ANALYSE STATISTIQUE DES ENQUETES ANALYTIQUES

M. Sanaa [1]

Résumé

Cette présentation porte sur les principes de l'analyse des résultats d'une enquête analytique. Après la présentation des différentes mesures et des tests d'association entre un facteur étudié et la maladie, l'auteur développe les notions de facteur d'interaction et de facteur de confusion ainsi que leur prise en compte au moment de l'analyse. Les principaux modèles statistiques utilisés en épidémiologie et leur adaptation à l'épidémiologie animale sont brièvement présentés.

Summary

This presentation deals with the main principles followed to analyse the results of an analytical study. First the different measures and the association tests between a studied factor and the disease are presented. Then the notions of interaction factor and of confounding factor and how to take them into account during the analysis are developped. The main statistical models used in epidemiology and their adaptation to animal epidemiology are rapidly presented.

Les enquêtes analytiques ont pour objectif de déterminer le rôle d'un ou de plusieurs facteurs dans l'étiologie d'une maladie [Toma et al., 1991]. Leurs méthodes se fondent sur des comparaisons entre des groupes. Ces comparaisons permettent de tester et d'estimer la force d'association entre un supposé facteur causal et la maladie. Dans ce qui suit, le supposé facteur causal sera appelé facteur étudié ou exposition ; celle-ci peut être individuelle (par exemple, biologique, comportementale ou génétique) ou environnementale (physique, chimique, ou zootechnique liée aux pratiques d'élevage).

Classiquement, la recherche des causes d'une maladie est fondée sur le principe que si un facteur X est une cause de la maladie M alors X et M seront associés (dans le sens statistique du terme). En d'autres termes, si X est un facteur causal, alors la probabilité de la maladie sachant que le facteur X est présent $\{P(M|X)\}$ est supérieure à la probabilité de la maladie sachant que le facteur X est absent $\{P(M|\overline{X})\}$. Cela ne constitue qu'une première étape ; il est ensuite nécessaire d'interpréter la signification de la relation observée. Une association observée peut être fallacieuse, causale ou non-causale.

[1] E.N.V.A., Laboratoire d'épidémiologie et de gestion de la santé animale, 7 avenue du Général de Gaulle, 94704 Maisons-Alfort cedex, France

Une association fallacieuse peut être observée à cause d'un biais de sélection ou d'information lors de la réalisation de l'étude. Les différentes sources de biais sont développées dans deux autres textes de ce numéro spécial.

Une association non-causale peut être observée lorsque l'exposition est une conséquence de la maladie ou lorsque la maladie et le facteur étudié sont tous les deux associés à un tiers facteur C, connu ou inconnu. Dans ce cas, en mesurant l'association entre l'exposition et la maladie on aura, par inadvertance, mesuré celle entre C et la maladie. Le tiers facteur C est appelé facteur de confusion.

L'interprétation des résultats d'une enquête analytique est toujours délicate, en raison de l'absence d'une maîtrise totale des conditions d'observation. En effet, contrairement aux études expérimentales, les enquêtes analytiques sont par définition des études d'observation où l'épidémiologiste n'intervient pas dans le courant des événements. En expérimentation, le tirage au sort des individus qui seront affectés au lot des exposés et au lot des non exposés au facteur étudié, apporte une garantie déterminante sur la comparabilité de ces deux lots. Dans le modèle expérimental, on suppose que tous les autres facteurs susceptibles de modifier le risque de la maladie sont répartis également entre les deux lots. Les effets des tiers facteurs sont en principe annulés et si l'on observe une différence entre les deux groupes elle pourrait être attribuée au facteur étudié.

Cependant, le modèle expérimental est simpliste car il n'envisage le rôle que d'un seul facteur. Alors qu'en général les maladies sont multifactorielles. Pour que la maladie apparaisse, il faut l'intervention d'un complexe de plusieurs facteurs (appelés facteurs de risque). Le modèle expérimental s'adapte mal au cas où la maladie est multifactorielle car il devient très difficile de maîtriser l'affectation de plusieurs facteurs dans une seule étude et il n'est pas possible d'étudier les interrelations entre les différents facteurs. Seules les études épidémiologiques sont capables de fournir des informations sur ces interrelations et les inter-agissements de ces facteurs sur la maladie. L'inconvénient majeur des enquêtes épidémiologiques est l'absence de "comparabilité" entre les groupes comparés car on n'est jamais à l'abri d'un tiers facteur entraînant des biais sur le test et l'estimation de l'association entre le facteur étudié et la maladie. L'analyse des résultats des enquêtes analytique est, de ce fait, plus complexe que celle des études expérimentales. Elle nécessite le recours à des méthodes statistiques multivariées permettant de tester et d'estimer l'association d'un facteur étudié en fonction des effets de tous les facteurs de confusion potentiels.

Dans les enquêtes analytiques, la prise en compte des éventuels facteurs de confusion peut se faire soit au moment même de la planification de l'étude (stratification) en équilibrant la répartition du ou des facteurs de confusion connus entre les groupes à comparer (entre les exposés et les non exposés dans une enquête de cohorte, entre les cas et les témoins dans une enquête cas/témoins), soit au moment de l'analyse des données des résultats (ajustement) si les informations concernant l'exposition aux facteurs de confusion potentiels ont été recueillies. Les méthodes de stratification semblent n'être efficaces que dans les enquêtes de cohorte, et on est ramené le plus souvent à ajuster sur les facteurs de confusion même après une stratification. Avant de présenter les principaux modèles statistiques utilisés en épidémiologie, on abordera les principes des mesures et des tests de l'association entre l'exposition et la maladie ; ensuite on introduira les notions d'effets d'interaction et de confusion.

I - MESURE ET TEST DE L'ASSOCIATION ENTRE L'EXPOSITION ET LA MALADIE

Les données recueillies dans une enquête analytique devraient permettre d'étudier l'association sur deux aspects :

- Le premier concerne l'évaluation de la force et de l'importance de l'association entre le facteur étudié et la maladie. Ce qui nécessite le choix d'une mesure d'association. Le choix du type de mesure dépendra du choix d'un modèle, additif ou multiplicatif, qui lui même peut être lié au type de l'étude entreprise (de cohorte ou cas/témoins).

- Le second concerne l'évaluation de la stabilité de l'association ; cela relève des tests statistiques.

Dans cette partie, on suppose que le facteur étudié et la maladie sont caractérisés par des variables dichotomiques. Pour le facteur d'exposition, les deux modalités sont notées par E+ pour les individus exposés et E- pour les individus non exposés. Pour la maladie, les malades sont notés par M+ et les non malades par M-. On n'envisagera dans cette présentation que les cas où la fréquence de la maladie est mesurée par son risque (noté par R), c'est-à-dire la probabilité de survenue de la maladie pendant une durée d'observation fixée. Le lecteur peut consulter l'article de Goldberg et Leclerc [1990] pour les cas où la fréquence de la maladie est mesurée par l'incidence instantanée ou le taux d'incidence.

A - MESURE D'ASSOCIATION

La mesure d'association exprime la force ou l'intensité de la relation statistique entre un facteur étudié et la maladie. C'est une comparaison directe des mesures de fréquence pour différentes valeurs ou catégories du facteur étudié.

Afin d'établir plus clairement les définitions des différentes mesures d'association, on distinguera les deux types d'enquêtes (les enquêtes prospectives, les enquêtes cas/témoins). Les enquêtes de prévalence ou transversales ne seront pas traitées dans cette présentation. Ces dernières souffrent de nombreuses limitations qui ont été exposées dans le premier article sur la méthodologie d'enquête. Il est en effet difficile de classer ces enquêtes dans les études analytiques, car il est souvent nécessaire de confirmer leurs résultats à l'aide d'autres types d'enquêtes.

On traitera seulement le cas où le facteur de risque et la maladie sont des caractères dichotomiques.

1 - ETUDE DE COHORTE

Les résultats d'une enquête analytique peuvent être résumés dans un tableau de contingence.

