de Mantel Haenszel est

$$\frac{3 \times 4 + 17 \times 3 + 16 \times 2 + 15}{4 \times 1 + 1 \times 2 + 1 \times 3 + 1 \times 4} = 8,46$$

et la variance de son logarithme

$$\frac{4^2 \times 3 + 3^2 . 17 + 2^2 \times 16 + 1^2 \times 15 + 8,46^2 (4 + 2^2 . 1 + 3^2 \times 1 + 4^2 . 1)}{(4 \times 3 + 3 \times 17 + 2 \times 16 + 1 \times 15)} = 0.218$$

qui donne un intervalle de confiance de (3.4 ; 21.1) très voisin de celui du maximum de vraisemblance, mais beaucoup plus simple à obtenir.

3-6. Test de $\Psi = 1$

Il peut être basé sur le rapport $\frac{\log \hat{\Psi}}{\sigma(\log \hat{\Psi})}$ si on ne s'intéresse pas à l'estimation du risque-relatif, construit bien plus simplement à partir du raisonnement suivant :

Considérons les T_m $(k+1)$uplets où figurent exactement m exposés ; ils sont de 2 types, ceux où le malade est exposé (en proportion théorique $p_m = \frac{m\Psi}{m\Psi + k - m + 1}$) et ceux où le malade n'est pas exposé (en proportion théorique $q_m = \frac{k-m+1}{m\Psi + k - m + 1}$). Sous l'hypothèse nulle, ces proportions sont $p_m = \frac{m}{k+1}$ et $q_m = \frac{k-m+1}{k+1}$

Ainsi sous H_o, $n_{1,m-1}$ a pour espérance $\frac{mT_m}{k+1}$ et pour variance $T_m \frac{m(k-m+1)}{(k+1)^2}$ et $\Sigma n_{1,m-1}$ a pour espérance et pour variance la somme de ces espérances et variances.

Le test est ainsi basé sur la statistique

$$\frac{\sum_{m=1}^{k} \left(n_{1,m-1} - m \frac{T_m}{k+1} \right)}{\frac{1}{k+1} \sqrt{\Sigma T_m m (k-m+1)}} \qquad (5)$$

qui sous H_o est normale réduite.

Dans l'exemple précédent

$$\frac{\log \hat{\psi}}{\sigma(\log \hat{\psi})} = \frac{2.073}{\sqrt{0.177}} = 4,93 \qquad \frac{\log \hat{\psi}_{M-H}}{\hat{\sigma}(\log \hat{\psi}_{MH})} = \frac{2.135}{\sqrt{0.218}} = 4,57$$

quant à (5) il vaut

$$\frac{(3 - \frac{1\times7}{5}) + (17 - \frac{2\times18}{5}) + (16 - \frac{3\times12}{5}) + (15 - \frac{4\times16}{5})}{\frac{1}{5}\sqrt{7 \times 4 + 18 \times 2 \times 3 + 17 \times 3 \times 2 + 15 \times 4 \times 1}} = 5,58$$

Les 3 résultats sont très voisins.

Il faut remarquer que la méthode (5) s'applique immédiatemen s'il y a un nombre variable de témoins par malade.

4 - PLUSIEURS FACTEURS DE RISQUE DE NATURE QUELCONQUE

Supposons que dans une enquête cas-témoins, où chaque malade es apparié à un témoin, plusieurs facteurs de risque, qualitatifs ou quantitatifs, sont simultanément considérés. Appelons x_i et y_i les vecteurs des valeurs de ces facteurs respectivement chez le malade et le témoin de la paire i.

La vraisemblance conditionnelle pour la paire i est (formule (3) du Chapitre XIV) est

$$\frac{e^{\beta x_i}}{e^{\beta x_i} + e^{\beta y_i}} = \frac{e^{\beta(x_i - y_i)}}{1 + e^{\beta(x_i - y_i)}} \qquad (6)$$

La vraisemblance totale est le produit sur i de ces vraisemblances élémentaires. La quantité (6) est la probabilité de succès dans un modèle logistique, quand la variable prend la valeur $x_i - y_i$. Ainsi pour estimer les paramètres β, on peut utiliser les programmes généraux d'estimation des paramètres de la fonction logistique, en prenant pour valeurs des variables " explicantes " $x_i - y_i$ (différences malades-témoins) et pour valeurs de la variable "expliquée" (succès ou échec) toujours la valeur "succès".

5 - GENERALISATION A UN NOMBRE QUELCONQUE (ET VARIABLE) DE TEMOINS PAR MALADE [3].