Dans une enquête prospective, ce sont les effectifs N_1 et N_0 d'individus exposés et non exposés qui sont *a priori* fixés par l'investigateur (tableau I).

Tableau I : Tableau de contingence : étude de cohorte

	Maladie		
Exposition	***M+***	***M-***	**total**
E+	a	b	N_1
E-	c	d	N_0
total			N

La première étape de l'analyse consiste à calculer les risques dans les deux groupes à comparer : risque de survenue de la maladie chez les exposés et risque chez les non exposés. Pour chaque catégorie d'exposition, on estime un risque (respectivement R_1 et R_0) :

$$R_1 = \frac{a}{N_1} \quad \text{et} \quad R_0 = \frac{c}{N_0}$$

Dans une deuxième étape, l'analyse va consister à comparer les deux risques. Il s'agit de mesurer la différence entre les deux groupes. Généralement, dans une enquête de cohorte on mesure l'association entre l'exposition et la maladie, à l'aide de la différence de risque (ΔR) ou du risque relatif (RR).

❒ DIFFERENCE DE RISQUE

La différence de risque (ΔR) est la différence entre le risque observé dans le groupe des exposés au facteur étudié et le risque observé dans le groupe des non exposés :

$$\Delta R = R_1 - R_0 = \frac{a}{N_1} - \frac{c}{N_0}$$

La différence de risque représente le risque en excès associé à l'exposition au facteur de risque étudié ; le risque observé chez les non-exposés est considéré comme le risque de base auquel on compare le risque dans le groupe exposé au facteur. Lorsque le facteur E n'est pas associé à la maladie M, on a : $R_1 = R_0$ *et* $\Delta = 0$.

Le choix de la différence de risque comme mesure d'association suppose que le modèle choisi est additif. Dans ce modèle, la quantité ΔR qui caractérise l'effet du facteur E est supposée être constante même lorsqu'on passe d'une population à une autre, c'est-à-dire même si la fréquence R_0 de la maladie varie. Dans de telles conditions, si on a deux facteurs A et B dont les différences de risque sont notées par ΔR_A et ΔR_B, la différence de risque liée à l'exposition conjointe des deux facteurs A et B est :

$$\Delta R_{AB} = \Delta R_A + \Delta R_B$$

Une différence de risque peut être négative. Dans ce cas, le facteur étudié serait en fait un facteur protecteur.

❒ RISQUE RELATIF

Le risque relatif est le rapport du risque observé dans le groupe des exposés au facteur étudié sur le risque observé dans le groupe des non exposés :

$$RR = \frac{R_1}{R_0} = \frac{a/N_1}{c/N_0}$$

Lorsque le facteur E n'est pas associé à la maladie M, on a : $R_1 = R_0$ *et* $RR = 1$.

Si on mesure la relation avec RR, on suppose que le modèle est multiplicatif. Dans ce cas on suppose que le rapport des risques (RR) reste constant d'une population à une autre. Comme pour la différence de risque, on peut montrer que si on a deux facteurs A et B dont les RR sont notés par RR_A et RR_B, le risque relatif lié à l'exposition conjointe des deux facteurs A et B est :

$$RR_{AB} = RR_A \times RR_B$$

Exemple:

Le tableau II présente des résultats fictifs inspirés d'une enquête réalisée dans une population de vaches laitières multipares. L'événement étudié est la présence de kystes ovariens entre le 40ème et le 305ème jour de lactation. Le facteur étudié est la présence d'antécédents de kystes lors des lactations précédentes.

Tableau II : Répartition des cas de kystes ovariens chez 148 240 vaches laitières multipares

Antécédents de kystes	Kyste ovarien +	Kyste ovarien -	Total	*R*
+	431	4.961	5.392	*0,14*
-	2 995	142.250	145.245	*0,04*

D'après les données du tableau II, l'excès de risque de développer un kyste ovarien en rapport avec des antécédents de kystes ovariens serait de :

$$\Delta = 0,14 - 0,04 = 0,10$$

Le risque de développer un kyste ovarien chez les vaches sans antécédent de kyste ovarien, par rapport à celui chez les vaches avec des antécédents de kystes ovariens, est d'après les données du tableau II de :

$$RR = \frac{0,14}{0,04} = 3,5$$

En d'autres termes, les vaches avec des antécédents de kystes ovariens ont 3,5 fois plus de risque de développer un kyste ovarien que les vaches sans cet antécédent.

Un risque relatif peut être inférieur à 1. Dans ce cas, le facteur étudié serait, en fait, un facteur protecteur.

2 - ETUDE CAS/TEMOINS

Les résultats d'une enquête cas/témoins peuvent être résumés dans un tableau de contingence.

Dans ce type d'enquête, ce sont les effectifs M_1 et M_0 de cas et de témoins qui sont *a priori* fixés par l'investigateur (tableau III).

Tableau III : Etude cas/témoins :
cas où le facteur de risque (E) est dichotomique

Exposition	*Maladie* *M+ (cas)*	*M- (témoins)*	total
E+	a	b	
E-	c	d	
Total	M_1	M_0	N

❒ ODDS RATIO

D'après sa conception, l'étude cas/témoins ne permet pas d'estimer le risque de survenue de la maladie. Mais elle permet d'estimer les probabilités suivantes :

P(E+/ CAS) et P(E+/ TEMOIN)

qui sont les probabilités conditionnelles d'être exposé au facteur selon qu'on est un cas ou un témoin.

A partir de ces deux probabilités, on va calculer un odds (anglicisme signifiant pari ou cote) dans chacun des groupes. L'odds (ou la cote) est un concept différent de celui de risque de survenue d'un événement. La cote est le rapport de la probabilité de survenue d'un événement à son complément. C'est donc le rapport entre le nombre d'exposés et le nombre de non exposés observés dans le groupe considéré.

Dans le groupe des cas, cette quantité vaut (tableau III) :

$$\frac{P(E+|cas)}{1-P(E+|cas)} = \frac{P_{E_1}}{1-P_{E_1}} = \frac{a/M_1}{c/M_1} = \frac{a}{c}$$

qui traduit la chance relative pour un cas d'avoir été exposé par rapport à celle de ne pas l'avoir été. On peut dire aussi que a/c est une mesure de la cote en faveur de l'exposition contre la non-exposition chez les cas*.

On peut calculer cette même quantité dans le groupe des témoins (tableau III) :

$$\frac{P(E+|témoin)}{1-P(E+|témoin)} = \frac{P_{E_0}}{1-P_{E_0}} = \frac{b/M_0}{d/M_0} = \frac{b}{d}$$

qui traduit la chance relative pour un témoin d'avoir été exposé par rapport à celle de ne pas l'avoir été. On peut dire aussi que b/d est une mesure de la cote en faveur de l'exposition contre la non-exposition chez les témoins.

Dans une enquête cas/témoins, la mesure de l'association entre le facteur étudié et la maladie se fait habituellement à l'aide du rapport de la cote d'être exposé pour les cas sur la cote d'être exposé pour les témoins. Ce rapport est le plus souvent appelé odds ratio (OR). D'après le tableau III :

$$OR_E = \frac{P_{E_1}/1-P_{E_1}}{P_{E_0}/1-P_{E_0}} = \frac{a/c}{b/d} = \frac{ad}{bc}$$

Lorsqu'il n'y a pas d'association entre la maladie M et le facteur E, on doit s'attendre à observer chez les deux groupes les mêmes cotes d'exposition : a/c=b/d, ce qui donne un odds ratio égal à 1. Par contre, si E est un facteur de risque, on doit s'attendre à observer une cote d'exposition plus élevée chez les cas que chez les témoins, et par conséquent un odds ratio supérieur à 1. L'association entre le facteur étudié et la maladie est d'autant plus forte que la valeur de l'odds ratio est plus élevée.

Dans une enquête de cohorte, on peut calculer le rapport des cotes de maladie : rapport de la cote d'être malade chez les exposés sur la cote d'être malade chez les non exposés. D'après le tableau I :

$$OR_M = \frac{R_1/(1-R_1)}{R_0/(1-R_0)} = \frac{a/b}{c/d} = \frac{ad}{bc}$$

On constate que les deux odds ratio, OR_E et OR_M, se calculent de la même façon par le rapport des produits croisés des effectifs des tableaux de contingence (tableaux I et II) : ad/bc. On peut montrer sans grande difficulté que les deux odds ratio sont équivalents.

Finalement, l'odds ratio peut être estimé dans tous les types d'enquête.

Comme dans le cas du RR, si on mesure la relation avec un OR, on suppose que le modèle est multiplicatif. Dans ce cas, on suppose que le rapport des cotes OR reste constant d'une population à une autre. Comme pour le RR, on peut montrer que si on a deux facteurs A et B dont les OR sont notés par OR_A et OR_B , le risque relatif lié à l'exposition conjointe des deux facteurs A et B est :

$$OR_{AB} = OR_A \times OR_B$$

❒ Relation entre le risque relatif et l 'odds ratio

L'odds ratio peut être considéré comme une estimation du risque relatif dans les quatre situations suivantes :

---

* La notion de cote se retrouve dans les courses ou les paris : quand un cheval est coté 5 contre 1, cela veut dire que 5 parieurs l'ont joué perdant contre 1 gagnant.

1. Lorsque la maladie est rare, l'odds ratio calculé dans une enquête de cohorte est numériquement très proche du RR. Cette propriété découle du fait que lorsque la fréquence de la maladie est faible, R_1 et R_0 sont petits et $(1-R_1)$ et $(1-R_0)$ sont proches de 1 :

$$OR = \frac{R_1 / (1-R_1)}{R_0 / (1-R_0)} \approx \frac{R_1}{R_0} = RR$$

Il faut souligner, d'une part, que la rareté de la maladie est à vérifier au niveau de la population et non pas sur l'échantillon et, d'autre part, que cette propriété de l'OR ne s'étend pas à la fréquence de l'exposition. En effet, si l'exposition est rare l'OR calculé dans une enquête cas/témoins sera proche du rapport des fréquences d'exposition chez les cas et les témoins qui n'est pas égal au RR :

$$OR = \frac{P_{E_1} / 1-P_{E_1}}{P_{E_0} / 1-P_{E_0}} = \frac{P_{E_1}}{P_{E_0}} \neq RR$$

2. Dans une enquête cas/témoins, on peut démontrer aussi que si la maladie est rare, l'odds ratio est numériquement très proche du risque relatif que l'on aurait pu calculer si on réalisait une enquête de cohorte.

3. Dans une étude cas/témoins, le recrutement des malades et des non malades n'est pas effectué en général à partir d'une cohorte bien définie, mais plutôt à partir d'une population constituée de malades et non malades. Cette population est supposée dynamique et stable : c'est-à-dire que les flux d'entrée et de sortie peuvent être considérés constants au cours de la période de recrutement. On distingue généralement deux types de protocoles selon que les cas sont recrutés parmi les cas incidents (les cas sont inclus dans l'étude lors de leur apparition dans la population) ou parmi les cas existants ("cas prévalents").

   Si l'étude est menée auprès des "cas incidents" dans une population dynamique mais stable durant la période de recrutement, alors l'odds ratio calculé dans l'enquête cas/témoins est la même mesure que le risque relatif.

4. Si l'étude cas/témoins porte sur des "cas prévalents", si la population est dynamique et stable, si la durée moyenne de la maladie chez les individus exposés est la même que celle chez les non exposés, alors l'odds ratio est égal au risque relatif.

   L'odds ratio calculé dans une enquête cas/témoins avec recrutement des "cas prévalents" peut être exprimé en fonction des durées moyennes de la maladie chez les exposés et les non exposés (respectivement D_1 et D_0), et du risque relatif :

$$OR = \frac{P_1 / (1-P_1)}{P_0 / (1-P_0)} = \frac{D_1}{D_0} \times \frac{ID_1}{ID_0} = \frac{D_1}{D_0} \times RR$$

   Si $(D_1 \neq D_0)$ l'écart entre l'OR et le RR est d'autant plus important que la durée de la maladie est plus longue chez les exposés que chez les non-exposés. Dans de telles conditions, on se pose la question sur la nature du facteur : est-il réellement un facteur de risque ou un facteur pronostic de la maladie ? Pour cette raison, l'hypothèse d'indépendance entre l'exposition et la durée de la maladie est essentielle dans les enquêtes cas/témoins à recrutement prévalent. Si cette hypothèse ne peut être vérifiée ou admise, on ne pourra pas exclure le fait que la relation observée soit due à des biais de sélection.

B - TEST DE L'ASSOCIATION

Le simple calcul d'une différence de risque, d'un risque relatif ou d'un odds ratio n'est autre qu'une estimation ponctuelle de ce que pourrait être la vraie mesure d'association entre l'exposition et la maladie. Cette estimation ne donne aucune idée sur la fiabilité du résultat obtenu.

On peut alors utiliser un test statistique permettant d'évaluer la vraisemblance de l'association observée, c'est-à-dire que l'on estime la probabilité que l'observation faite puisse être expliquée par le simple hasard.
La stratégie d'un test statistique est celle de réfutation d'hypothèse, c'est-à-dire le rejet de l'hypothèse soumise à l'épreuve. On ne peut pas alors confirmer directement une hypothèse mais on peut rejeter son contraire. La méthode des tests statistiques consiste alors à formuler deux hypothèses :

- D'une part, l'hypothèse préalable selon laquelle il n'existe pas d'association entre l'exposition et la maladie, appelée hypothèse nulle et que l'on note par H_0 ; elle sous-entend que l'association observée n'est due qu'aux fluctuations inhérentes à toute procédure d'échantillonnage ;

- Et d'autre part, l'hypothèse alternative correspondant à l'existence d'association entre l'exposition et la maladie, notée par H_1.

Dans une enquête de cohorte, H_0 peut être exprimée différemment selon la mesure d'association retenue, à savoir la différence de risque ($H_0 : \Delta R = 0$) ou le risque relatif ($H_0 : RR = 1$). Dans une enquête cas/témoins, la seule formulation possible est en fonction de l'odds ratio ($H_0 : OR = 1$).

Quel que soit le type d'enquête ou la mesure d'association retenue, le test statistique sera le même. En effet, dans tous les cas, le test revient à comparer deux proportions observées (proportions des malades chez les exposés et les non-exposés, ou proportion des exposés chez les cas et les témoins, selon le type de l'étude). On peut utiliser alors les tests classiques de comparaison de deux pourcentages, écart réduit ou χ^2 de Pearson, sous réserve de vérification des conditions d'application [Schwartz, 1969].

Dans l'analyse des résultats d'enquêtes épidémiologiques, on utilise habituellement le test χ^2 de Mantel-Haenszel. Sa formule est la suivante :

$$\chi^2_M = \frac{[a-A]^2}{V(A)} = \frac{\left(a - \frac{N_1 M_1}{N}\right)^2}{\left(\frac{M_0 M_1 N_0 N_1}{N^2(N-1)}\right)} = \frac{(ad-bc)^2(N-1)}{M_0 M_1 N_0 N_1}$$

$$\text{où} \begin{cases} A = E(a) = \dfrac{N_1 M_1}{N} \\ V(A) = \dfrac{N_1 N_0 M_1 M_0}{N^2(N-1)} \end{cases}$$

L'avantage du test χ^2 de Mantel-Haenszel est qu'il est également utilisé dans les analyses stratifiées (cf. partie prise en compte d'un facteur de confusion).

Le résultat d'un test statistique, à savoir l'acceptation ou le rejet de l'hypothèse nulle, repose sur la comparaison entre la valeur du test obtenue et une valeur critique provenant d'une table statistique spécifique à chaque type de test. Pour le test χ^2, on utilise la table de la loi du χ^2.