Si pour le malade i il existe k_i témoins appariés (k_i peut ne pas être constant par suite de données manquantes etc ...) et si x_i désigne les valeurs des facteurs de risque chez le malade, et $y_i^{(1)} \ldots y_i^{(k_i)}$ chez les témoins, la vraisemblance du (k_i+1) uple considéré est, toujours selon la même formule (3) du Chapitre XIV.

$$\frac{e^{\beta x_i}}{e^{\beta x_i} + e^{\beta y_i^1} + \ldots + e^{\beta y_i^{k_i}}} = \frac{1}{1 + e^{\beta(y_i^1 - x_i)} + \ldots + e^{\beta(y_i^{k_i} - x_i)}}$$

L'estimation des β et les tests portant sur leurs valeurs s'obtiennent selon la méthode habituelle.

6 - CAS PARTICULIER D'UN FACTEUR D'EXPOSITION A PLUSIEURS CLASSES

6-1. Le modèle

Soit l'exemple suivant emprunté à [4]. On veut rechercher par une enquête cas-témoins avec appariement, s'il existe un lien entre une maladie et le groupe sanguin. Les données (fictives) en nombre de paires figurent dans le tableau ci-dessous.

malades	témoins				
	O	A	B	AB	
O	64	18	8	3	93
A	66	74	14	6	160
B	4	2	4	2	12
AB	12	10	12	2	36
	146	104	38	13	

La variable étant à 4 classes on peut définir, ainsi qu'il a été dit au Chapitre XIII, 3 variables x_2, x_3, x_4 qui prennent

respectivement les valeurs 0 ou 1 selon que le groupe sanguin est A ou non A, B ou non B, AB ou non AB. Ainsi O est codé 0-0-0 ; A est codé 1-0-0 ; B est codé 0-1-0 et AB est codé 0-0-1.

Considérons une paire (ij) où i = 1,2,3,4 est le groupe sanguin du malade et j = 1,2,3, 4 est le groupe sanguin du témoin

$\beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$ valant 0, β_2, β_3, β_4 selon que le sujet est O, A, B, AB il est facile de voir que la vraisemblance (6) se réduit à :

$$\frac{e^{(\beta_i - \beta_j)}}{1+e^{(\beta_i-\beta_j)}} = \frac{\psi_i/\psi_j}{1+\psi_i/\psi_j} = \frac{\psi_i}{\psi_i+\psi_j} .$$

où l'on a posé $\psi_i = e^{\beta_i}$ (1 = 2,3,4) et $\psi_1 = 1$

La vraisemblance des paires (ij) est ainsi $\left(\frac{\psi_i}{\psi_i+\psi_j}\right)^{n_{ij}}$ et la vraisemblance de toutes les observations est

$$\prod_{i<j} \frac{\psi_i^{n_{ij}} \psi_j^{n_{ji}}}{(\psi_i+\psi_j)^{N_{ij}}} \qquad (N_{ij} = n_{ij} + n_{ji})$$

Pour estimer les ψ on peut estimer les β en utilisant les programmes généraux d'estimation de la fonction logistique, comme indiqué au paragraphe 5.

On trouve $\psi_1 = 1$, $\hat{\psi}_2 = 3.503$, $\hat{\psi}_3 = .559$, $\hat{\psi}_4 = 4.669$ qui conduisent pour le logarithme de la vraisemblance à une valeur de -81,40.

6-2. Test de la relation maladie x facteur

a) Tester cette relation revient à tester l'hypothèse nulle H_o : $\psi_2 = \psi_3 = \psi_4 = 1$. Sous H_o, la vraisemblance vaut

$\prod_{i<j} \left(\frac{1}{2}\right)^{N_{ij}}$; son logarithme est $- \log 2 . \sum_{i<j} N_{ij}$

$\sum_{i<j} N_{ij}$ est le nombre de paires discordantes, dans l'exemple 157. Le logarithme de la vraisemblance sous H_o vaut donc -108,82.

Le test du rapport des maximum de vraisemblance de H_o s'obtient en considérant la quantité

$2\ (108,82 - 81,40) = 54,84$ qui sous H_o est un χ^2 à 3 d.d.l.

On en conclut que la relation entre le groupe sanguin et la maladie est très hautement significative.

b) On peut construire un autre test de H_o (c'est la forme "score test", à laquelle il a été fait allusion au Chapitre XIII).

Considérons les totaux marginaux $N_{k.}$ et $N_{.k}$. Comme sous H_o, les n_{ij} ne sont pas "très" différents des n_{ji}, il en est de même de $N_{k.}$ et $N_{.k}$. Le test revient donc à comparer les 2 distributions marginales.

Cherchons les moyennes des $N_{.k}$ et leur matrice de variances-covariances sous H_o

$$N_{.k} = n_{1k} + n_{2k} + \ldots + n_{kk} + \ldots$$

Sous H_o $E(n_{1k}) = \frac{n_{1k} + n_{k1}}{2}$ et var $(n_{1k}) = \frac{n_{1k} + n_{k1}}{4}$ (binomiale de paramètres $\frac{1}{2}$).

tandis que $E(n_{kk}) = n_{kk}$ et var $(n_{kk}) = 0$ (on se place conditionnellement à $n_{1k} + n_{k1}$ fixé comme on le fait dans la méthode des couples).