Toute décision aboutissant au rejet de l'hypothèse nulle implique un certain risque d'erreur appelé risque de première espèce ou risque α, souvent fixé à 5 p.100. Cela veut dire que l'on accepte une probabilité de 5 p.100 de rejeter l'hypothèse nulle alors qu'elle est vraie. La valeur critique d'un test est déterminée en fonction du risque d'erreur α, plus ce risque est faible plus la valeur critique est élevée.

Habituellement, lorsqu'on rejette l'hypothèse nulle, on conclut à l'existence d'une association statistique entre le facteur étudié et la maladie, et cette conclusion est donnée avec un certain degré de signification appelé valeur de *P*. Le degré de signification exprime la probabilité d'obtenir une mesure d'association supérieure ou égale à celle obtenue dans l'enquête alors qu'en réalité il n'y a en fait aucune association réelle entre l'exposition et la maladie. Plus la valeur de *P* est faible plus l'association est incompatible avec l'hypothèse nulle, et plus le test est significatif.

Un test statistique ne doit pas être utilisé comme un résumé de l'association entre l'exposition et la maladie. En effet, il n'apporte aucune information ni sur la force de l'association ni sur son sens ; il ne donne en fait qu'une indication sur le rôle éventuel du hasard dans le résultat observé.

La valeur d'un test statistique est proportionnelle à la taille de l'échantillon de l'étude. L'absence d'un résultat significatif peut résulter soit de l'absence d'association soit de la faible taille de l'échantillon. On parle dans ce dernier cas de manque de puissance. La figure 1 présente des exemples d'évolution de la valeur du test de χ^2 en fonction de la taille de l'échantillon. Plus la taille de l'échantillon est grande, plus la valeur du test est élevée et la valeur de *P* faible.

Un test statistique dépend également de la force d'association. D'après la figure 1, le test χ^2 est plus significatif, à taille d'échantillon égale, lorsque l'association est plus forte (OR = 3 *vs.* OR = 1,5).

Figure 1 : Evolution de la valeur du test de χ^2 et son degré de signification (valeur de P) en fonction de la taille de l'échantillon

Figure 1a : Valeurs du test de χ^2

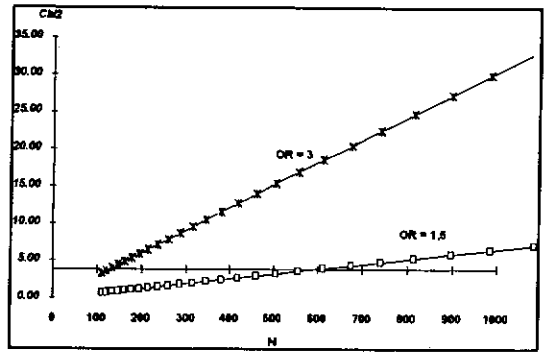
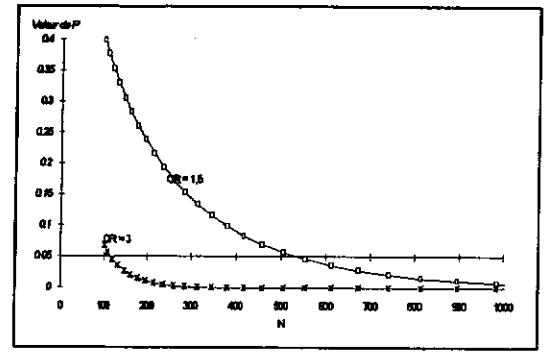


Figure 1b : Valeurs de P



Dans les deux graphiques l'axe des abscisses coupe l'axe des ordonnées au seuil de signification du test (à 3,84 pour la figure 1a et à 0,05 pour la figure 1b)

Compte tenu des remarques évoquées ci dessus, il est préférable de présenter en plus du résultat du test de χ^2 l'estimation ponctuelle de l'importance de l'association (ΔR, RR ou OR) assortie de son intervalle de confiance.

C - CALCUL DE L'INTERVALLE DE CONFIANCE DE LA MESURE D'ASSOCIATION

❒ INTERVALLE DE CONFIANCE D'UNE DIFFERENCE DE RISQUE

L'intervalle de confiance à 95 p.100 d'une différence de risque se calcule selon la formule suivante :

$$IC \text{ à } 95 \text{ p.100} = \Delta R \pm 1{,}96\sqrt{\frac{R_1(1-R_1)}{N_1} + \frac{R_0(1-R_0)}{N_0}}$$

❒ INTERVALLE DE CONFIANCE D'UN RISQUE RELATIF

Le risque relatif et l'odds ratio ont des valeurs possibles comprises entre $[1,+\infty]$ dont la distribution n'est pas normale. Cependant, leur logarithme a une distribution proche de la loi normale. Du fait de cette propriété, on calculera tout d'abord l'intervalle de confiance de Ln(RR) (logarithme népérien de RR), puis on en déduit celui du risque relatif en prenant les exponentielles des bornes de l'intervalle précédent.

L'intervalle de confiance de Ln(RR) est :

$$IC \text{ à } 95 \text{ p.100} = Ln(RR) \pm 1{,}96\sqrt{\frac{b}{N_1 a} + \frac{d}{N_0 c}}$$

ou directement :

$$IC \text{ à } 95 \text{ p.100} = RR \times \exp\left(\pm 1{,}96\sqrt{\frac{b}{N_1 a} + \frac{d}{N_0 c}}\right)$$

❒ INTERVALLE DE CONFIANCE D 'UN ODDS RATIO

Pour les mêmes raisons évoquées pour le risque relatif, on calculera tout d'abord l'intervalle de confiance de Ln(OR) (logarithme népérien du OR), puis on en déduit celui de l'odds ratio en prenant les exponentielles des bornes de l'intervalle précédent.

L'intervalle de confiance de Ln(RR) est :

$$IC \text{ à } 95 \text{ p.100} = Ln(OR) \pm 1{,}96\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

ou directement :

$$IC \text{ à } 95 \text{ p.100} = OR \times \exp\left(\pm 1{,}96\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}\right)$$

II - INTERACTION ET FACTEUR DE CONFUSION

Dans les enquêtes analytiques, on est souvent amener à étudier l'association entre la maladie et plusieurs facteurs. Plusieurs situations sont alors possibles : soit on s'intéresse simultanément au rôle étiologique de plusieurs facteurs, soit un ou des facteurs étiologiques sont déjà bien identifiés et l'on souhaite étudier le rôle spécifique d'un nouveau facteur, soit on souhaite tenir compte d'un facteur non étiologique lié par exemple aux conditions de diagnostic de la maladie et au facteur étudié.

Les deux dernières situations seront développées dans la partie facteur de confusion.

La situation où l'on étudie la relation entre l'exposition à plusieurs facteurs et la maladie sera traitée dans la partie effet d'interaction.

Afin de clarifier les notions d'interaction et de facteur de confusion, on se place dans le cas où l'on s'intéresse à l'étude de deux facteurs étiologiques dichotomiques.

A - EFFET D'INTERACTION

❒ DEFINITION

Dans le sens strict du terme, l'interaction est un phénomène qui permet à deux ou plusieurs facteurs étiologiques de se constituer comme *complexe causal* où l'action de chaque facteur sur la maladie devient un stimulus pour un autre et réciproquement. La présence des autres facteurs augmente l'effet du facteur étudié sur la maladie. On parle aussi de synergie entre les facteurs.

Plus généralement, on parle d'interaction entre deux facteurs étiologiques si la mesure de l'association d'un des deux facteurs se trouve modifiée en présence de l'autre facteur. L'interaction peut être positive (synergie) ou négative (antagonisme) selon que l'effet du facteur se trouve augmenté ou diminué en présence du ou des autres facteurs.

Par exemple, l'étude de Martin et al. [1989] montre une modification de l'association entre la séroconversion au virus B.V.D. (*bovine virus diarrhoea*) et la maladie respiratoire bovine en présence de séroconversion à *Pasteurella haemolytica* cytotoxique. La force de cette association, mesurée par l'odds ratio, était de 2,86 en absence de séroconversion à *Pasteurella haemolytica* cytotoxique et de 1,17 en présence de séroconversion à *Pasteurella haemolytica* cytotoxique.

La modification de l'association n'est pas un biais. C'est une information utile sur l'effet d'un facteur sur la maladie permettant d'identifier des sous-populations à haut risque et définir des actions préventives plus adaptées.

❒ DETECTION DE L 'INTERACTION

L'absence d'interaction entre deux facteurs s'écrit différemment selon le modèle utilisé :

Dans le ***modèle additif***, l'absence d'interaction s'écrit :

$$R_{11} - R_{10} = R_{01} - R_{00} \text{ soit}$$
$$R_{11} = R_{10} + R_{01} - R_{00}$$

Les indices sur les R notent la présence (1) ou l'absence (0) des facteurs X_1 et X_2. Ainsi, est le risque de la maladie dans le groupe des individus exposés au facteur X_1 et non exposés au facteur X_2.

L'absence d'interaction peut être testée à l'aide du test ' T ' présenté par Hogan et al. [1987] où $= R_{11} - R_{10} - R_{01} - R_{00}$. En absence d'interaction, la valeur de ' T ' devrait être proche de zéro.

Ce test d'interaction a le même principe que le test ε [Schwartz, 1969] :

$$\varepsilon = \frac{T}{\sqrt{V(T)}} \approx \textit{loi normale centrée réduite}$$

où la variance V(T) est égale à :

$$V(T) = \sum_{i=0}^{1}\sum_{j=0}^{1} \frac{R_{ij}(1-R_{ij})}{N_{ij}}$$

Lorsqu'on est dans le cas de ***modèle multiplicatif***, en utilisant le risque relatif comme mesure de l'association, on obtient :

$$RR_{11} = RR_{10} \times RR_{01}$$

où RR_{11} est le risque relatif de l'exposition aux deux facteurs, RR_{10} le risque relatif du facteur X_1 en absence de X_2 et RR_{01} le risque relatif de X_2 en absence de X_1. Le groupe non exposé à aucun des facteurs est pris comme groupe de référence.

De même, si l'odds ratio est la mesure choisie, on obtient :

$$OR_{11} = OR_{10} \times OR_{01}$$

Lorsque le modèle est multiplicatif, on dispose de plusieurs approches permettant de tester l'interaction. On peut citer les tests de décomposition χ^2 [Fleiss, 1981], et les tests d'homogénéité odds ratio ou du risque relatif entre les strates. Les tests d'interaction sont généralement peu puissants. Les conséquences pratiques en sont moins gênantes lorsqu'on cherche à vérifier une homogénéité plutôt qu'à détecter une interaction.

Les trois expressions de l'absence d'interaction ne sont pas équivalentes. Il n'y a en fait absence d'interaction que pour une mesure donnée. L'absence d'interaction pour une mesure n'implique pas son absence pour les autres. Toutefois, si la maladie est rare, les expressions en fonction du risque relatif et de l'odds ratio sont approximativement les mêmes. De ce fait, on distingue généralement deux sortes d'interactions : "additive" et "multiplicative" selon que le modèle choisi est additif ou multiplicatif.

B - FACTEUR DE CONFUSION

❒ Definition

Un facteur de confusion est un tiers facteur associé à la fois au facteur étudié et à la maladie, qui déforme la réalité de l'association entre l'exposition et la maladie.

Une forte association entre le facteur étudié et la maladie observée au cours d'une analyse brute peut être partiellement ou totalement due à un facteur de confusion. Un facteur de confusion peut également masquer, voire inverser, le sens d'une association.

Exemple

Soit une étude visant à rechercher si un vêlage difficile augmente la fréquence des troubles de fécondité chez la vache allaitante. Dans cette étude, des vaches avec vêlage difficile et sans vêlage difficile sont comparées et aucune différence nette n'est apparue entre ces deux groupes quant à la fréquence des troubles de fécondité. On peut supposer que les vaches avec vêlage difficile sont vraisemblablement plus jeunes que les vaches avec vêlage facile et comme les problèmes de fécondité augmentent avec l'âge, l'effet "condition de vêlage" est masqué. Dans cet exemple, l'âge est dit facteur de confusion de l'effet "condition de vêlage" sur les troubles de fécondité. L'analyse des résultats de cette enquête doit alors tenir compte de l'âge.

❒ Les conditions de confusion

Dans une enquête de cohorte, pour qu'un facteur C soit facteur de confusion, il faut à la fois qu'il soit :

- Associé à la maladie par classe du facteur étudié,
- Et associé au facteur étudié dans l'ensemble de la population (comprenant à la fois les malades et les non malades).

Dans une enquête cas/témoins, il y a confusion lorsque C est à la fois :

- Associé à la maladie par classe du facteur étudié,
- Et associé au facteur d'exposition conditionnellement au groupe des cas et au groupe de témoins. C'est-à-dire que la relation entre l'exposition et le facteur C doit être vérifiée aussi bien dans le groupe des cas que dans le groupe des témoins.

❒ Exemple

On s'intéresse à l'étude de l'association entre l'état de propreté des animaux et la contamination du lait de tank par *Listeria monocytogenes* dans une population de 2000 élevages bovins laitiers dont la répartition est indiquée dans le tableau IV. Cet exemple est une adaptation des résultats de l'étude de Sanaa et al. [1993].

Tableau IV : Répartition de la population des 2.000 élevages bovins laitiers selon l'exposition (propreté des animaux) et la contamination du lait de tank

Propreté des animaux	Contamination du lait		total
	oui	non	
sales	87	849	936
propres	13	105	1.064
	RR = 7,61		

Le risque de contamination du lait de tank par *Listeria monocytogenes* est 7,61 fois lorsque les animaux sont sales :

$$RR = \frac{87/936}{13/1064} = 7,61$$

La répartition de cette population en deux catégories (ou strates) selon la qualité de l'ensilage distribué aux animaux est présentée dans le tableau V :

Tableau V : Répartition de la population des 2.000 élevages en fonction de la qualité de l'ensilage

strate 1 : pH de l'ensilage < 4,0			
Propreté des animaux	Contamination du lait oui	non	total
sales	55	207	262
propres	5	133	138
	$RR_1 = 5,79$		

strate 2 : pH de l'ensilage > 4,0			
Propreté des animaux	Contamination du lait oui	non	total
sales	32	642	674
propres	8	918	926
	$RR_2 = 5,50$		

Les valeurs du risque relatif dans chacune des strates (RR_1 = 5,79 et RR_2 = 5,50) sont différentes de celle du risque relatif brut (RR = 7,61). Pour cette raison, le pH de l'ensilage est considéré facteur de confusion pour la relation entre la propreté des animaux et la contamination du lait de tank par *Listeria monocytogenes*. La différence entre le risque relatif brut RR et les risques relatifs conditionnels (RR_1 et RR_2) résulte de la réunion de deux phénomènes :

- D'une part, la qualité de l'ensilage est associée à la propreté des animaux. Le tableau VI, que l'on reconstitue à partir des données du tableau V, montre que la proportion des élevages distribuant un ensilage de mauvaise qualité est de 28 p.100 dans le groupe des élevages exposés (animaux sales) et 13 p.100 dans les élevages non exposés (animaux propres),

- D'autre part, le risque de contamination du lait est plus élevé lorsque l'ensilage est de mauvaise qualité (pH < 4,0), le RR = 6,00 (Tableau VII).

Tableau VI : Répartition de la population des 2.000 élevages bovins laitiers selon l'exposition (propreté des animaux) et la qualité de l'ensilage (pH)

pH de l'ensilage	Etat de propreté des animaux sale	propre
pH < 4,0	262	138
pH > 4,0	674	926
Total	936	1.064

Tableau VII : Répartition de la population des 2000 élevages bovins laitiers selon la qualité de l'ensilage (pH) et la contamination du lait de tank

pH de l'ensilage	Contamination du lait oui	non	total
pH < 4,0	60	340	400
pH > 4,0	40	1.560	1.600
	RR = 6,00		

C - PRISE EN COMPTE D'UN FACTEUR DE CONFUSION : ANALYSE STRATIFIEE

L'analyse de l'association entre un facteur d'exposition et la maladie nécessite dans la majorité des cas la prise en compte d'un facteur de confusion. On présente ici les méthodes simples de test et de mesure d'association ajustés sur un facteur de confusion.