De ces valeurs on déduit aisément que

$$E(N_{.k}) = \frac{N_{.k} + N_{k.}}{2}$$

$$\text{var}(N_{.k}) = \frac{N_{.k} + N_{k.}}{4} - \frac{n_{kk}}{2}$$

Cherchons maintenant la covariance de $N_{.k}$ et $N_{.\ell}$

$\text{cov}(n_{1k} + n_{2k} + \ldots + n_{lk} + \ldots, n_{1\ell} + n_{2\ell} + \ldots + n_{k\ell}) =$

$\text{cov}(n_{\ell k}, n_{k\ell}) = -\text{var}(n_{lk}) = -\frac{N_{\ell k}}{4}$

On vérifiera que la matrice V de variances covariances dans le cas de l'exemple précédent est

$$V = \left|\begin{array}{rrr:r} 27,75 & -21 & -3 & -3,75 \\ -21 & 29 & -4 & -4 \\ -3 & -4 & 10,5 & -3,5 \\ \hdashline -3,75 & -4 & -3,5 & 11,25 \end{array}\right|$$

Comme la somme des $N_{.k}$ est la même que celle des $N_{.k'}$ comparer les 2 vecteurs constitués par les distributions marginales se ramène à comparer seulement 3 composantes de ces vecteurs par exemple les 3 premiers.

La statistique de test est alors

$$(O-E)' \; V_2^{-1} \; (O-E)$$

où O est le vecteur des 3 premiers $N_{.k}$ observés (146 ; 104 ; 38), E le vecteur des valeurs attendues correspondantes (119,5 ; 132 25) et V_2 la matrice 3 x 3 de variances covariances correspondante.

Sous H_o, la statistique a une distribution de χ^2 à 3 degrés de liberté (nombre de modalités de E moins un). On trouve que

$$\chi^2_3 = 79,8$$

Cette valeur est assez éloignée numériquement de la valeur trouvée par l'autre test (54,84), mais à ce degré de signification, les différences ne sont guère importantes.

6-3. Test du modèle

On a obtenu comme valeur du risque relatif Ψ_2, du groupe A par rapport au groupe O la valeur 3,503.

En fait les données permettent d'avoir directement des estimations de Ψ_2.

Par exemple les paires O-A fournissent $\frac{66}{18} = 3,67$. Mais d'autres estimations sont possibles : ainsi le risque relatif

de A par rapport à B est $\frac{14}{2} = 7$ et le risque relatif de B par rapport à O est 4/8 =•5, qui fournit pour ψ_2 une nouvelle estimation 7 x .5 = 3,5, etc ...

Tester le modèle c'est voir s'il y a cohérence entre les différentes estimations. Ceci peut se faire en comparant les effectifs observés aux effectifs calculés sous le modèle.

L'effectif calculé de la case (ij) est tout simplement $N_{ij} \frac{\hat{\psi}_i}{\hat{\psi}_i \hat{\psi}_j}$. On obtient ainsi le tableau des effectifs calculés.

	O	A	B	AB
O	-	18,654	7,697	2,646
A	65,346	-	13,801	6,865
B	4,303	2,199	-	1,497
AB	12,354	9,135	12,503	-

Deux statistiques de test sont possibles:

a) la première est $G^2 = 2 \sum n_{ij} \log \frac{n_{ij}}{\hat{n}_{ij}}$ (c'est le test du rapport des maximum de vraisemblance)

b) la seconde est $\sum \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$

qui sous H_o sont toutes les 2 distributions selon un χ^2 à $\frac{(k-1)(k-2)}{2}$ d.d.l (k est le nombre de modalités du facteur).

On trouve ici que $G^2 = 0.50$ avec 3 d.d.l. Le modèle colle particulièrement bien aux données (fictives, rappelons-le).

Resterait à interpréter un résultat significatif : si des estimations du même ψ calculées sur des paires différentes sont différentes, c'est qu'il existe une interaction avec les facteurs sur lesquels est fondé l'appariement.

CHAPITRE XV - REFERENCES

[1] - O.S. MIETTINEN
Estimation of relative risk from individually matched series.
1970, Biometrics, 26, 75-86.

[2] - J. CONNETT, A. EJIGOU, R. McHUGH, N. BRESLOW
The precision of the Mantel-Haenszel estimation in case-control studies with multiple matching.
1982. Am. Journ. Epid. 116, 875-877.

[3] - N.E. BRESLOW, N.E. DAY; K.T. HALVORSEN, R.L. PRENTICE, C. SABAI.
Estimation of multiple relative risk functions in matched case-control studies.
1978. Am. Journ. Ep. 108, 299-307.

[4] - M.C. PIKE, J. CASAGRANDE, P.G. SMITH
Statistical analysis of individually matched case-control studies in epidemiology : factor under study a discrete variable taking multiple values.
1975. Brit. J. Prev. Soc. Med. 29, 196-201.