On suppose que le facteur de confusion, C, est qualitatif à p classes : $C_1, C_2, \ldots, C_i, \ldots, C_p$. Dans chaque classe i (ou strate i), on retiendra les notations suivantes :

$$\left.\begin{aligned} R_{1i} &= P(M+|E+,C_i) \\ R_{0i} &= P(M+|E-,C_i) \end{aligned}\right\}$$

lorsqu'il s'agit d'une enquête de cohorte, et

$$\left.\begin{aligned} P_{1i} &= P(E+|M+,C_i) \\ P_{0i} &= P(E+|M-,C_i) \end{aligned}\right\}$$

lorsqu'il s'agit d'une enquête cas/témoins.

RR_i et OR_i notent respectivement le risque relatif et l'odds ratio pour la strate i.

Pour chaque strate, par exemple la ième, on peut construire le tableau de contingence suivant (tableau VIII) :

Tableau VIII : Tableau de contingence : strate i

Exposition	Maladie M+	Maladie M-	
E+	a_i	b_i	N_{1i}
E-	c_i	d_i	N_{0i}
	M_{1i}	M_{0i}	N_i

Selon le type de l'enquête, les différents paramètres de mesure d'association RR_i ou OR_i peuvent être calculés pour chaque strate. On peut également tester l'association entre le facteur étudié et la maladie, et calculer l'intervalle de confiance du paramètre de mesure pour la strate i. L'analyse statistique effectuée dans la strate i est dite analyse partielle ou conditionnelle à la strate i.

❒ Test d'ajustement de Mantel-Haenszel

Le test χ^2 d'ajustement de Mantel-Haenszel est le test le plus fréquemment utilisé. Il suppose l'absence d'interaction entre le facteur étudié et le facteur de confusion, c'est-à-dire que les odds ratios ou les risques relatifs calculés dans les différentes strates soient égaux.
Sa formule est la suivante :

$$\chi^2_{MH} = \frac{\left(\sum_{i=1}^{p} a_i - \sum_{i=1}^{p} A_i\right)^2}{\sum_{i=1}^{p} V(A_i)} \text{ où } \begin{cases} E(a_i) = \dfrac{N_{1i}M_{1i}}{N_i} \\ V(A_i) = \dfrac{N_{1i}N_{0i}M_{1i}M_{0i}}{N_i^2(N_i-1)} \end{cases}$$

Le résultat du test de Mantel-Haenszel se compare toujours aux valeurs de la distribution du χ^2 à un degré de liberté, et ce quel que soit le nombre de strates [Mantel et Haenzsel, 1959].

❒ Mesure d'association apres ajustement sur un facteur de confusion

Comme pour le test d'ajustement de Mantel-Haenzel, les mesures d'association ajustées, ou communes, supposent l'absence d'interaction entre le facteur étudié et le facteur de confusion. Lorsqu'on met en évidence un effet d'interaction, l'analyse se limitera logiquement à la présentation des résultats par strate. La mesure de l'association sur l'ensemble de la population serait forcément imparfaite puisqu'elle masquerait la variabilité de la force d'association entre les strates.

○ Estimation du risque relatif commun

Lorsqu'il y a absence d'interaction, le risque relatif commun peut être calculé selon la méthode suivante :

$$RR_c = \exp\left[\frac{\sum w_i Ln(RR_i)}{\sum w_i}\right]$$

RR_i est le risque relatif calculé dans la strate i et $Ln(RR_i)$ est son logarithme népérien. w_i est le poids de la strate i qui est l'inverse de la variance du logarithme népérien du risque relatif de cette strate :

$$w_i = 1\Big/\left(\sqrt{\frac{b_i}{N_{1i}a_i} + \frac{d_i}{N_{0i}c_i}}\right)$$

La variance du logarithme du risque relatif commun est égale à l'inverse de la somme des poids de chacune des strates :

$$V(Ln(RR_c)) = 1/\sum w_i$$

○ Estimation de l'odds ratio commun

Lorsqu'il y a absence d'interaction, l'odds ratio commun peut être calculé selon la méthode suivante :

$$OR_c = \exp\left[\frac{\sum w_i Ln(OR_i)}{\sum w_i}\right]$$

OR_i est l'odds ratio calculé dans la strate i et $Ln(OR_i)$ est son logarithme népérien. w_i est le poids de la strate i qui est l'inverse de la variance du logarithme népérien du odds ratio de cette strate :

$$w_i = 1\Big/\left(\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}\right)$$

La variance du logarithme de l'odds ratio commun est égale à l'inverse de la somme des poids de chacune des strates :

$$V(Ln(OR_c)) = 1/\sum w_i$$

❒ Exemple

On reprend les données du tableau V.

Le tableau IX présente les détails de calcul du test d'ajustement de Mantel-Haenszel et de la mesure d'association entre la propreté des animaux et la contamination du lait par *L. monocytogenes* après ajustement sur la qualité de l'ensilage.

Tableau IX : Test de Mantel-Haenszel et mesure d'association après ajustement sur un facteur de confusion (données du tableau V)

	pH de l'ensilage	
	< 4,0 : *Strate 1*	> 4,0 : *Strate 2*
a_i	55,00	32,00
A_i	39,30	16,85
$V(A_i)$	11,55	9,51
$\chi^2_{MH} = \frac{[(55+32)-(39,30+16,85)]^2}{11,55+9,51} = 44,06$		
RR_i	5,79	5,50
$Ln(RR_i)$	1,76	1,70
$V(LnRR_i)$	0,21	0,15
w_i	4,83	6,51
$RR_c = \exp\left(\frac{4,83 \times 1,76 + 6,51 \times 1,70}{4,83 + 6,51}\right) = 5,62$		

Le risque relatif ajusté sur l'effet qualité ensilage (RR_c = 5.62) est inférieur au risque relatif brut (RR = 7,61). Le rapport entre le risque relatif brut et le risque relatif ajusté, appelé rapport de confusion, est de 1,35. Il montre un écart de 35 % entre ces deux mesures.

❒ Strategie de l 'analyse stratifiee

Les différentes étapes à suivre dans une analyse stratifiée sont les suivantes :

1. *Réalisation d'une analyse brute* : Les résultats de cette première étape doivent fournir une estimation ponctuelle de la force d'association, son intervalle de confiance et le degré de significativité (valeur de *P*) du test statistique employé.

2. *Stratification de l'échantillon* : Elle consiste à construire un tableau de contingence pour chacune des modalités de la variable supposée être un facteur de confusion ou d'interaction. Une analyse brute est effectuée par la suite dans chacune des strates.

3. *Recherche de l'interaction* : Elle se fait en comparant les estimateurs de la force d'association obtenus dans les différentes strates. Pour que le facteur soit considéré facteur d'interaction, il faut que le test statistique de l'interaction soit significatif et que les différences entre les mesures soient importantes en termes épidémiologiques.

4. *Ajustement* : Cette étape dépendra du résultat de la précédente. Si on a mis en évidence une interaction, l'analyse s'arrête. Si on décide que les résultats de la troisième étape sont compatibles avec une absence d'interaction, l'analyse continue.

Cette dernière étape consiste à calculer l'estimateur ajusté et à le comparer avec l'estimateur brut de la force d'association (calculé en 1). Il n'existe malheureusement pas de méthode permettant de tester la différence entre les estimateurs ajusté et brut. On utilise alors des règles empiriques : si la différence d'un estimateur par rapport à l'autre dépasse 15 ou 20 p.100, on conclut que le facteur est vraisemblablement facteur de confusion. Dans de tels cas, on présente la mesure ajustée, avec son intervalle de confiance et le résultat du test d'ajustement de Mantel-Haenzsel. Lorsque les différences entre les résultats de l'analyse brute et de l'analyse stratifiée sont minimes (c'est-à-dire < 15 p.100), il est acceptable de présenter les résultats bruts pour un souci de simplification.

III - LES PRINCIPAUX MODELES STATISTIQUES UTILISES EN EPIDEMIOLOGIE

Dans les parties précédentes, on a uniquement envisagé le cas d'un facteur qualitatif à deux classes, étudié soit seul (analyse brute), soit en fonction d'un facteur de confusion qualitatif à deux ou à plusieurs classes (analyse stratifiée). Cependant, dans la plupart des cas, on a à analyser simultanément l'association entre la maladie et plusieurs facteurs qualitatifs ou quantitatifs. L'analyse stratifiée devient impossible dans de tels cas, et on a recours aux méthodes statistiques multivariées.

Les méthodes statistiques multivariées consistent à exprimer une variable Y décrivant la maladie, comme une fonction de plusieurs autres variables caractérisant les facteurs de risque ou

de confusion potentiels. En terminologie statistique, la variable Y est appelée variable dépendante (ou à expliquer), tandis que les autres sont dites variables indépendantes (ou explicatives).

Le choix des variables indépendantes dépend de la nature et du but de l'enquête. Dans une enquête analytique, le but est de comprendre l'apparition de la maladie. L'objectif de l'analyse est alors de mesurer et de tester le lien entre l'exposition et la maladie après avoir pris en compte les effets des facteurs de confusion potentiels ou connus. On inclut donc dans le modèle les facteurs d'exposition étudiés et un certains nombre de facteurs de confusion potentiels ou connus.

Dans une enquête analytique dont le but principal serait de prédire (par exemple, l'issue de la maladie en fonction de certaines variables), les variables sont retenues dans le modèle en fonction de leur qualité prédictive.

Un modèle statistique peut s'écrire sous la forme suivante :

$$E(Y|X_1,X_2,...,X_p) = f(X_1,X_2,...,X_p)$$

où $E(Y|X_1,X_2,...,X_p)$ est la valeur moyenne de Y pour des valeurs particulières des variables indépendantes (X), lorsque Y est quantitative, sa fréquence lorsque Y est qualitative à deux classes et f une fonction mathématique des variables indépendantes (X).

Un modèle statistique permet de mesurer l'association entre Y et une variable indépendante tout en prenant en compte l'effet des autres facteurs. Les tests et les mesures d'une analyse statistique multivariée sont des tests et des mesures ajustés.

Les principaux modèles utilisés en épidémiologie sont : les modèles linéaires, les modèles logistiques et les modèles de survie.

A - LES MODELES LINEAIRES

Les modèles linéaires sont utilisés lorsque la variable décrivant la maladie est quantitative : c'est le cas par exemple du poids de l'animal, de sa production, d'un dosage biologique ...

Le modèle linéaire s'écrit :

$$E(Y|X_1,X_2,...,X_p) = \alpha + \beta_1X_1 + \beta_2X_2 + ... + \beta_pX_p$$
$$= \alpha + \sum_{i=1}^{p} \beta_iX_i$$

où α est la moyenne générale de la variable Y, β_i le coefficient mesurant la relation entre Y et X_i. β_i peut être interprété comme l'accroissement de Y correspondant à l'accroissement d'une unité de X_i. L'utilisation du modèle linéaire suppose que la variable Y suit une loi normale.

Il existe plusieurs variantes du modèle linéaire en fonction de la nature des variables indépendantes $(X_1,X_2,...,X_p)$.

On parle de modèle d'analyse de variance (ANOVA) lorsque toutes les variables sont qualitatives, de modèle de régression multiple lorsque toutes les variables sont quantitatives, et de modèle d'analyse de covariance (ANCOVA) lorsque les variables sont mixtes.

B - LE MODELE LOGISTIQUE

Le modèle logistique, ou régression logistique, peut être utilisé lorsque Y est binaire (absence/présence de la maladie).

Le modèle linéaire s'écrit :

$$P(M+|X_1,X_2,...,X_p) = \frac{\exp\left(\alpha + \sum_{i=1}^{p}\beta_iX_i\right)}{1+\exp\left(\alpha + \sum_{i=1}^{p}\beta_iX_i\right)}$$

En utilisant la transformation logit on obtient :

$$Logit\left[P(M+|X_1,X_2,...,X_p)\right] = Ln\frac{P(M+|X_1,X_2,...,X_p)}{1-P(M+|X_1,X_2,...,X_p)}$$
$$= \alpha + \sum_{i=1}^{p}\beta_iX_i$$

La fonction logit est le logarithme népérien de l'odds d'une proportion.

Deux conditions doivent être vérifiées pour utiliser le modèle logistique : la variable indépendante doit être binaire (malade/non malade) et les informations sur ce statut doivent être recueillies à une date fixe. La dernière condition est vérifiée dans les enquêtes de cohorte lorsqu'on fixe une date de point pour l'analyse et dans les enquêtes cas/témoins.

L'interprétation des paramètres du modèle logistique dépend de la nature des variables indépendantes et de leur codage.

Variable qualitative binaire : elle est codée 0 en absence de l'exposition et 1 en présence de l'exposition. Le facteur qualité de l'ensilage, pH < 4 / pH > 4 est une variable de cette sorte. Lorsqu'on choisit ce codage, le paramètre β correspondant à cette variable est le logarithme de l'odds ratio mesurant l'association entre l'exposition et la maladie.

Variable qualitative à plusieurs classes : c'est le cas par exemple de la saison de vêlage.

Le codage le plus utilisé est celui qui transforme la variable à k classes en (k-1) variables à deux classes. Si la variable saison de vêlage est à 4 classes ; automne, hiver, printemps et été, on créera trois nouvelles variables : hiver, printemps et été. Les trois nouvelles variables (X1, X2, X3) doivent permettre de repérer les quatre catégories initiales (tableau X) :

Tableau X : Codage d'une variable à 4 classes

variable initiale	hiver X1	printemps X2	été X3
automne	0	0	0
hiver	1	0	0
printemps	0	1	0
été	0	0	1

Le paramètre β de la variable X1 est le logarithme de l'OR mesurant l'accroissement de l'odds du risque entre le groupe des vaches qui mettent bas l'hiver et celui des vaches qui mettent bas l'automne. Celui de la variable X2, est le logarithme de l'OR entre les vaches qui mettent bas l'hiver et celles qui mettent bas l'automne ...

Variable quantitative : lorsqu'on utilise la valeur réelle de la variable, le paramètre s'interprète comme l'accroissement de l'odds du risque correspondant à l'accroissement d'une unité de X.

C - LES MODELES DE SURVIE

Les modèles de survie sont utilisés lorsque la variable Y est binaire (malade/non malade) et qu'on s'intéresse à la date de survenue de la maladie. Ce modèle est souvent utilisé dans l'analyse des études pronostiques et dans les études prospectives où le risque de survenue de la maladie évolue au cours du temps.
Le modèle le plus utilisé est celui du Cox. Ce modèle permet de décrire et d'analyser la relation entre le risque instantané d'apparition de la maladie et un ensemble de variables indépendantes. Il se base sur l'hypothèse des risques proportionnels c'est-à-dire que le rapport des risques entre deux individus est indépendant du temps et ne dépend que des caractéristiques des individus (leur exposition). Le principe de ces méthodes et leur application dans la recherche biomédicale sont développés par Hill et al. [1990].

L'ensemble des modèles qui viennent d'être présentés supposent que les observations sont indépendantes. Cette condition d'application n'est pas vérifiée dans l'ensemble des enquêtes épidémiologiques animales, et ce notamment lorsqu'on s'intéresse à la fois à des facteurs d'exposition à l'échelle individuel et à l'échelle d'élevage. Les observations recueillies sur les animaux appartenant à un même élevage ne peuvent pas être considérées indépendantes. Il existe en effet un effet troupeau (ou élevage) dont il faut tenir compte au moment de l'analyse.

IV - PRISE EN COMPTE DE L'EFFET ELEVAGE

Les études épidémiologiques vétérinaires sont fréquemment confrontées à l'analyse de données provenant de populations naturellement regroupées dans des élevages. Les individus appartenant à un même élevage partagent un certain nombre de facteurs tels la technicité de l'éleveur, l'alimentation, la génétique ou des facteurs d'environnement. Ces facteurs peuvent influencer le développement de la maladie au sein de l'élevage ou le succès d'interventions préventives ou curatives sur les animaux. C'est pourquoi, dans de nombreuses enquêtes épidémiologiques en élevage, on ne peut considérer les réponses des individus d'un même élevage comme indépendantes.

Si l'on connaît depuis longtemps l'existence d'un effet troupeau, sa prise en compte dans l'analyse d'enquêtes épidémiologiques est beaucoup plus récente, notamment lorsqu'il s'agit de variable binaire. Une étude récente de McDermott et al. [1994] a montré que sur 67 articles d'épidémiologie vétérinaire, 31 d'entre eux ignoraient la présence d'un éventuel effet élevage. Plusieurs auteurs ont montré que l'utilisation du modèle logistique ordinaire donnait des estimations biaisées des paramètres, de leurs écarts-types ainsi que des résultats de test des paramètres lorsque l'indépendance des observations recueillies n'était pas satisfaite [Curtis et al., 1993 ; Goelema et al., Atwill et al., 1993].

On peut expliquer ce biais intuitivement de la façon suivante : supposons que l'on soit dans la situation extrême où les réponses de tous les individus d'un même élevage soient identiques (effet élevage maximal). Le nombre réel d'observations indépendantes correspond au nombre d'élevages que nous appellerons I. Supposons maintenant qu'à l'inverse, les observations soient totalement indépendantes (absence d'effet élevage). Toutes les observations sont informatives et leur nombre correspond au nombre total d'individus N. Lorsqu'il existe en pratique un effet élevage, on est dans une situation intermédiaire entre I et N. Utiliser la régression logistique ordinaire revient à se placer dans la situation fausse où le nombre d'observations indépendantes est égal à N. Ceci explique que l'on sous-estime les variances des paramètres et que l'on conclut à tort à la signification de certains paramètres, notamment ceux concernant des facteurs d'élevage. L'étude de McDermott et al. [1994] a montré que parmi 31 études épidémiologiques vétérinaires n'ayant pas pris en compte le facteur élevage, 26 d'entre elles donnaient des conclusions erronées.

Il existe plusieurs approches permettant de prendre en compte une corrélation intra-groupe.

Une première approche, fréquemment utilisée, consiste à réduire les réponses individuelles recueillies dans un élevage à un unique paramètre. L'information recueillie sera résumée par une moyenne s'il s'agit d'une variable continue ou par un pourcentage s'il s'agit d'une variable discontinue. Cette variable sera ensuite modélisée en fonction des différents paramètres d'élevage. Cette méthode est appelée "two stage" ou "derived variable analysis", ce qui signifie "analyse en deux étapes". La condition implicite d'utilisation de cette méthode est l'absence de variables explicatives individuelles.

Plus généralement, on distingue deux autres approches permettant de tenir compte de l'effet élevage : celle modélisant une réponse spécifique à chaque individu (cluster-specific model) et celle modélisant une réponse moyenne de la population (population-average model).

Avec le premier type de modèles (cluster-specific model), la probabilité de la survenue de la maladie est modélisée en fonction des facteurs d'exposition étudiés et de paramètres α_i spécifiques au groupe i. Cette approche inclue les modèles logistiques à effets mixtes [Mauritsen, 1984] et le modèle logistique stratifié [Breslow et Days, 1980].

Dans la seconde approche (population-average model), c'est la probabilité marginale de la survenue de la maladie qui est modélisée au delà des groupes en fonction des facteurs d'exposition étudiés [Liang et Zeger, 1986].

V - CONCLUSION

Les méthodes présentées dans cet article sont bien adaptées aux situations où l'on peut facilement isoler une variable à expliquer (la maladie), et un petit nombre de variables explicatives pertinentes ; ceci est souvent le cas dans les études d'épidémiologie analytique. Cependant, dans certaines études les connaissances sur le sujet étudié ne sont pas très avancées et on s'écarte des conditions idéales d'utilisation de ces méthodes. On se retrouve ainsi avec un nombre très élevé de variables explicatives dont toutes ne seront pas pertinentes : certaines sont redondantes, d'autres ont des associations non significatives avec la maladie. Dans cette situation, que l'on peut qualifier de "partiellement exploratoire",

l'utilisation directe des modèles statistiques multivariés n'est pas la bonne solution, même si les logiciels et les moyens de calcul le permettent. Le grand nombre de modèles à essayer et la multiplicité des tests font que le risque de montrer des associations devient grand. Une solution à ce problème serait par exemple l'utilisation en première étape de méthodes descriptives multivariées permettant de mieux comprendre les liens entre les variables explicatives et de choisir parmi elles les variables les plus pertinentes à introduire dans les modèles.

BIBLIOGRAPHIE

Atwill E. R. , Rodriguez L., Hird D. W. and Rojas O.- Environmental and host factors associated with seropositivity to New Jersey and Indiana vesicular stomatitis viruses in Costa Rican cattle. *Prev. Vet. Med.*, 1993, 15, 303-314.

Breslow N. E. and Days N.E.- Statistical methods in cancer research. *Lyon : International Agency for Research on Cancer*, 1980.

Curtis C. R., Mauritsen R. H., Kass P. H., Salman M.D. and Erb H. N.- Ordinary versus random-effects logistic regression for analysing herd-level calf morbidity and mortality data. *Prev. Vet. Med.*, 1993, 16, 207-222.

Fleiss J.L.- Statistical methods for rates and proportions. Wiley, New York. 2e edition, 1981.

Goelema J. O., Jansen J. and Frankena K.- Effect of extra binomial variation on parameter estimates and standard errors in application of multiple logistic regression to veterinary epidemiology. *Proceedings of the 6th International Symposium on Veterinary Epidemiology and Economy,Ottawa, Canada,* 1991, 87-89.

Goldberg M. et Leclerc A.- Epidémiologie à visée étiologique, concepts, mesures et méthodes d'investiagation. *Epidémiol. Santé Anim.*, 1990, 18, 1-30.

Hill C., Com-Nougue C., Kramar A., Moreau T., O'Quigley J., Senoussi R., Chastang C.- Analyse statistique des données de survie. Inserm, Médecine Sciences, Flammarion, 1990, Paris.

Hogan M.D., Kupper L.L., Most B.M. and Haseman J.K.- Alternatives to Rothman's approach for assessing synergism (or antagonism) in cohort studies. *Am. J. Epidemiol.*, 1987, 108, 60-67.

Liang K. Y. and Zeger S. L.- Longitudinal data analysis using generalized linear models. *Biometrika*, 1986, 73, 13-22.

Mantel N. and Haenszel W.- Statistical aspects of the anlysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.*, 1959, 22, 719-748.

Martin S.W. , Bateman K.G., Shewen P.E., Rosendal S. and Bohac J.E.- The frequency, distribution and effects of antibodies to seven putative repiratory pathogens on respiratory disease and weight gain in feedlot calves. *Can. J. Vet. Res.*, 1989, 53, 355-362.

Mauritsen R. H.- Logistic regression with random effects. PhD Thesis, University of Washington, Seattle, 1984.

Mc Dermott J.J, Schukken Y.H, and Shoukri M.M.- Study design and analytic methods for data collected from clusters of animals. *Prev. Vet. Med.*, 1994, 18, 175-191.

Sanaa M., Poutrel B., Ménard J.L. and Serieys F.- Risk factors associated with contamination of raw milk by *Listeria monocytogenes* in dairy farms. *J. Dairy* Sci., 1993, 76 : 2891-2898.

Schwartz D.- Méthodes statistiques à l'usage des médecins et des biologistes, Flammarion, Paris, 1969, 318 p.